HHS Public Access

Author manuscript

Proc IEEE Int Conf Big Data. Author manuscript; available in PMC 2020 March 17.

Published in final edited form as:

Proc IEEE Int Conf Big Data. 2019 December ; 2019: 4067–4070. doi:10.1109/bigdata47090.2019.9006234.

bench4gis: Benchmarking Privacy-aware Geocoding with Open Big Data

Daniel R. Harris,

Institute for Pharmaceutical Outcomes and Policy, College of Pharmacy, Center for Clinical and Translational Sciences, University of Kentucky, Lexington, KY USA

Chris Delcher

Institute for Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Kentucky, Lexington, KY USA

Abstract

Geocoding, the process of translating addresses to geographic coordinates, is a relatively straight-forward and well-studied process, but limitations due to privacy concerns may restrict usage of geographic data. The impact of these limitations are further compounded by the scale of the data, and in turn, also limits viable geocoding strategies. For example, healthcare data is protected by patient privacy laws in addition to possible institutional regulations that restrict external transmission and sharing of data. This results in the implementation of "in-house" geocoding solutions where data is processed behind an organization's firewall; quality assurance for these implementations is problematic because sensitive data cannot be used to externally validate results. In this paper, we present our software framework called bench4gis which benchmarks privacy-aware geocoding solutions by leveraging open big data as surrogate data for quality assurance; the scale of open big data sets for address data can ensure that results are geographically meaningful for the locale of the implementing institution.

Keywords

geographic information systems; geospatial analysis; big data applications

I. Introduction

Geocoding is the process of digitally translating a street address to longitude and latitude coordinates on the Earth's surface; geocoding facilitates geospatial analysis such as descriptive counts per geographic unit and distance calculations between point coordinates. Patient privacy is the primary concern when geocoding healthcare data; in the U.S. specifically, the Health Insurance Portability and Accountability Act (HIPAA) considers address and geographic coordinates as protected health information and consequently imposes restrictions on data use, transmission, and sharing. Geospatial analysis with healthcare data can be an expensive, elaborate process where great care and diligence is

daniel.harris@uky.edu.

needed to fully comply with regulations and data security requirements [1]. Geographic privacy must be maintained at two different levels: during geocoding and data matching [2], [3] and during analysis and aggregation [4]–[6]. The benefits of integrating geospatial information into healthcare analytics include helping clinicians and researchers better understand contextual information about a patient's environment and potentially leading to specific targeted interventions [1]; by considering location as a vital sign, physicians may also leverage knowledge of geographical social-determinants of health when promoting certain healthre-sources while being sensitive to a patient's access to medical care [7]. Additionally, there is additional evidence that geospatial information is also useful for secondary research use in clinical data warehouses [8].

There are three types of well-known geocoding strategies [9]; we discuss each of these with respect to healthcare data:

1. *"in-house" geocoding*: This refers to any system or solution implemented behind an institution's firewall where protected health information must remain. Other than controlling user access, there are no regulatory limitations since addresses are not shared beyond the firewall and all computation occurs locally.
2. *geocoding web-services*: This refers to geocoding offered as a web-service. Geocoding web services, such as the ArcGIS World Geocoding service and Google Maps Geocoding service, offer an application programming interface (API) which provides geocoding functionality; these typically require a subscription or a charge-per-request to do large scale computation. In order for these web services to be used with protected health information, a business associates agreement (BAA) must be established, outlining data sharing and compliance needs; there is additional cost and time associated with this as it is a legal process.
3. *pass-through cloud services*: This refers to leveraging existing BAAs in conjunction with hosted webservices. Institutions may already have BAAs with cloud providers, such as Azure and Amazon, and may be able to satisfy local privacy regulations by obfuscating the origin of the geocode request by passing the address through a cloud-hosted web service, which makes the external request.

Web-services for geocoding, such as those mentioned in strategies 2 and 3 above, are convenient to use, but restrictions on data usage due to privacy concerns limit their usefulness in practice. Furthermore, charge-per-request models of geocoding limit their usefulness with larger data sets for institutions requiring economical processing; free rate-limited web-services which only allow for a few thousand requests per day are not be useful for data sets that contain millions of records. These two concerns suggest that "in-house" geocoding solutions are the most appropriate option for processing large-scale healthcare data. Despite this, the accuracy of "in-house" solutions cannot be easily tested since the results of processing sensitive data cannot be externally verified. In this paper, we discuss "in-house" solutions for geocoding and our software framework, bench4gis, which streamlines benchmarking geocoding solutions by leveraging open big data sets when sensitive, private data sets cannot be externally validated.

II. Open Big Data for Addresses

OpenAddresses is an online repository containing over half a billion geocoded addresses, which are annotated with longitude and latitude coordinates [10]. These addresses represent a wide variety of countries from every continent except Antarctica. Table I shows the count of addresses available across the four regions of the United States of America.

Given the scale of the OpenAddresses data and large breadth of coverage, we hypothesize that OpenAddresses data can act as a surrogate for sensitive data in testing and benchmarking local geocoding solutions; the computed geographic coordinates from an in-house geocoding implementation can be compared to the original source coordinates to measure how well the system is performing. The performance of geocoders may vary according to both geographic region of the source coordinates and the quality of the reference data being used; geocoding the address data from OpenAddresses may reveal any geographical or workflow issues for a given “in-house” implementation. We borrow from other comparative analyses of geocoding and use metrics such as hit rate and point-to-point accuracy [11]. For US states, coverage varies; sparsely populated states may need to consider a region also containing neighboring states.

III. Geocoding

Commercial solutions for geocoding exist, but we focus on open-source options due to their wide-spread availability and easy adoption. These geocoding solutions are not specifically designed for use with healthcare data, but offer software capable of processing addresses “in-house” without calling external web-services, making them practical options for geocoding sensitive data sets. Our framework for leveraging big open data for benchmarking will apply to any geocoding solution.

A. Open-source Geocoders

Both open-source communities and private companies have published open-source geocoding solutions, including but not limited to PostGIS TIGER Geocoder [12], OpenStreetMap’s Nominatim [13], and Gisgraphy [14]. From academia, geocoding tools such as DeGAUSS [15], [16], Texas A&M Geocoder [17], [18], and EaserGeocoder [19] attempt to streamline geocoding for specific research purposes. For our benchmarking framework, we utilized the PostGIS TIGER Geocoder and OpenStreetMap’s Nominatim, which have significantly different geocoding workflows. We chose these two geocoders for our experiments and benchmarking code development because they represent two large classes of geocoders: database-driven and web-services-driven; our benchmarking code will be applicable to any geocoders within these classes.

The PostGIS TIGER Geocoder is an extension for PostGIS [12], which adds support for spatial and geographic objects to the open-source PostgreSQL database [20]. Geocoding is implemented as a function which can be called within SQL queries or PL/pgSQL code; geospatial reference data is taken from U.S. Census Bureau’s TIGER (Topologically Integrated Geographic Encoding and Referencing) database [12]. Healthcare data is typically stored within relational databases, making a SQL-based solution both convenient

and intuitive. The limitation is that TIGER is only meaningful for geocoding U.S. addresses due to the dependency on U.S. Census Bureau data; however, PostGIS is fully capable of loading geospatial data from any source [12], [21].

OpenStreetMap's Nominatim is a web-application that supports geocoding via locally-hosted web-services; these services offer an API for geocoding and reverse geocoding [13]. Nominatim also uses PostGIS but with OpenStreetMap data loaded through the osm2pgsql tool; support for TIGER data is also available [13]. The geographic coverage is customized at load time; we chose data from the U.S. region, though international data is also available. Batch processing with Nominatim must use a programming language capable of making API calls. In the next section, we describe our bench4gis software and how it relates to these two types of geocoders.

B. bench4gis

Our software framework for benchmarking geocoding solutions using big open data, bench4gis, is available as open-source software [22]. As illustrated in Figure 1, two different workflows are supported: one for database-driven geocoding and one for web-services-driven geocoding. Both workflows start at the same point: downloading OpenAddresses data as a reference data set; a region is specified and any filtering of sub-regions occurs. The source files from OpenAddresses that did not get filtered out are then merged into a comma-separated values (CSV) file which can be input into either of the two supported workflows.

Database-driven geocoders leverage a database to store geospatial reference data and analytical functions, such as PostgreSQL [20] with the PostGIS extension [12]. With bench4gis, we packaged example data definition language (DDL) statements as a precursor to our code for loading the OpenAddresses data and managing batch geocoding. The loading code is optional if the user would prefer to load the OpenAddresses data using whatever techniques or tools they would prefer to use. Once the data is loaded, we provide code for generating a random table sample using PostgreSQL's Bernoulli table sampling method [20]; alternatively, the entire table of source addresses may be used for processing. After a sample is selected, we provide code for batch geocoding; we found that geocoding in batches of 1000 addresses performed reasonably well for our hardware configuration.

Web-services-driven geocoders depend on hosting a web-server, such as Apache's HTTP server, and a web-application, such as OpenStreetMap's Nominatim. Access is obtained through either a web-browser for individual requests or through API programming for batch processing. GeoPy is a geocoding library for Python that offers the ability to query a Nominatim instance by specifying the domain of the geocoding server; batch processing is supported by writing Python code to cycle through a data set and submit geocoding requests for each address [23]. bench4gis reads directly from the CSV source file aggregated from the included setup scripts and provides code for initializing a geocode worker process for a given address; Python's multiprocessing library then manages $N - 1$ worker processes where N is the number of CPUs. As addresses finish geocoding, additional geocode workers are triggered until there are none left unprocessed.

Once OpenAddresses addresses are geocoded using the “in-house” implementation, bench4gis also includes code for geospatial comparison of source data to computed data. Currently, point-to-point distance from source coordinates to computed coordinates is used to calculate accuracy; a query that generates general descriptive statistics is included.

C. Benchmarking and Discussion

We implemented the PostGIS TIGER geocoder and OpenStreetMap’s Nominatim to process a data set with protected health information. Our sensitive data set covers healthcare data for our state-wide hospital system and is geographically located in the southern region of the US; as seen in Table I, OpenAddresses contains 92.9 million addresses for the southern region. We chose the most geographically-relevant subset from OpenAddresses to process and test our “in-house” geocoding abilities. From the southern U.S. subset of OpenAddresses we used bench4gis to randomly choose and geocode a test set of one million addresses. We then compared the computed coordinates from both geocoders to the given coordinates in OpenAddresses by calculating point-to-point distance; the results are summarized in Table II. Geocoding is not guaranteed to return an address [9]; the hit rates in Table II indicate that the majority of the addresses were able to be geocoded. For successfully geocoded addresses, we report general summary statistics; extreme outliers are evident from the discrepancies between the minimum, median, average, and maximum distance from the reference coordinates; the majority of the addresses were computed within a mile (1609m) of the reference data for either geocoder. Although Nominatim had near exact matches, the TIGER geocoder performed better for addresses in our region; other regions may experience different results. The acceptable difference between source location and computed location varies by application: an analysis involving rural distances is less sensitive to error when compared to analysis involving smaller city block distances. For many healthcare applications, 70–80% geocoding accuracy to the street level is considered acceptable [24].

Because outliers challenge the usefulness of summary statistics, we also provide a box plot in Figure 2 where the whiskers extend to 1.5 times the interquartile range; extreme outliers are cropped from view for readability. This shows that the TIGER geocoder produced coordinates within a smaller range for point-to-point differences with the reference data.

There are limitations to using big open data as a reference. We use OpenAddresses data as a surrogate data set to test the accuracy of “in-house” geocoding solutions, but since OpenAddresses data is large, it potentially contains mistakes such as missing or erroneous fields. These mistakes impact the ability to successfully geocode OpenAddresses data but they also represent real world issues. Common mistakes we observed: invalid zip codes and street numbers, such as 0; typos in street names, such as *Crossgates* instead of *Crossgate*; confusion between city name and census place name.

IV. Conclusion and Future Work

We presented bench4gis, our software framework for using big open data to test and evaluate “in-house” geocoding solutions. When geocoding protected, private data, external validation of computed locations is typically not permitted; we have demonstrated that using big open

data as a surrogate data set can inform how well an implementation is performing with respect to a specific geographic region. As future work, we wish to explore how large identified differences from benchmarking with big open data might inform or identify potential mistakes in processing private, sensitive data. We wish to explore the timeliness of records in OpenAddresses and the impact of new roads. We also wish to define a workflow to support institutional research for geospatial analysis.

Acknowledgment

The project described was supported by the NIH National Center for Advancing Translational Sciences through grant number UL1TR001998. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1]. Buckingham WR, “The potential and pitfalls of geocoding electronic health records.” WMJ: official publication of the State Medical Society of Wisconsin, vol. 111, no. 3, pp. 107–111, 2012. [PubMed: 22870555]

[2]. Christen P, “Privacy-preserving data linkage and geocoding: Current approaches and research directions,” in Sixth IEEE International Conference on Data Mining-Workshops (ICDMW’06) IEEE, 2006, pp. 497–501.

[3]. ——, “Geocode matching and privacy preservation,” in International Workshop on Privacy, Security, and Trust in KDD. Springer, 2008, pp. 7–24.

[4]. Curtis A, Mills JW, Agustin L, and Cockburn M, “Confidentiality risks in fine scale aggregations of health data,” Computers, Environment and Urban Systems, vol. 35, no. 1, pp. 57–64, 2011.

[5]. Zandbergen PA, “Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data,” Advances in medicine, vol. 2014, 2014.

[6]. Murad A, Hilton B, Horan T, and Tangenberg J, “Protecting patient geo-privacy via a triangular displacement geo-masking method,” in Proceedings of the 1st ACM SIGSPATIAL International Workshop on Privacy in Geographic Information Collection and Analysis ACM, 2014, p. 6.

[7]. Sunderland N, Bristed H, Gudes O, Boddy J, and Da Silva M, “What does it feel like to live here? exploring sensory ethnography as a collaborative methodology for investigating social determinants of health in place,” Health & Place, vol. 18, no. 5, pp. 1056–1067, 2012. [PubMed: 22722015]

[8]. Gardner BJ, Pedersen JG, Campbell ME, and McClay JC, “Incorporating a location-based socioeconomic index into a de-identified i2b2 clinical data warehouse,” Journal of the American Medical Informatics Association, vol. 26, no. 4, pp. 286–293, 2019. [PubMed: 30715327]

[9]. Rivera B and Hoffman M, “Technical strategies for real-time geocoding in healthcare,” in 2018 IEEE International Smart Cities Conference (ISC2) IEEE, 2018, pp. 1–5.

[10]. Openaddresses: the free and open global address collection. Accessed Sept. 1, 2019 [Online]. Available: http://openaddresses.io

[11]. Roongpiboonsopit D and Karimi HA, “Comparative evaluation and analysis of online geocoding services,” International Journal of Geographical Information Science, vol. 24, no. 7, pp. 1081–1100, 2010.

[12]. Postgis: Spatial and geographic objects for postgresql. Accessed Sept. 1, 2019 [Online]. Available: https://postgis.net

[13]. Nominatim. Accessed Sept. 1 2019 [Online]. Available: http://www.nominatim.org

[14]. Gisgraphy: Open source geocoder and address database. Accessed Sept. 1, 2019 [Online]. Available: https://www.gisgraphy.com/

[15]. Brokamp C, Wolfe C, Lingren T, Harley J, and Ryan P, “Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies,” Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 309–314, 2017.

[16]. Brokamp C, “Degauss: Decentralized geomarker assessment for multisite studies.” J. Open Source Software, vol. 3, no. 30, p. 812, 2018.

[17]. Goldberg DW, “The usc webgis open source geocoding platform,” USC GIS, 2008.

[18]. Texas a&m geocoder: free online geocoding service. Accessed Sep. 29, 2019 [Online]. Available: http://geoservices.tamu.edu/Services/Geocode/About/

[19]. Rashidian S, Dong X, Jain SK, and Wang F, “Easergeocoder: integrative geocoding with machine learning,” in Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems ACM, 2018, pp. 572–575.

[20]. Postgresql: the world’s most advanced open-source database. Accessed Sept. 1, 2019 [Online]. Available: https://www.postgresql.org/

[21]. Rowlingson B, Lawson E, Taylor B, and Diggle PJ, “Mapping english gp prescribing data: a tool for monitoring health-service inequalities,” BMJ open, vol. 3, no. 1, p. e001363, 2013.

[22]. Harris/bench4gis. Accessed Sep. 29, 2019 [Online]. Available: https://bitbucket.org/_harris/bench4gis

[23]. Geopy: Geocoding library for python. Accessed Sept. 1, 2019 [Online]. Available: https://github.com/geopy/geopy

[24]. Dolan C and Delcher C, “Monitoring health inequities and planning in virginia: poverty, human immunodeficiency virus, and sexually transmitted infections,” Sexually transmitted diseases, vol. 35, no. 12, pp. 981–984, 2008. [PubMed: 18685545]

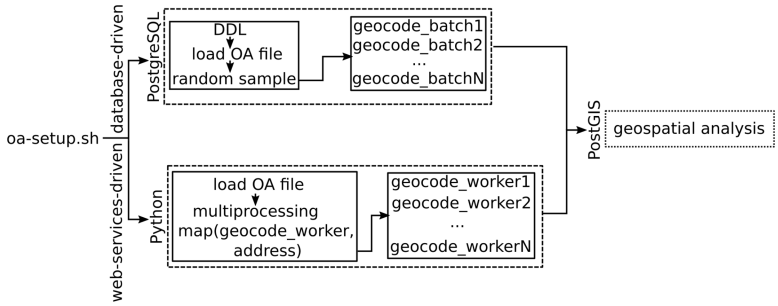

Fig. 1.
bench4gis workflows for OpenAddresses (OA) data

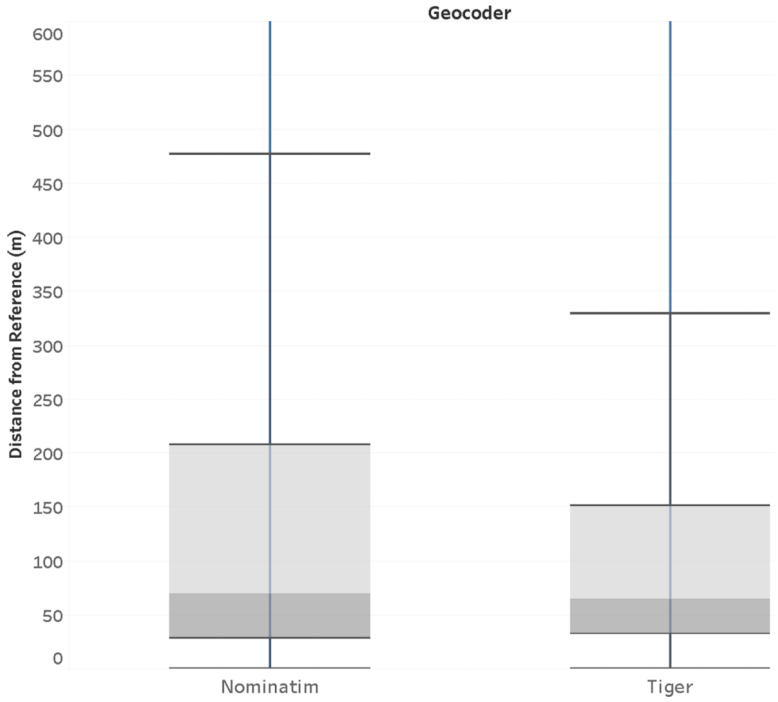

Fig. 2.
Box-plot of point-to-point distances between computed location and reference location.

TABLE I

Addresses per U.S. Region

U.S. Region	Addresses Available
Northeast	23,392,560
Midwest	36,804,438
South	92,918,156
West	43,722,713

TABLE II

Summary of Geocoding Results

Geocoder	Hit Rate	Distance from Reference (meters)					Proportion of Distances Under Threshold				
		Min.	Median	Max	Avg.	Std. Dev.	≤ 50m	≤ 100m	≤ 300m	≤ 600m	≤ 1609m
PostGIS TIGER	.99	.17	65	3,595,808.29	11,050.04	64,526.49	.40	.65	.83	.87	.90
Nominatim	.90	0	70	6,244,995.99	34,931.30	148,945.29	.36	.54	.71	.75	.77