Positional Accuracy of Online Geocoding Services: Case Study of Bhaktapur District

Er. Amrit Karmacharya

KEYWORDS

Geocoding, Nominatim, Accuracy, Positioning, Location Based Service

ABSTRACT

Data is food for Information Systems and location data is basis for GIS services. Geocoding services provide this data by converting Street Addresses like NCIT, Balkumari to corresponding geographic coordinates. These coordinates are then used in data processing to deliver services. Many services (especially location based), FourSquare, Uber, depend on these services for operation. Nepalese market is also increasingly using these services like Tootle, sarathi cab. Till now nobody knows how accurate the result of such services are. There exist multiple places with same name. Some names are not actual but referential. Accuracy of the services depend on the underlying database, the method used, the actual geographic location and also the actual query. Different methods yield different results. The assessment of accuracy and suitability of the geocoding services has not been conducted, yet they are being used extensively. The objective of this study is to compare the positional difference between two Geocoding methods, OpenStreetMap (OSM) Nominatim Service and Google Geocoding Services and compare them with standard government datasets to measure their discrepancy. For reference, settlement data from NGIID was used. Addresses were first geocoded to street level and positional difference in the results were calculated using havensire formula. The discrepancies were categorized into intervals of 100m. Out of 267 location points, 118 result were found in Nominatim, whereas only 86 were found in Google. Average discrepancy for Nominatim result was 175m and for Google it was 1810m. Comparisons show minimal difference in median and minimum values, while there were larger differences in the maximum value. Nominatim delivered comparatively accurate results and found more addresses than google. Google on the other hand gave huge mismatches for some cases. The study found out that the databases are missing in both cases as shown by the no of "not found" cases and that the results from Nominatim are more reliable than that of Google because of its hierarchical matching system and user friendly interface.

1. INTRODUCTION

Geocoding is the processing of matching a description of a location to geographic coordinates. With the advances in web technologies and location based mapping, the traditional Geocoding tools provided in desktop GIS software are being increasingly replaced by online geocoding services. The Web geocoding services from various providers offer users an easier way to geocode place names to location coordinates in multiple text formats like extensive Markup Language (XML), JavaScript Object Notation (JSON), or Comma Separated Values(CSV).

Geocoding gives result in form of coordinate pair, usually latitude and longitudes pair. It may also give out extra information as to the shape and size of the features if the features were linear or areal. But mostly the result is in form of a point. The accuracy of a geocoding depends on the database used to perform the search and its hierarchical model.

Result of geocoding depends on the data used, Nepal government has published an Index of Geographical Names for the whole country. Google maps provides geocoding services but the sources of its data are unknown. OpenStreetMap mobilizes volunteers and local community to collect data directly on the field and provides free service for geocoding. In the current situation, the data from Nepal Government is not dense enough to locate places. The data from Google seem to be accurate but have not been verified. Also Google deliberately uses Easter Eggs (false information mixed with original data to identify if data is being stolen) which compromises its accuracy. Google is the most popularly used geocoding service in Nepal. OpenStreetMap data is unevenly distributed over the data, areas with active volunteers are better mapped whereas areas without volunteers are empty.

The assessment of accuracy and suitability of the geocoding services has not been conducted yet they are being used extensively.

1.1. Objective of the study

The objective of the study are as follows:

- To compare the positional difference in results provided by different services.

1.3 Limitations of the Study

Limitation of the study is as follows

- Address in rural areas do not have precisely defined boundaries, so the assessment of accuracy is based on human interpretation.
- Because of unavailability of accurate field data for reference, the results are comparative analysis only.
- The study is limited to settlements only. Geocoding application in other sectors like house numbering, street level geocoding, point of interest matching, have not been conducted.

2. METHODOLOGY

2.1 Source of Data

The address data was collected from the Topographical Base Map Data. National Geographic Information Infrastructure Project (NGIID) distributes the data. The data collected was of the Bhaktapur. The other data of google and open street map are accessed from the web.

2.2. Data Preparation

The data from various sources are in various projection system. The data from NGIID was from UTM system and the data from other geocoding services are in WGS 84 system. So, all of the data from NGIID was converted to WGS 84 for uniformity.

All postal addresses were preprocessed before geocoding to improve standardization and quality. We reviewed the data for misspelled address information and remedied any incorrect home addresses (e.g. incorrect names). In addition, we removed all extraneous characteristics and standardized the spelling. We removed address which were inside the sheet but outside the study area.

2.3. Preparation of comparison table

Comparison table (shown below) was prepared to compare the discrepancies between the different systems. The table consists of location. Its coordinates as given by the 3 different providers. the distance range between the derived coordinates computed using havensire formula.

		NGIID		OpenStreetMap				Google maps			
Sn	Name	Lon	Lat	Lon	Lat	discrepancy (km)	Remarks	Lon	Lat	discrepancy(km)	Remarks

2.4. Data Filing

For comparing the location information from three sources the location data from 3 sources are excelled. For this work, different sources have different system of acquiring the data.

Firstly, from NGIID the data is available in GIS format which can be converted to different format and as we required, the latitude and longitude can be generated and exported to excel file.

For the Nominatim (free geocoding services which uses geographic data from free and open OpenStreetMap project), we can enter the name of location in search box, then it will provide with the number of matches. There may be more matches, so for exact match we can input the location name with the higher level address as well. It will then provide with the area and lat, long of the centroid.

For the Google Geocoding service, there is a web application which can help us to input and output data from google easily. The app is available here http://googlemaps.github.io/js-v2-samples/geocoder/singlegeocode.html. The process is similar as in Nominatim except that the application only shows single result which is usually the first result returned by Google service. In such a case it is important to manually judge whether the location is the desired one or not.

3. DATA ANALYSIS

A total of Two hundred sixty-seven (267) address were searched and matched using the above mentioned procedure. The distance between the location provide by the two different services were compared using the havensire formula. Havensire formula gives distance between two set of coordinates which are in latitude longitude format and gives output in metric system. It takes into account for the distortion due to the curvature of earth and different scale factor at different latitude values. The values given by NGIID were used as a standard data and discrepancy were calculated from other two sources. The discrepancy was then categorized into interval of 100m. There is no standard fixed value to specify how large an area is related to a location. it is big in village areas while small in crowded areas.

4. RESULT

Of the 267 addresses matched, OpenStreetMap found 118 results, whereas Google found 86 results. There is huge variation seen in between these matches. OpenStreetMap matched addressed in the range 100 m to 17 km. Google did the same with range of 70 m to 1034km. This huge difference is because google does not provide user interaction in searches. The high difference is obviously error but they cannot be identified correctly. The average discrepancy in OpenStreetMap is 175 m and the average discrepancy in Google is 1810 m.

Discrepancy Between	0 - 100 m	100 - 200 m	200 - 300 m	300 – 400	400 - over	Total	Not Found
NGIID - OSM	0	15	36	11	56	118	149
NGIID - Google	1	8	8	4	65	86	181
OSM - Google	2	8	4	4	34	52	215

Table 1: Table of Discrepancy

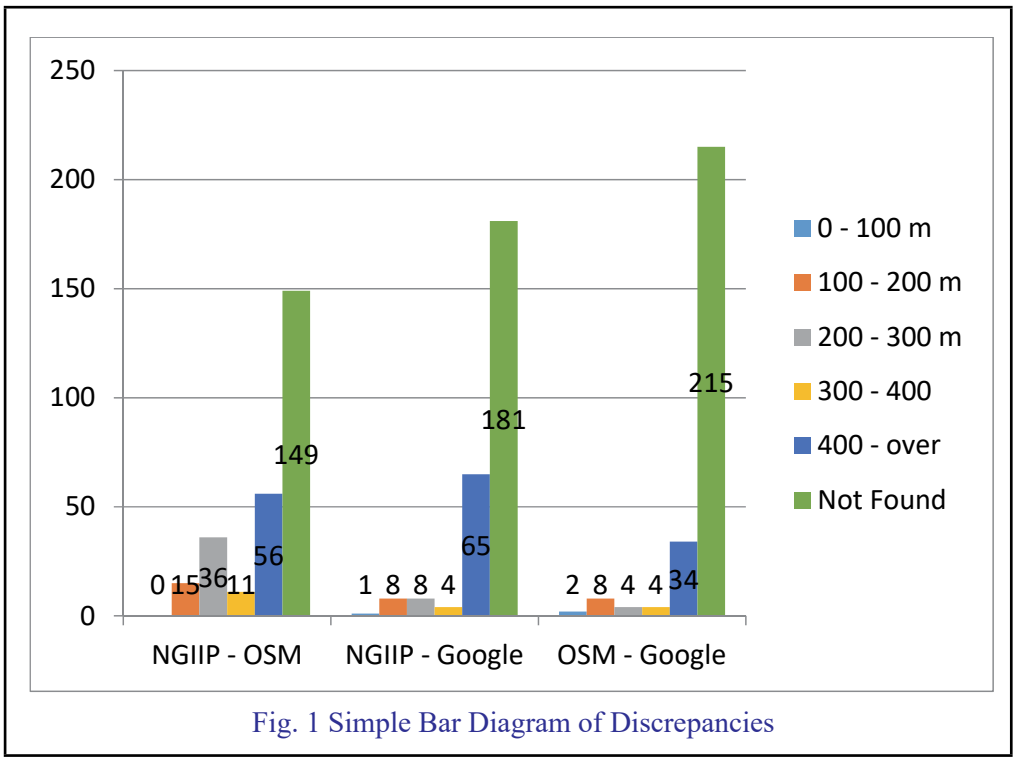

Fig. 1 Simple Bar Diagram of Discrepancies

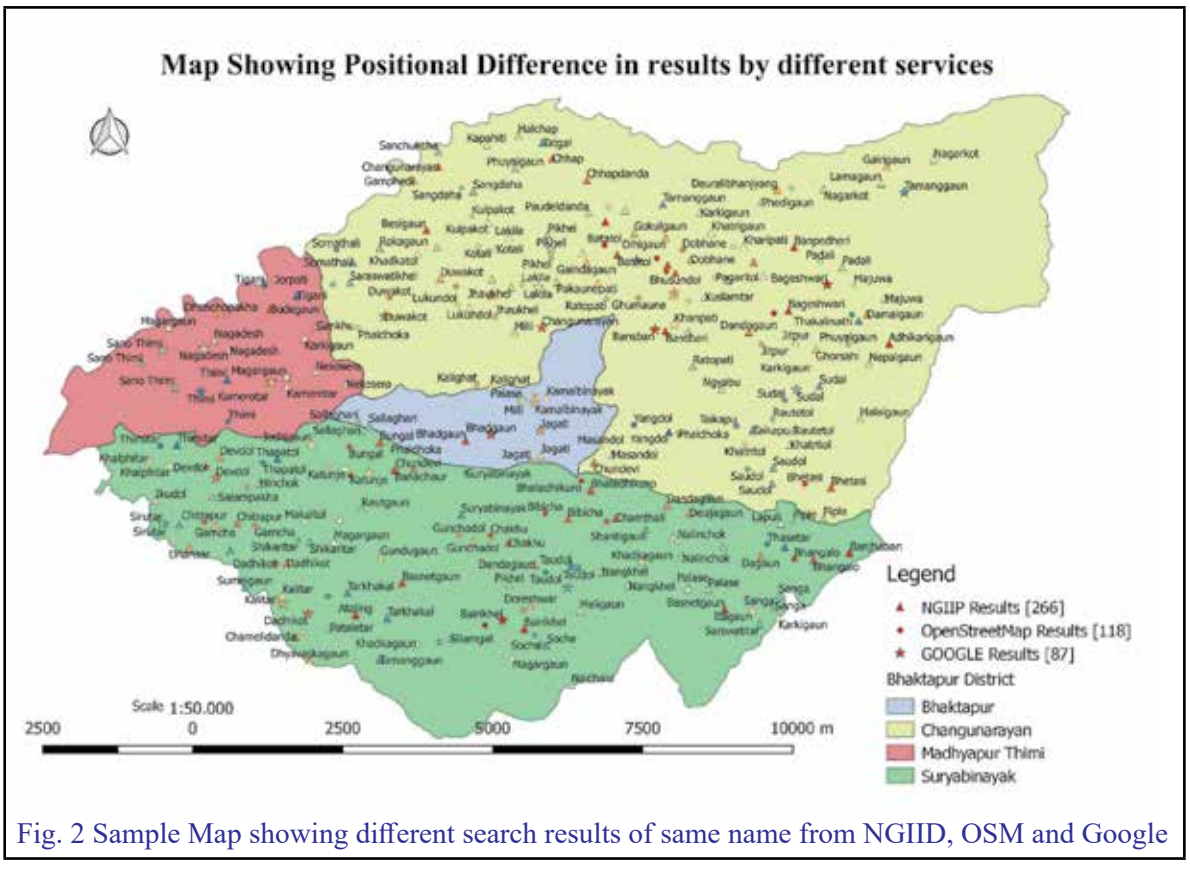

Fig. 2 Sample Map showing different search results of same name from NGIID, OSM and Google

Among the Results found in both OSM and Google, the discrepancy chart below shows there is very not a good match in the results given by both services. Maximum error by OSM is below 20 km whereas Maximum error by Google is around 1000 km. This is because, OSM results are hierarchical, i.e. it will give only results which match the hierarchy while google provides best match from all over the world. This means it will give results that will tend to match the search at least one of the word in the searched location.

5. CONCLUSION

There are many Geocoding services available but they all have one or more of the following limitations: (i) allows only geocoding one address at a time; (ii) requires the creation of a user account; or (iii) includes multi-page navigation before arriving at the geocoding interface. Nominatim is extremely user friendly and does not have these restrictions. Importantly, while a number of studies have evaluated the geocodes produced by Google, much less research has evaluated the geocodes produced by several of these alternative software packages. Since the accuracy of geocodes in part depends on the quality of the street reference maps used to generate the coordinates. The "true" geographic location of each address can be determined through aerial imagery or with global positioning systems (GPS) receiver data. Though these are gold standards, this was not practical nor a central focus of the study. In addition, Bing and Yahoo (the two companies that can be used to produce geocodes) maintain extensive geographic databases, which are frequently updated, ensuring strong address-matching capabilities and a sufficiently high positional accuracy. The street base map data used by the different geocoding services plays a large part in determining accurate address matches. The mapping companies Tele Atlas and NAVTEQ map and sell these base map data to companies like ESRI, Google and Yahoo, which then include them in their geocoding services. We do not know of any such professional companies in Nepal, but Both Google and Open street MapOSM use crowdsourcing to collect data. Open Street Map is a volunteered powered organization and Google map maker also collects data from crowd. Therefore, the base map data used by the different geocoding services at any given point may vary in quality and completeness. The quality and completeness may also vary by geographic region. Thus, it is important to also document (if possible) what base map data the geocoding service used. However, even if two geocoding services use the exact same base map data, different address-matching sensitivity settings built into the geocoder may produce different positional placements. Further, while error might be introduced due to incorrect geocodes (with correctly recorded addresses), error can also arise due to the quality of the collected addresses. For this reason, we manually cleaned the addresses for this study prior to geocoding. Although we geocoded the same addresses that had been cleaned, it is likely that the editing of the addresses impacted the geocoding findings. Additionally, we used interactive geocoding to investigate ties in order to yield the highest possible match rate and increase the positional accuracy. Our use of interactive matching is likely to have affected the geocodes included in this study. It is also important to note that, in addition to the settings used, different programs, or even different versions of the same geocoding software, might produce different results. Since each of the elements discussed can influence the results, we suggest that future projects take these aspects into consideration when geocoding and examining differences between geocoding methods. In conclusion, although this study indicates that positional differences between the two geocoding methods examined exist, the medians of the differences found with Google and Nominatim were minimal and most addresses were placed only a short distance apart. Although future research should compare the positional difference of Nominatim to criterion measures of longitude/latitude (e.g. with GPS measurement), we feel that Nominatim is a free and powerful alternative when geocoding addresses.

REFERENCES:

Bakshi, R., Knoblock,C. A.,& Thakkar, S.(2004). *Exploiting online sources to accurately geocode addresses.* In D.Pfoser, I. F. Cruz, and M. Ronthaler, eds., ACM-GIS '04:Proceedings of the 12th ACM International Symposium onAdvances in Geographic Information Systems, WashingtonD.C., November 2004, 194-203.

Bonner Han D., Nie, J., Rogerson, P., Vena, J.E., & Freudenheim, J.L. (2003). *Positional accuracy of geocoded addresses in epidemiologic research.Epidemiology* 2003, 14(4), 408-412.

Boscoe, F. P., Kielb,C. L., Schymura, M. J.,Bolani, T. M. (2002). *Assessing and improving census track completeness*. Journal of Registry Management, 29(4),117-20.

Davis Jr., C.A., &. Alencar, R.O. (2011). *Evaluation of the quality of an online Geocoding resource in the context of a large Brazilian city.* Transactions in GIS, 15(6): 851-868.

Dearwent, S.M., Jacobs, R.R., & Halbert, J.B. (2001). *Locational uncertainty in georeferencing public health data sets.* J Expo Anal Environ Epidemiol, 11(4):329-334.

Gatrell, A.C. (1989). *On the spatial representation and accuracy of address-based data in the United Kingdom.* Int. Journal of Geographical Information Systems, 3(4): 335-48.

Hay, G., Kypri, K., Whigham, P., &Langley, J., (2009). *Potential biases due to geocoding error in spatial analyses of official data.* Health Place, 15, 562-567.

Henley, A.C., & Heiss, G. (2006). *Accuracy of commercial geocoding: assessment and implications.* EpidemiolPerspectInnov,3, 8.

Lixin, Y. (1996). *Development and evaluation of a framework for assessing the efficiency and accuracy of street address geocoding strategies.* Ph.D. thesis, University at Albany, State University of New York, Rockefeller College of Public Affairs and Policy, 1996.

Ward, M H., Nuckols, J.R., Giglierano, J., Bonner, M.R., Wolter, C., Airola, M. Mix, W, Colt, J.S., & Hartge, P. (2005). *Positional accuracy of two methods of geocoding. Epidemiology,*16(4): 542-7.

Zhan, F.B., Brender, J.D., De Lima, I., Suarez, L., & Langlois, P.H. (2006). *Match rate and positional accuracy of two geocoding methods for epidemiologic research*. Ann Epidemiol,16, 842-849.

Author's Information

Name	:	Er. Amrit Karmacharya
Academic Qualification	:	B.E. in Geomatics Engineering
Organization	:	Survey Department
Current Designation	:	Survey Officer
Work Experience	:	5 years
Published Papers/Articles	:	1
Email ID	:	akarmacharya8@gmail.com