Address Geocoding Services in Geospatial-based Epidemiological Analysis: A Comparative Reliability for Domestic Disease Mapping

Monir, N..,[1] Abdul Rasam, A.R.,[1,2]* Ghazali, R.,[1] Suhandri, H. F.[3] and Cahyono, A.[4]
[1]Centre of Studies for Surveying Science and Geomatics, Faculty of Architecture, Planning and Surveying, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia
[2]Environmental and Social Health (ESH) Group, Health and Wellbeing Excellence Entities, Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia, E-mail: rauf@uitm.edu.my
[3]Faculty of Civil Engineering and Built Environment, Universiti Tun Hussein Onn Malaysia (UTHM), Johor, Malaysia
[4]Cartography and Remote Sensing Study Program, Department of Geographic Information Science, Faculty of Geography, Universitas Gadjah Mada, Yogyakarta, Indonesia
Corresponding Author

DOI: https://doi.org/10.52939/ijg.v17i5.2029

Abstract
Recently, geographical information system (GIS) has emerged as an important tool for many cases in epidemiology. Through the process, it begins with geocoding, i.e. assigning geographic coordinates to an address on a map. This process is a bridge between spatial information and its attribute data. Fortunately, some open geocoding services are available. The paper aims to examine the mapping reliability of some online geocoding services to map the spread of tuberculosis (TB) in Sarawak, Malaysia towards practical implementation in the domestic health department. The features examined the common platforms, namely QGIS, Google Map (API), and ArcGIS Online, were selected and explored in terms of the following variables; positional quality, speed, cost, and coverage. Based on our exploratory experiment, ArcGIS Online offers relevant mapping features for the local geocoding services of the TB locations compared to the other two platforms. But the chosen geocoding methods or services may depend on the nature of the project, cost restrictions, and the experience of an analyst. Comparison of the positional accuracy with manual reference methods (e.g GPS measurement and manual digitizing) could be further studied.

1. Introduction

Environmental epidemiology requires a reliable exposure assessment of both temporal and spatial components. In response to these challenges, epidemiological studies are increasingly using residential addresses of study participants and GIS to improve the characterization of environmental exposures and examine their association with human health risks for a large variety of disease conditions. Disease mapping is a visual representation of intricate geographic data that provides a quick overview of said information. Mainly used for health GIS application such as explanatory purposes, disease maps can be represented to survey high-risk areas and help policy and resource allocation in site areas (Krieger et al., 2003; Clarke et al., 1996 and Moore and Carpenter, 1999).

The process of geocoding and assigning a georeferenced location to the study subject's residential addresses is one of the first steps in GIS-based epidemiological studies such as disease mapping. The quality of geocoding depends on the completeness and the level of positional accuracy information of locating addresses. Completeness can be defined as the proportion of addresses that can be geocoded and depends on the quality of the collected data on addresses, while the positional accuracy reflects the level of proximity of geocoded objects to their true location (Goldberg, 2011). Many geocodes are available, whether free or openly online, but from many aspects does, not sure which one is most suitable to use and easy to access and accurate at the same time. The suitability is important for a healthy person to represent data of disease mapping and be familiar with access to the service that's available, whether free and open source to simplify their work and represent the data technically.

In the context of tuberculosis epidemiology, the suitability geocoding services need to be investigated in terms of positional quality, ethical use of address (detection), time and cost for better residential address of the cases that are related to effective surveillance and control of the disease. This is important to precisely understand the patterns of disease spread over space and time. Visualization of such data on maps enables health officials to obtain, analyse and understand real-time disease patterns compared to the tubular or report forms. Displaying locations on maps serves as an exploration of data and can lead to cluster analysis of disease incidences. Generating maps for different time periods is helpful for understanding the disease progression over time. Some health departments in developing countries still registered disease cases not in longitude and latitude information, in particular for early versions of data record management.

The latitude and longitude information of the cases can be generated since they comprise locational addresses that are essential for disease geocoding and mapping. Existing manual geocoding techniques need to be also converted to digital techniques to facilitate geocoding work, involving hundreds of disease cases. As a result, it is necessary to have an automatic batch geocoding software which could help to geocode addresses automatically (Faure et al., 2017). Finally, representation geocoded notified cases and registered cases on maps within a designated time. Furthermore, the geolocation-updated information assists healthcare officials to understand patterns of incidence and spread of tuberculosis. In Malaysia, local researchers have applied the geocoding in TB disease mapping (Mahsin et al., 2021; ; Abdul Jalil and Abdul Rasam, 2021, Abdul Rasam et al., 2020, Azewan and Abdul Rasam, 2020 and Abdul Rasam et al., 2016), but specific studies on suitable geocoding techniques or tools still need to be conducted to assist researchers to create disease maps properly, especially among healthcare communities.

Selecting a suitable geocoder perhaps is able to facilitate healthcare officials in a better platform and would allow in depth spatial health analysis. As an initial step to achieve this goal, the performance of selected online geocoding services on a dataset of address is examined. Previous studies have demonstrated that GIS can be used to correctly measure geographical health applications in a cost efficient manner (Fortney and Warren, 2000 Hurley et al., 2003 and Krieger, 2003). The knowledge of accurate geocoding to public health practitioners is essential in order to consider potential biases and limitations in disease mapping (Edwin Chow et al., 2016, Hurley et al., 2003, Moore and Carpenter, 1999 and Roongpiboonsopit and Karami 2010). Roongpiboonsopit and Karami (2010) added any errors connected with the geocoded addresses will be disseminated to the future decisions, activities, modelling, and analysis. The result reliability of these online software platform are is also verified since is developed by specialized teams and there is no unauthenticated modification (Sood and Soni, 2016).

Therefore, the study was performed to explore the common and available free and open online address geocoding using a suitable method for geocoding of subject residences. This study consists of three objectives; i. to examine the existing platform of freely online geocoding services for TB disease mapping, ii. to analyse the geocoding capabilities of selected freely online platform for disease mapping of TB and iii. to map the TB distribution in the study area using selected freely online geocoding platform.

2. Methodology

This study used a methodology that consists of several phases, including planning, data collection, data processing and analysis as shown in a diagram at Figure 1. These phases are parallel with the concept and framework as explained in a GIS function, where converting text-based postal address data into digital geographical coordinates. ArcGIS Online, Google Map and QGIS were selected because they offered free and open services and were commonly or familiarly easy to process the spatial dataset. ArcGIS Online uses organization type under GeoUiTM Portal. QGIS is version 3.10.1 for the Windows environment. Lastly, Google Maps that is mapped with Google, My Maps using google account.

2.1 Phase of Planning and Study

The findings were carried out to examine the capabilities and availability of the selected free online address geocoding services. During this phase, each of the free online address geocoding services was explored and reviewed by doing research on the internet, journaling and exploring its capabilities for mapping. This method is one of the efficient ways to start this planning phase and can decide the most relevant free online address geocoding service for disease mapping. Three websites are used to find the top three common software for geocoding (ArcGIS Online, QGIS and Google Maps) of TB cases in the state.

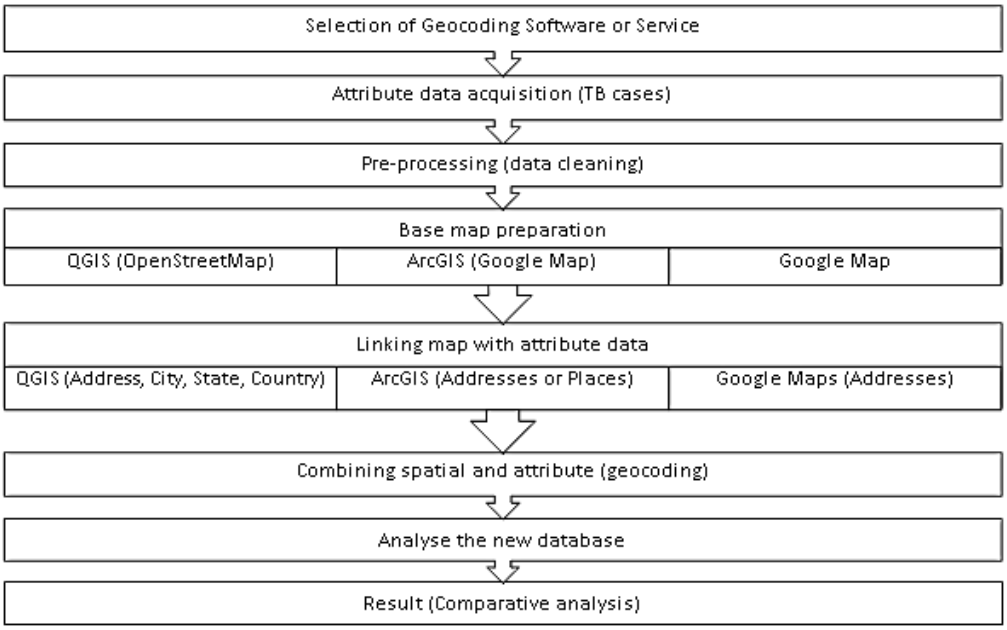

Figure 1: Research Methodology

The selection for free online address geocoding service for disease mapping is very fundamental in making this study meet the aim and results. When selecting the best free online address geocoding, the requirement of the service that involves the element of mapping for disease needs to be considered in order to ensure this project fulfils the requirement. All the software that has been found is licensed either as free service or free and open source service. Three software have been selected and used (ArcGIS Online, QGIS and Google Maps). ArcGIS Online (https://www.esri.com) is selected as the common online address geocoding service for disease mapping. ArcGIS Online is a cloud-based mapping, analysis and data storage system hosted by Esri that can be used to create, share and manage maps, scenes, layers, apps and other geographic content. ArcGIS Online is a global platform and in order to provide base maps that are useful to a global community of users the default projection has to work globally. ArcGIS Online uses the WGS 1984 Web Mercator (Auxiliary Sphere) coordinate system as the default Projected Coordinate System (PCS).

QGIS is selected as the second common free online geocoding service for disease mapping. QGIS is an open source software and it is easy to access as compared to commercial software ArcGIS. Although a continuous development of plugins in QGIS, at present time it is not as much developed as ArcGIS is. QGIS has less processing time and better rendering capabilities. Google Map Geocoding is the third common free online geocoding that is selected is Google Map Geocoding. Google Maps has become everyone's favourite source for navigation, traffic, and transit and location information. Behind the scenes, it is because of Google Maps' rich geocoded database containing millions of data points. And it is as simple as typing a place in the address bar and Google Maps will take you to the location.

2.2 Data Collection and Software Selection

The data that were used is tuberculosis cases in Sarawak in 2018. For this case study, the data used were limited only to the Kuching area that involved 30 cases from the overall of the cases. Each address element was entered into separate fields of Microsoft Excel file. In Figure 2, the field is divided into several columns that contain the country (Negara Asal), state (Negeri), city (Bahagian/Kawasan/Daerah) and address (Alamat Kediaman (Seperti Dalam CDCIS)). The file was saved as CSV (Comma delimited).

2.3 Data Processing

The data processing is into two consecutive steps, namely address data cleaning and address geocoding. The previous step, address data cleaning, intends to improve standardization and quality of geocoding. The address of subjects was verified manually for the spelling of street and district names. Address fields of subjects were also completed such as missing or incomplete postal code, district name, street name and street number.

	A	B	C	D	E
	Negara Asal	Negeri	Bahagian/Kawa	Daerah	Alamat Kediaman (Seperti Dalam CDCIS)
	MALAYSIA	SARAWAK	KUCHING	KUCHING	
	MALAYSIA	SARAWAK	KUCHING	KUCHING	
	MALAYSIA	SARAWAK	KUCHING	KUCHING	
	MALAYSIA	SARAWAK	KUCHING	KUCHING	
	MALAYSIA	SARAWAK	KUCHING	KUCHING	
	MALAYSIA	SARAWAK	KUCHING	KUCHING	
	INDONESIA	SARAWAK	KUCHING	KUCHING	

Figure 2: Address data (*Specific patient addresses are null due to ethical data protection)

Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi 1 \cdot \cos\varphi 2 \cdot \sin^2(\Delta\lambda/2)$$

$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$

$$d = R \cdot c$$

where φ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km)

Figure 3: Haversine Formula

Assistance from Google map was used to find the incomplete address data. The later step, address geocoding, is based on a linear interpolation of an address within the address range for the street segment in a reference street file. There were 30 sample addresses in the database system geocoded after the cleaning step was passed. This step undergoes trial-and-error process to obtain the expected solid result. The purpose of this repeating process is to obtain reliable results for the comparative evaluation of the domestic TB mapping using ArcGIS Online, Google Map and QGIS.

2.4 Mapping and Analysis

This phase is the process to evaluate the capabilities of the selected free and open software. The selected software is evaluated based on the following aspects: coverage (detection), quality (accuracy), speed (time) and cost. For the spatial coverage or detection, the cleaned data were searched and matched using the selected geocoding. The match rate for each geocoding service is compared and analysed. The high rate of match is considered as more closely reliable or practical geocoding than the other platforms. The locations of the dot points (cases) on the area were also confirmed whether they are placed on the study area or not using the base map provided in the respective platform. The base map data used by different geocoding services also to find the quality and the completeness.

The other aspect of the evaluation is quality or position. Accuracy is used to describe the closeness of a measurement to true value. The results on match rate and similarity. Match rate is the proportion of input addresses that retrieves a geocode from the geocoding system. This similarity is defined as the distance measure between geocodes from two services. For example, if the distance between each geocode from Google Maps and QGIS is greater than the distance between geocodes from Google Maps and ArcGIS Online, then can be concluded that the geocode from Google Maps has more similarity to the geocode from ArcGIS than that of QGIS. The similarity evaluation was only focussed on the addresses that had a matching geocode across all three services.

After searching and matching processes through the geocoding services. The distance between the locations provided by the two different services were compared using Haversine formula (Figure 3). Haversine formula needs distance between two sets of coordinates, which are in latitude and longitude format, and gives output (distance) in the metric system. In other words, the shortest distance on the earth's surface is performed by computing the great-circle distance between two points. It takes into account the distortion due to the curvature of earth and different scale factors at different latitude values. Movable Type Scripts are used to calculate distance between latitude and longitude. The page presents a variety of calculations for latitude/longitude points, as well as the formulas and code fragments. The computation process in this step is performed on the isometric surface, *i.e.* under an assumption that the globe is a perfect sphere. Although the earth geometric model in common practical geodetic application is approached by

ellipsoidal, the spherical model deviates only by 0.3% from the actual reference ellipsoid.

Speed in time is related to the operating performance of a geocoding system and defines characteristics of the geocoding system that affect how fast records can be processed. In most modern computing environments in use today, per-record processing speed is of little concern as many commercially available geocoding systems can process on the order of millions of records per hour. However, if large volumes of data must be continually processed or re-processed, speed may be an issue that can be used to discriminate between geocoding systems. An extreme example would be a need for real-time geocoding in health emergency scenarios such as disease outbreaks. The geocode data needed immediately to help resolve or understand a phenomenon as it is unfolding on the ground to assist in the decision-making process, determine where resources are needed and identify a course of action to pursue to save lives. The last aspect is cost, in which the true cost of a geocoding system can be a difficult thing to quantify. However, some aspects of the geocoding system cost are easy to quantify. The prices for a software license for the geocoding system, the price of a license for the required reference data layer, and the price for a support contract are examples of one-time (or yearly) fixed costs that can readily be obtained from a software vendor or assumed to be zero for open source-software.

3. Results and Discussion

3.1 Existing Online Geocoding Services for TB Disease Mapping

There are several geocoding services available both commercial-license and free-license software. In this contribution, three geocoder or address matchers from ten different developers were studied based on their positional quality, detection, time and cost factors. Based on the main website sources (e.g https://gisgeography.com/geocoders/,https://www.programmableweb.com/news/7-free-geocoding-apis-google-bing-yahoo-and-mapquest/2012/06/21, https://rapidapi.com/collection/geocoding-location-apis) describe that seven free geocoding API and compare the geocoder on features, speed and limits. For geocoder timing, each API is checked every ten minutes for every week. The address used to be the same for all decoders. Google Geocoding API and Cloud made Geocoding API were the only two with 100% uptime. In terms of speed, the fastest two were Bings Map Geocode and Google Geocoding. The other website explains that there are seven geocoding and reverse geocoding services for pinpointing addresses. Next website reference describes 36 top geocoding and location API. The geocoding is evaluated based on the timeliness, positional quality and detection.

The following list (Table 1) presents some geocoders or address matchers used in many GIS applications, including Google Map, QGIS, Here, Esri, PBBI's Geocoding (Phitney Bowes), US Census, and Bing Location API. According to the review, all platforms are almost the same, but the difference is the description of evaluation for each geocoder. From the reviews done, most of the website selected three main geocoder services which are QGIS, ArcGIS (Esri) and Google Map Geocoding. This three selected online geocoding service is explored and evaluated based on criteria of positional quality, timeliness, detection and cost.

Table 1: Summarized list of the geocoding service platforms

Service	License	Description (Geocoding capacity)
ArcGIS Online	Free and Open Source	1,000,000/month Location-weighting unknown
Quantum QGIS	Free	Google Map only 2,500/day OpenStreetMap has no limits
Google Maps	Free	2,500/day
HERE Maps Geocoding	Free	10,000 per day
US Census Geocoder		Only for United State area
Bing Location API	Free	125,000/year
MapQuest Geocoding API	Free	15,000/month Location weighting Test area
CloudMade Geocoding	Free	100,000/month
Pitney Bowes Geocoding	Free	Storing geocodes 30 days Location-weighting unknown
OpenAddresses GeoLocated	Free	5000/day

Table 2. Address matched

Software / Services	Addresses Matched	Match Rate
ArcGIS Online	30/30	100 %
Google Maps	29/30	97 %
QGIS	16/30	53 %

Table 3. The distances of two difference geocode

Address	Distance (km)		
	QGIS to ArcGIS Online	ArcGIS Online to Google Maps	QGIS to Google Maps
Address 1	0.320	0.151	0.177
Address 2	0.062	0.772	0.138
Address 3	0.415	0.004	0.416
Address 5	0.413	0.241	0.193
Address 6	0.271	5.362	5.126
Address 7	0.157	0.014	0.144
Address 8	1.653	0.943	1.036
Address 9	0.135	2.170	2.155
Address 11	1.034	8.133	7.408
Address 12	0.699	0.066	0.764
Address 13	0.209	0.050	0.255
Address 14	4.449	0.074	4.473
Address 15	6.127	1.383	7.249
Address 16	0.199	0.062	0.255
Address 18	7.284	0.029	7.256

3.2 Comparative Analysis of Selected Online Geocoding Capabilities for TB Disease Mapping

The capability of the selected free and open platform (ArcGIS Online, Google Map and QGIS) is evaluated in this part. The selected platform were evaluated based on coverage (detection), quality (position), speed (time) and cost.

3.2.1 Coverage (Detection)

Match rate was at its lowest level (53%) when using QGIS geocoding service (Table 2). The other two geocoding services retrieved most of the geocodes (97% - 100%). After checking one by one, by using ArcGIS Online, all 30 address dot points are still in the area (Kuching, Sarawak). Using Google Map the 29 dot points are still in the area. While using QGIS only 16 dot points were still in the area. In terms of coverage, Google Maps, and ArcGIS Online (using Google map has higher coverage. OpenStreetMap (OSM) in QGIS, on the other hand, has an average coverage. For instance, many important places such as hospitals, government buildings, parks and others will be missing from the map and an operator will have an additional task to include and edit the missing places. Google map is very detailed in its coverage down to the smallest streets and shops. In website categories, Google map is also ahead of OpenStreetMap in many categories including business and more than 200 other categories. Google and OpenStreetMap (in QGIS) use crowdsourcing to collect data. OpenSteetMap is a volunteer power organization and Google map maker also collects data from the crowd. Therefore, the base map data used by different geocoding services at any point may vary in quality and completeness. Thus, it is important to also document what data the geocoding service used. Different address-matching sensitivity settings built into the geocoder may produce different positional placements. Manually cleaned the addresses for this study prior to geocoding. Although geocoded the same addresses that had been cleaned, it likely impacted the geocoding finding. For example, in QGIS, the only detectable matched addresses are 7 out of 30 before data cleaning, however the geocoding has improved to 16 matched addresses after data cleaning. This revealed that a procedure of manual standardization was performed in order to enhance the quality of the results provide a more precise geographic representation of health related events (Baldovin et al., 2015)

3.2.2 Positional quality (Quality of reference maps)

For this exploration, only 15 addresses were geocoded with all the geocoding services (ArcGIS, QGIS and Google Maps) for quality evaluation of the cases positions/distances. The distances between the locations provided by the three differences services (Table 3) were compared based on the point of location. The average distance between a pair of geocodes from different services is larger when comparing geocodes from Google Maps and QGIS

(Table 4). In other words, the lowest similarity was found between geocodes from Google Maps and QGIS. The quality of geocodes in part depends on the quality of the street reference maps used to generate the coordinates. The smaller gap of distance between ArcGIS online and Google Maps, showing that these geocoder comprise a better quality of base/reference map for the cases placements. The actual geographic location of each address can be also determined through a global positioning system (GPS). The further study can be considered the standard of accuracy especially for disease mapping.

3.2.3 Speed (Time)

Three-time processing attempts for each service were recorded for each service to find the average of the time required to do the address-matching process. The result showed that the fastest time for geocoding processing is ArcGIS Online, followed by Google Maps and QGIS (Table 5). For example, ArcGS Online has taken less than 5 seconds to complete the 30 addresses location-matching of the cases, whereas Google Maps and QGIS required more than 30 seconds to carry the same processes. This study has suggested that the process of location-matching is strongly affected by the networking, and it is therefore in maintaining the consistency of data, processing is done at the same speed of the network and same the desktop or laptop.

3.2.4 Cost

There are many geocoding services available with specific characteristics, in particular cost factors. In QGIS, the geocoding process is quite flexible because the web service (OpenStreetMap) used is free and there is no limit to the use of space compared to the use of Google which requires cost and space due to the API key license. Open Street Map or Nominatim used in QGIS is user friendly and does not have these restrictions. ArcGIS Online is available for personal use through a free public account or organization-wide through an annual subscription. The license costs $100 for a 12-month period. This annual fee allows for the home use or personal use of the ArcGIS suite of GIS software along with its most popular extensions. Though Google Map is free, there are certain charges incurred when one makes use of Google mapping services. There is the cost of privacy in addition to not being able to control whatever is displayed on the map. OpenStreetMap on the other hand, always will be free to users, developers and companies. Nevertheless, According to Swift et al., (2008) the study findings indicate that price alone is not a reliable indicator of geocode correctness. Centrus, the most expensive geocoder tested, produced USCB misclassifications on various input addresses, all of which were handled correctly by the less expensive commercial versions (ESRI) as well as the free internet web services.

3.3 Geocoding Services for TB Disease Mapping

The ArcGIS has seen much use in spatial analytics and modelling in different perspectives and is one of the most advanced and reliable geospatial analytical tools available (Table 1) as illustrated in Figure 4. However, QGIS, an open-source GIS tool, has become very popular in the field of geospatial analytics. In this study, QGIS is used to plot geocoded locations on a map using QGIS version 3.10.1 for the Windows environment. Within the QGIS environment, the open layers plugin provides options for selecting Google Maps and OSM as base maps on which to plot geocoded locations of the cases (Figure 5). These locations were plotted on OSM only. Plugin in QGIS in an effort to balance the capabilities of ArcGIS, especially in offering better rendering capabilities. However this platform still lacks in terms of optimal geocoding processing (Figure 6). QGIS is an open source software and it is easy to access as compared to commercial software ArcGIS, Google Maps has a street view option.

Table 4. Average distances (km) of geocodes between geocoding services

	ArcGIS Online	QGIS	Google Maps
ArcGIS Online	-	-	-
QGIS	1.562	-	-
Google Maps	1.297	2.469	-

Table 5. Time processing for 30 addresses location-matching

Software / Services	**Time**	**Comment**
ArcGIS Online	Less than 5 seconds	Fast
Google Maps	Less than 30 seconds	Medium
QGIS	More than 1 minute	Medium

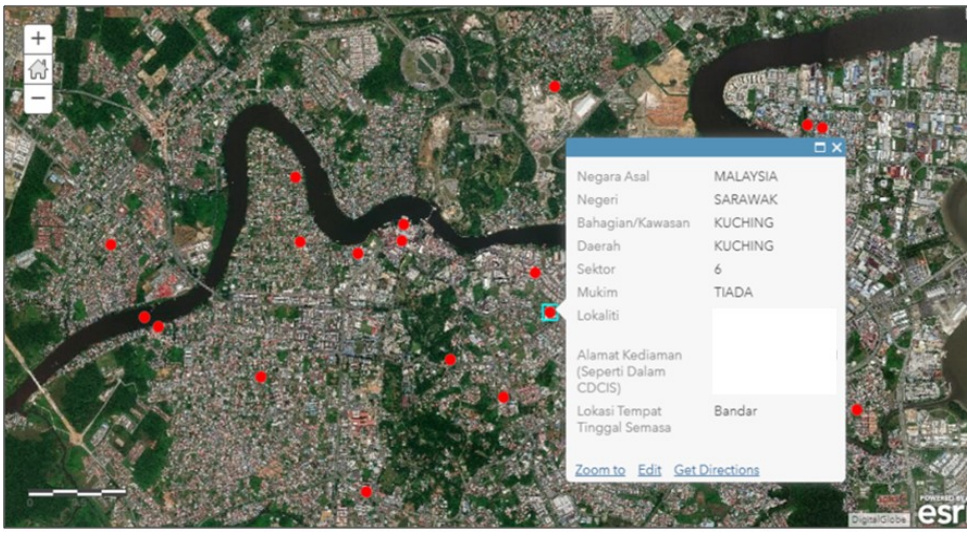

Figure 4: Geocoding of TB using ArcGIS Online

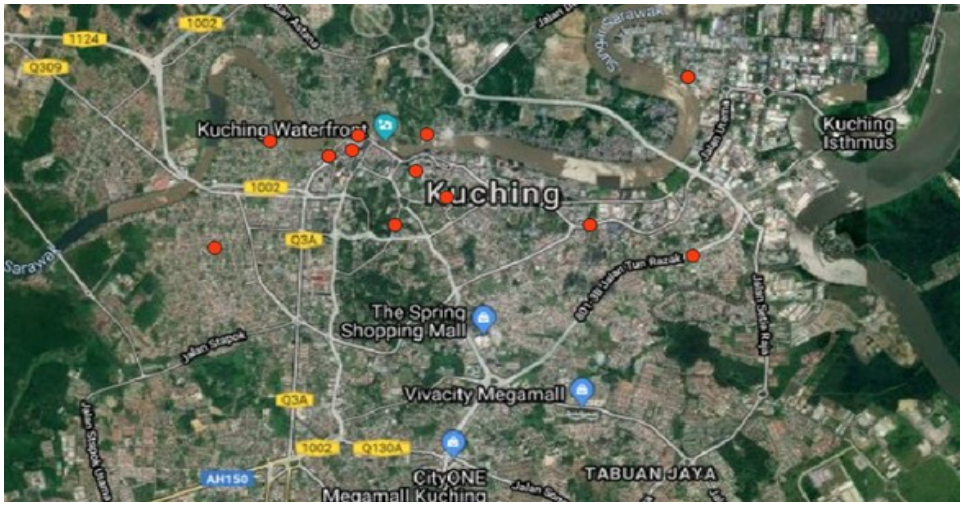

Figure 5: Geocoding of TB using QGIS

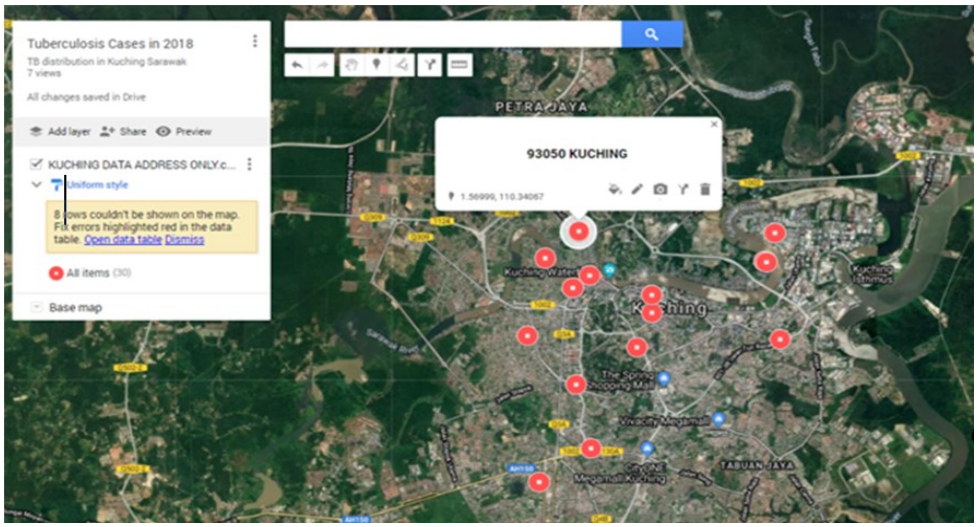

Figure 6: Geocoding of TB using Google Maps

Table 6. Summarized result of the geocoding processes

Software	Free or license	Programming Requirement	Online or locally installed	Relative output quality, speed and coverage	Key Benefits
ArcGIS Online with ESRI World Geocoding Service	Unlimited via site license	No	Online	High Fast Global	Easy to use once figure out the workflow but limited metadata output with results. ~30 in less than 10 seconds
QGIS (OpenStreetMap Geocoding API)	Free	Yes, MMQGIS plug-in	Installed	Medium Moderately fast Global but spotty	Free, no API key needed. ~30 in 1 minutes
Google Maps	Flexible price	No	Online	High Moderately fast Global	Flexible price ~30 in less than 30 seconds

**mileage may vary – results based on limited testing and local hardware, software and network configuration*

Prior to conducting GIS and spatial analyses, health researchers and practitioners often need to geocode address-based data and many do so "in-house". However, with more user-friendly, geocoding services available, the more user-friendly, geocoding services available, the important decisions regarding geocoding sensitivity may be hidden from the user (which might not be identifiable to a non-specialist) and the geocodes obtained may be inaccurate (which can lead to substantial exposure misclassification). The study of free software for geocoding requires a lot of time and decision. In order to find the suitable software or service, the selection must fulfil requirements especially in terms of pricing and other relevant aspects. In the context of this study, the overall result of the respective geocoding processes is shown in Table 6. Although Swift et al. (2008) indicated that no one geocoder stands out as above and beyond the others, with each having their strong and weak points, in this study found there were slight advantages to ArcGIS Online in terms of address matching and time completeness.

In regard to review on software or service available in the market. A finding was conducted to identify the free and open online service for geocoding. Therefore, there about a hundred software was generally reviewed by study based on the main website references. Then 10 software were specifically reviewed to find the three suitable and common services for geocoding that have capabilities to fulfil the requirement such as cost required and others. This is important for healthcare officials to analyse data of TB in an easier way. All the reviewed software is only licensed as free software or and open source software. One of the most important parts for this study is address matching or geocoding using three selected services (or software). Selection of the Suitable Software or Service The findings indicate that ArcGIS Online is the best way for mapping based on the coverage (detection), quality (accuracy), speed (time) and cost. However, it is important to note that the chosen geocoding method may depend on the nature of the project, cost restrictions and the skills of the analyst.

Quantum GIS uses the Google API to geocode addresses, but only allows for entry of one address at a time unless a custom programme is allowed for more. Although many of free geocoder and allow many addresses to be geocoded, all have one or more of the following limitations:

a) Allows only geocoding one address at a time
b) Requires the creation of a user account
c) Includes multi page navigation before arriving at the geocoding interface

ArcGIS, Quantum GIS and Google Maps are user friendly. Although I do not know the actual locations of each address, I am confident that the geocodes produced by this service are generally positional accurate. Since the accuracy of geocodes in part depends on the quality of the street reference maps used to generate the coordinates, most up-to-date maps are used like from the ArcGIS Online World Geocoding Service which uses the most recent commercial street data. This study should include geocoded addresses with the highest positional accuracy. The true geographic location of each address can be determined through aerial imagery or with global positioning systems (GPS)

receiver data. However these are superior standards, this was not practical nor a central focus of the study.

The base map data used by the different geocoding services play a large part in determining accurate address matches. 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 vary by geographic region. Thus it is important to also document what base map data the geocoding service used.

However, even if 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 in other words could also be due to incorrect addresses such as incorrectly spelled street names). For this reason, the addresses for this study are manually cleaned prior to geocoding. Although geocoded the same addresses that had been cleaned, it is likely that the editing of addresses impacted the geocoding findings for example improved the match rate and probably also increased the positional accuracy.

4. Conclusion and Future Direction

Geocoding is useful in public health tracking because it tells us where disease, specific demographics, environmental issues, or other health-related factors are concentrated geographically. Several health departments still registered disease cases in manual systems, especially old cases recorded without longitude and latitude information. Thus, it is necessary to have online batch geocoding software which could assist to geocode addresses automatically. For this disease mapping of tuberculosis (TB) study, three common platforms utilized for the domestic applications, namely QGIS, Google Map, and ArcGIS Online, were explored in terms of positional quality, time of speed processing, cost and spatial coverage. Experimentally, these platforms have special features, but ArcGIS Online has a slight advantage based on the aspects evaluated, especially in coverage and time. Although this study indicated that positional differences between the three geocoding methods examined exist, the differences found with ArcGIS were minimal and most addresses were placed only a short distance apart. The selected geocoding services are a free/open and powerful alternative when geocoding addresses, a much relevant task for healthcare officials or health researchers and practitioners with limited experience in this field. The knowledge of accurate geocoding to public health practitioners is essential in order to consider potential biases and limitations in disease mapping that are important for the future decisions and analysis. This exploratory result could be also utilised as a rough guideline for finding a service that meets the needs of an application that may depend on the nature of the project, cost restrictions and the skills of the analyst. However, future research should compare the positional difference of services to criterion measures of longitude/latitude using manual reference methods (GPS measurement or manual digitization) for better accuracy evaluation. In addition, the study can further apply statistical evaluation methods, covering a big area with the number of the cases/addresses, and getting perspective from software experts on the platform capabilities.

Acknowledgements

The authors gratefully acknowledge the assistance of the Ministry of Higher Education (MOHE) and Universiti Teknologi MARA Selangor for providing Fundamental Research Grant Scheme (FRGS) 600-IRMI/FRGS 5/3 (093/2019). The research is registered in the National Medical Research Register, Malaysia (ID: NMR R -15-2499-24207). The authors are also thankful to the Ministry of Health Malaysia for providing TB datasets used in this study.

References

Abdul Jalil, I. and Abdul Rasam, A. R., 2021, Social Network Analysis of Spatial Human Mobility Behaviour in Infectious Disease Interaction: An Exploratory Evidence of Tuberculosis in Malaysia, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci,* XLIII-B4-2021, 55–61, https://doi.org/10.5194/isprs-archives-XLIII-B4-2021-55-2021

Abdul Rasam, A. R., Shariff, N. M. and Dony, J. F., 2016, Identifying High-Risk Populations of Tuberculosis Using Environmental Factors and GIS Based Multi-Criteria Decision Making Method. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Science*, Vol. XLII-4/W1, 9-13. https://doi.org/10.5194/isprs-archives-XLII-4-W1-9-2016, 2016

Abdul Rasam, A. R., Mohd Shariff, N., Dony, J. and Ling, O. H. L., 2020, Local Spatial Knowledge for Eliciting Risk Factors and Disease Mapping of Tuberculosis Epidemics. *Environment-Behaviour Proceedings Journal*, Vol. 5(SI2), 45-51. https://doi.org/10.21834/ebpj.v5iSI2.2522

Azewan, M. D. H. and Abdul Rasam, A. R., 2020, *Disease Mapping and Health Analysis Using Free and Open Source Software for Geospatial (FOSS4G): An Exploratory Qualitative Study of Tuberculosis*. In: Alias N., Yusof R. (eds) Charting the Sustainable Future of ASEAN in Science and Technology (Springer, Singapore). https://doi.org/10.1007/978-981-15-3434-8_43

Baldovin, T., Zangrando, D., Casale, P., Ferrarese, F., Bertoncello, C., Buja, A., Marcolongo, A. and Baldo, V., 2015, Geocoding Health Data with Geographic Information Systems: A Pilot Study in Northeast Italy for Developing a Standardized Data-Acquiring format. *Journal of Preventive Medicine And Hygiene*. Vol. 56(2), E88–E94.

Edwin Chow, T., Dede-Bamfo, N. and Dahal, K. R., 2016, Geographic Disparity of Positional Errors and Matching Rate of Residential Addresses among Geocoding Solutions. *Annals of GIS*, Vol. 22(1), 29-42. DOI: 10.1080/19475683.2015.1085437.

Faure, E., Danjou, A. M. and Clavel-Chapelon, F., 2017, Accuracy of Two Geocoding Methods for Geographic Information System-Based Exposure Assessment in Epidemiological Studies. *Environ Health*, Vol. 16, 15. https://doi.org/10.1186/s12940-017-0217-5.

Fortney, J., Rost, K. and Warren, J., 2000, Comparing Alternative Methods Measuring Geographic Access to Health Services. *Health Services & Outcomes Research Methodology*, Vol. 1, 173-84.

Goldberg, D. W, 2011, Advances in Geocoding Research and Practice. *Transactions in GIS*, Vol.15, 727-733.

Hurley, S. E., Saunders, T. M., Nivas, R., Hertz, A. and Reynold, P., 2003, Post Office Box Addresses. A Challenge for Geographic Information System-Based Studies. *Epidemiology*, Vol. 14, 386-391.

Krieger, N., 2003, Place, Space, and Health: GIS and Epidemiology. *Epidemiology*, Vol. 14, 384-385.

Krieger, N., Waterman, P. D., Chen, J. T, Soobader, M. J. and Subramanian, S. V., 2003b, Monitoring Socioeconomic Inequalities in Sexually Transmitted Infection, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures–The Public Health Disparities Geocoding Project (US). *Public Health Reports*, Vol. 118, 240-260.

Mahsin, W.H., Rasam, A.R, Saraf, N.M. and Khalid, N., 2021, Free and open GIS source software for spatial epidemiology and geospatial health in Malaysia: a comparative analysis of the software usability. *International Journal of Advanced Technology and Engineering Exploration*, Vol. 8(78), 584-599. http://dx.doi.org/10.19101/IJATEE.2020.762192

Moore, D. A. and Carpenter, T. E., 1999, Spatial Analytical Methods and Geographic Information Systems: Use in Health Research and Epidemiology. *Epidemiologic Reviews*, Vol. 21, 143-61.

Roongpiboonsopit, D. and Karami, H. A., 2010, Comparative Evaluation and Analysis of Online Geocoding Services. *International Journal of Geographical Information Science*, Vol. 24, 1081-1100.

Sood, G., Shipra and Soni, R., 2016, Comparative Study: Proprietary Software vs. Open Source Software. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4(11), 19032-19038. [Available Online At https://doi.org/10.15-680/IJIRCCE.2016].

Swift, J. N., Goldberg, D. W. and Wilson, J. P., 2008, Geocoding Best Practices: Review of Eight Commonly Used Geocoding Systems. Los Angeles, CA, University of Southern California GIS Research Laboratory Technical Report No 10.