Name:	Sylvia Cho
Affiliation:	Columbia University
Email:	sc3901@cumc.columbia.edu
Presentation type	Poster

Comparison and Evaluation on Online Geocoding Services

Sylvia Cho, MHS[1], Karthik Natarajan, PhD[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA

Abstract
Geographic data provides important information in healthcare research such as the spread of disease or geographic variations in healthcare access. As an ultimate goal to embed a geovisualization tool within OHDSI, we investigated which online geocoding service would be the most appropriate for our research purposes. The dataset includes address data of a cohort that we have been monitoring for up-to-date vaccine coverage. We found that the Google Maps, MapQuest, and the Data Science Tool Kit returns a high match rate, and reliable geocode results. Census Bureau geocoding service has a lower match rate compared to the others, but had a highly reliable result. Users should also consider that there are usage limits to Google Maps and MapQuest, whereas the Census Bureau and the Data Science Tool Kit does not have limits when choosing geocoding services.

Introduction
Geographic data has been widely used in the field of healthcare. For example, geographic data is important for disease surveillance where we can geovisualize the spread of disease.[1] Also, it could be helpful in understanding the geographic variation of healthcare access.[2] The OMOP Common Data Model (CDM) in the observational health data sciences and informatics (OHDSI) has a location table that shows a standard way to represent addresses. Adding latitude and longitude data to this table would allow in depth spatial health analysis. As an initial step to achieve this goal, we examined the performance of four online geocoding services on a dataset of addresses.

Methods
The dataset was selected from the immunization information system (IIS), EzVac which consists of a cohort of 14,531 patients that we monitor for vaccine coverage rates across the institution.[3] The dataset includes street, city, state, and zip code. All addresses were fully filled out. We compared four different geocoding services which are the US Census Bureau, Google Maps, MapQuest, and the Data Science Tool Kit. These were selected because they offered free service and were easy to process the large dataset. The geocodes of patient address was retrieved by using the application programming interface (API) to interact with the geocoding service websites. We used the batch geocoding service when using the Census Bureau, which accepts up to 1,000 addresses per file. Thus, we needed to split the original file into multiple files and merge them after the geocode was retrieved. For the other three geocoding services, we used the original file and geocoded the addresses serially. Our script was written in Python 2.7.10. We evaluated the results on match rate and similarity.[4] Match rate is the proportion of input addresses that retrieves a geocode from the geocoding system. Similarity is the distance measure between geocodes from two services. For example, if the distance between each geocode from Google Maps and MapQuest is greater than the distance between geocodes from Google Maps and the US Census Bureau, then we can say that the geocode from Google Maps is more similar to the geocode from the US Census Bureau than that of MapQuest. Only the addresses that had a matching geocode across all four services (N = 12,311) were evaluated on similarity.

Results
Match rate was lowest (84.77%) when using the US Census Bureau geocoding system (Table 1). The other three geocoding services retrieved most of the geocodes (99.9% - 100%). Similarity was generally greater between the geocodes from the US Census Bureau compared to the other three geocoding services. Similarity was lowest between the geocodes from MapQuest and the Data Science Tool Kit. In other words, the average distance between a pair of geocodes from different services is largest when comparing geocodes from MapQuest and Data Science Tool Kit (Table 2).

Table 1. Match Rate (N = 15,431)

	Match Rate
US Census Bureau	84.77%
Google Maps	99.938%
MapQuest	100%
Data Science Tool Kit	99.986%

Table 2. Average distance of geocodes between geocoding services (average miles; N =12,311)

	Census Bureau	Google Maps	MapQuest	Data Science Tool Kit
US Census Bureau	-	-	-	-
Google Maps	0.129	-	-	-
MapQuest	0.225	0.313	-	-
Data Science Tool Kit	0.166	0.26	0.36	-

Discussion

The US Census Bureau geocoding service has its strength in that it has no limit in the number of addresses that we can geocode in a certain time period. Also, the majority of the addresses are for residential areas, and few are available for commercial areas. The database is updated regularly at least once per year.[5] On the other hand, Google Maps has usage limits up to 2,500 free address requests per IP per day. It is also possible to exceed the usage limit if too many requests are sent per second.[6] MapQuest has usage limit up to 15,000 free address request per API key per month.[7] They offer both single line address geocoding and batch geocoding. However, batch geocoding accepts only up to 100 addresses per file. The Data Science Tool Kit does not have a usage limit, and its API uses data from the US Census Bureau and the OpenStreetMap.[8]

This research was the initial step for developing a geovisualization tool using healthcare data. We analyzed the match rate and similarity focusing only on free online geocoding services. High match rate alone does not indicate high accuracy of results. We compared the results across the four geocoding services. We assumed that if the distance between a geocode from one of the services compared to geocodes from other geocoding services was greater than 100km, it would be an unreliable result. We found that although the US Census Bureau geocoding service had the lowest match rate, its matched geocodes had less variation compared to the other services. In other words, the geocodes returned by the Census Bureau were approximately the same as the results from the other geocoding services. Google Maps, MapQuest, and the Data Science Tool Kit returned geocode results for almost all addresses, but a few number of geocodes were inconsistent when cross-compared with other services. However, considering the total number of addresses were over 12,300, overall the results seem reliable. Since the US Census Bureau provides reliable results and the batch geocoding service, developers who are interested in incorporating batch geocoding services into their system could pipeline the process so that another service can run the non-matching results from the US Census Bureau. For further studies, we plan to analyze the results on more geocoding quality metrics found in a framework study on evaluating geocoding systems.[9] We also plan to evaluate geocode discrepancies at a granular level, and furthermore, connect a visualization tool.

Conclusions

This research does not recommend one geocoding service over another. Each service has its own strengths and limitations. Thus, it would be important to consider the research goals and choose the most appropriate geocoding service. We believe that leveraging the OMOP CDM and the OHDSI tools can be very useful in performing spatial analysis. We hope to contribute our code if and when geolocation information is added to the OMOP CDM.

References

1. Lawson AB, Kleinman K. Spatial and syndromic surveillance for public health. Wiley Online Library; 2005;
2. Comber AJ, Brunsdon C, Radburn R. A spatial analysis of variations in health access: linking geography, socio-economic status and access perceptions. Int J Health Geogr. BioMed Central; 2011;10(1):1.
3. Vawdrey DK, Natarajan K, Kanter AS, Hripcsak G, Kuperman GJ, Stockwell MS. Informatics lessons from using a novel immunization information system. Stud Health Technol Inform. 2012;192:589–93.
4. Roongpiboonsopit D, Karimi HA. Comparative evaluation and analysis of online geocoding services. Int J Geogr Inf Sci. Taylor & Francis; 2010;24(7):1081–100.
5. United States Census Bureau. Census Geocoder [Internet]. Available from: https://www.census.gov/geo/maps-data/data/geocoder.html
6. Google Maps API [Internet]. Available from: https://developers.google.com/maps/
7. MapQuest Developer [Internet]. Available from: https://developer.mapquest.com/
8. Data Science Tool Kit [Internet]. Available from: http://www.datasciencetoolkit.org/
9. Goldberg DW, Ballard M, Boyd JH, Mullan N, Garfield C, Rosman D, et al. An evaluation framework for comparing geocoding systems. Int J Health Geogr. BioMed Central; 2013;12(1):1.