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

Background

- Geographic data has been widely used in the field of healthcare. For example, it can be used to geovisualize the spread of disease for disease surveillance.[1] More specifically, data from electronic health record can be used to geographically monitor diseases in real-time to identify the communities with healthcare needs.[2]
- The current OMOP Common Data Model (CDM) in the observational health data sciences and informatics (OHDSI) collaborative has a location table that shows a standard way to represent addresses.
- **Objective:** We suggest to add latitude and longitude data to this table for spatial health analysis on large scale observational data. Our motivation for adding this feature is to enable surveillance of vaccine coverage of our pediatric population. As an initial step to achieve this goal, we examined the performance of four online geocoding services on a dataset of addresses.

Methods

Dataset

- Cohort of 14,531 patients monitored for vaccine coverage rates selected from our institution's immunization information system (IIS), EzVac.[3]
- The dataset includes street, city, state, and zip code which follows the format of CDM version 5.

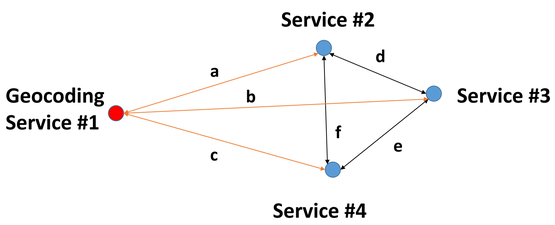
Geocoding services

- US Census Bureau : No usage limit, Batch geocoding (1000 addresses per file)
- Google Maps : Usage limit up to 2,500 free address requests per IP per day
- MapQuest : Usage limit up to 15,000 free address request per API key per month
- Data Science Tool Kit : No usage limit

Evaluation

- ***Match rate*** [4] ***:*** proportion of input addresses that retrieves a geocode from the geocoding system.
- ***Similarity*** [4] ***:*** distance measure between two geocode points from a pair of services.
  - Longer the distance, less similarity
  - Calculated only when the addresses had a matching geocode from both services.
- Counted the number of addresses where one of the geocoding services provides a geocode point that is more than 0.1 miles apart from that of the other services (Figure 1).
  - Threshold was set due to the dense population of New York City, and we wanted geocode points to be within a block or two of each other

Figure 1. Example of in-depth analysis on similarity



Analyzing similarity with a threshold of 0.1 miles:
a > 0.1 miles & b > 0.1 miles & c > 0.1 miles & d ≤ 0.1 miles & e ≤ 0.1 miles & f ≤ 0.1 miles

Geocode points for each services

Results

- Match rate was lowest (84.77%) when using the US Census Bureau geocoding system (Table 1).
- Similarity was generally greater between the geocodes from the US Census Bureau compared to the other three geocoding services (Table 2).
- Data Science Tool Kit had the most number of addresses that returned a geocode point more than 0.1 miles away from the corresponding geocodes acquired from other services (Table 3).

Table 1. Match rate (N = 14,531)

	Match Rate
US Census Bureau	84.77%
Google Maps	99.94%
MapQuest	100%
Data Science Tool Kit	99.99%

Table 2. Average distance between geocodes from geocoding services (average miles)

	Census	Google	MapQuest	DSTK
US Census Bureau	-	-	-	-
Google Maps	0.129 (N = 12,311)	-	-	-
MapQuest	0.225 (N = 12,318)	0.484 (N = 14,522)	-	-
Data Science Tool Kit (DSTK)	0.166 (N = 12,318)	0.965 (N = 14,521)	1.073 (N = 14,529)	-

Table 3. Number of addresses where geocodes are inconsistent

	Among addresses with match across all services (N = 12,311)	Among addresses with any non-matches (N = 2,200)	Total
US Census Bureau	52	-	52
Google	385	137	522
MapQuest	362	189	551
Data Science Tool Kit	792	609	1401

Conclusion

Findings

- Although the US Census Bureau geocoding service had the lowest match rate, its matched geocodes had more similarity (less variation) compared to the other services.
- Google Maps and MapQuest returned geocode results for almost all addresses, but a few number of geocodes were inconsistent when cross-compared with other services.
- The Data Science Tool Kit returned the highest number of inconsistent geocodes.
- Based on these findings, the users can choose the appropriate service that fits their goals.

Future Studies and Uses

- Connect a visualization tool to the updated database with geocodes.
- Use the geocodes that are populated in the updated table definition to visualize the geographic variations on whether people are up to date on vaccinations.
- Leveraging the OMOP CDM and the OHDSI tools would help other sites that are interested in geovisualizing their healthcare data.

Reference

1. Lawson AB, Kleinman K. Spatial and syndromic surveillance for public health. Wiley Online Library; 2005;
2. Laranjo L, Rodrigues D, Pereira AM, Ribeiro RT, Boavida JM. Use of Electronic Health Records and Geographic Information Systems in Public Health Surveillance of Type 2 Diabetes: A Feasibility Study. JMIR public Heal Surveill. JMIR Publications Inc.; 2016;2(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. Int J Geogr Inf Sci. Taylor & Francis; 2010;24(7):1081–100.

Contact: Sylvia Cho (sc3901@cumc.columbia.edu)