Engineering and Applied Science Research April – June 2019;46(2):86-97 Research Article

EASR

Engineering and Applied Science Research

https://www.tci-thaijo.org/index.php/easr/index

Published by the Faculty of Engineering, Khon Kaen University, Thailand

Quality analysis of online geocoding services for Thai text addresses

Dueanpen Manoruang and Duangduen Asavasuthirakul*

Department of Computer Science and Information Technology, Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand

Received 17 August 2018
Revised 1 January 2019
Accepted 8 January 2019

Abstract

A number of online geocoding services are now available enabling fast access to map-based geolocation. However, the quality of these services is uncertain, often being based on poor data, especially in developing countries such as Thailand. This paper reports on a comparative analysis of the quality of five such online geocoding services, with tests based on text addresses and points of interest (POIs) in Thailand. The geocoding service providers included in our tests were Google, MapQuest, Bing, Yahoo!, and OpenCage and the text inputs were in Thai. The quality of the geocoded results was measured using the match rates and the positional accuracy. Two experiments were conducted, each with a different input format: (i) text addresses collected from research participants (N = 1,511), and (ii) names of POIs sampled from a dataset of Thai academic institutes (N = 5,000). The quality of the services tested was compared statistically using the Friedman test and the Wilcoxon signed rank test. The results show that Google outperformed all other services for both text addresses and POIs. Google, Bing, Yahoo!, and OpenCage each had match rates over 90%, while MapQuest's match rate was 82%, but the positional accuracy of most services did not reach a high standard at rooftop levels. From this analysis, we identify geocoding issues that need to be addressed for further enhancement of the quality of the geocoding of addresses in Thailand. The knowledge obtained here also provides valuable insight into the geocoding issues facing Thailand and other developing countries, and it is hoped that this will benefit further research and the future development of high-quality geocoding tools.

Keywords: Geocoding, Online service, Quality, Match rate, Positional accuracy

1. Introduction

Information on addresses is commonly stored in databases and this may include information relating to personal data, landmarks, and points of interest (POIs). Typically, each address will reference only one particular geographic location or region. The process of converting text addresses into geographic coordinates is called geocoding [1]. As a result of rapid advances in technology, users can geocode text quickly and free of charge using online services, examples of which include Google Maps, Bing Maps, Yahoo! Maps, OpenCage and MapQuest. However, even though these online geocoding services make the process of conventional geocoding easy and relatively seamless, uncertainties associated with reference databases, geocoding algorithms, and positional errors remain [2]. Given these problems, processing geocoded coordinates of unknown quality leads to questions about the reliability of the process.

One problem preventing generalization of the geocoding process is the variety of addressing systems used globally. Addresses in most western countries, i.e., the United States, Europe [3] and Australia [4], are based on street networks and as such, the general address structure is comprised of a street number, street name, city, province or state, postal code and country. Within a city block, the numbering of buildings in western countries typically increases sequentially with the numbers of the buildings, rising as one travels along the street, with odd numbers on one side and even numbers on the other. This makes it possible to interpolate the distance between buildings with different addresses on the same street.

However, while in some east Asian countries, such as Japan, addresses are similarly based on blocks [3], with each block separated by streets, house numbers may be assigned arbitrarily within a block and it will therefore not be possible to extrapolate any distance related information from a particular address or set of addresses. Moreover, addresses may also lack a street name. Rather, house numbering systems may be based on construction dates. As a result, it can be very difficult to find a specific location based on its address. Finding a particular location within a block-based addressing system requires matching the block address and then returning the centroid of the block as the target location address, but this method may be inaccurate for larger blocks. The addressing systems used in India and in other eastern Asian countries, including China, Japan and Korea, are formatted to run from the largest administrative unit to the

*Corresponding author. Tel.: +66 5596 3262
Email address: duangduenr@nu.ac.th
doi: 10.14456/easr.2019.11

smallest, while in Korea, the addressing system is based on street networks [5-6]. In India, the addressing system is also based on street networks, but addresses are distributed irregularly along streets and the geographical locations cannot be identified systematically. Additionally, the structure of Indian addresses differs in rural and urban regions. Rural regions are characterized by districts within states, sub-district within districts, and villages within sub-districts, while urban regions are characterized by cities within states, localities within cities, and this is then followed by other information [7-8]. Beyond this, address string formatting in Asian countries is further complicated by factors such as differences in language and alphabets, and sometimes the lack of delimiters between words. Thus, because Asian countries may not use the street address format preferred in western countries, the range of house numbers on either side of a street segment may not be sequential and the address geocoding process will then not comply with TIGER/LINE or G-NAF standards. Street address formatting requires a street segment database for locating addresses and this may not work if addresses are distributed irregularly along streets, as they are in India and Thailand. However, most countries do not have their own reference database due to the difficulty of recording comprehensive coordinate points for all areas of the country, and such spatial data as there is may be distributed among several different governmental agencies.

In Thailand, the hierarchy of address components starts from the smallest administrative unit and runs progressively to the largest. Address details will typically include the house number, street name, district, municipality, name of jurisdiction (such as province), postal code, and country. However, unlike in some other countries, in Thailand, the postal code can identify the geographic boundary only at the provincial level, and old city layouts increase the difficulty of the geocoding process. The complexity of geocoding Thai addresses is compounded by the different structures of urban and rural addresses. Rural addresses are comprised of the house number, village number, street name, sub-district, district, province and postal code, but this structure causes most online geocoding services to return a false position. Beyond these problems, most studies on geocoding quality have been performed in western countries, where the majority of online geocoding services are based. So, these results cannot be generalized to include countries that have different addressing systems. The outcome of this is that in practice, users choose a geocoding service according to their ease of access, without knowledge of the quality of its results.

The purpose of this study was therefore to analyze the Thai address system and to compare the quality of the geocoding services available in Thailand that can convert addresses, place names, and POIs, stored in Thai text, into geographic coordinates. The online geocoding services tested were Google, Bing, Yahoo!, MapQuest, and OpenCage. Geocoded results were quantified based on their match rate, which is the rate at which specific locations were found, and the positional accuracy of reported locations, which is reported in meters calculated from the actual referenced location. The quality of the tested geocoding services was then statistically compared and errors (or dislocations) obtained from the experimental results were analyzed to help to make the geocoding process more suitable for use in Thailand and in other developing countries that may experience similar problems in their geocoding.

2. Background and related works

A number of different addressing systems are used in various countries. Some of the address formats used are summarized in Table 1. Japanese addresses have a block-based structure and these blocks are separated by streets, although Japanese streets do not have names [9-10]. Addresses in Japan may in fact contain more elements including block address, city block, groups of city blocks, city, major city, and country [3]. Korea uses a street address system. Its format is based on a road name and building number, with the entire address consisting of city or province, district or smaller city, town, road name, building number, and other components (floor, room number, or apartment number) [5]. Chinese addresses contain elements for country, province, city, district or county, sub-district or town, community or village, residential area, block, street number or building number, and room number [11-13, 3, 14]. The complexity of the address string format in China and poor address management has resulted in multiple address authorities, spread among various governmental agencies [13].

Table 1 Address formats used in western and Asian countries

	Address Format
Western Countries	
United States [3, 15]	House number, street number, street name, city, state, zip code, country
Australia [16]	Street (sub-unit, number, name), locality, state, postcode
Asian Countries	
Japan [17, 3, 10]	Country, province (*ken*), municipality (special wards, city or county), locality (district, town or village), city district, block, building number
Korea [5, 6, 18]	Country, city or province, district or smaller city, town, road name, building number, other components (floor, room number, apartment number)
China [3, 11, 13, 19]	Country, province, city, district or county, sub-district or town, community or village, residential area, block, street number or building number, room number
India [8]	-Rural addresses: country, state, district, village, road -Urban addresses: country, state, city, locality, sub-locality, sub-sub-locality, road, building
Thailand	-Rural addresses: House number, village number (*moo*), alley (*soi*), sub-district (*tambon*), district (*amphoe*), province (*changwat*), postal code, country -Urban addresses: House number, alley (*soi*), street name (*thanon*), sub-district (*tambon*), district (*amphoe*), province (*changwat*), postal code, country

Table 2 The five free online geocoding service providers

Service name*	Company	Limitation of non-commercial service
Google Maps	Google Inc.	2,500 requests per day
MapQuest Maps	MapQuest Inc.	15,000 transactions per month
Bing Maps	Microsoft Corporation	50 jobs per day
Yahoo! PlaceFinder	Yahoo! Inc.	5,000 queries per IP per day
OpenCage Geocoder	OpenCage Data Ltd.	2,500 queries per day

*All services provide global coverage with RESTful APIs

Thai addresses use a different language (Thai) and are usually written without a comma between the address elements, which go from the smallest to the largest unit. These usually consist of the house number, village number (moo) (for rural addresses), alley (soi) (for urban addresses), street name (thanon) (for urban addresses), sub-district (tambon), district (amphoe), province (changwat), and postal code. Address components can be extracted from the prefix of each address element. It should be noted that in Thailand, only house number is used as an address component. The building number is not a part of an address.

Thai postal codes consist of a five-digit number indicating the region, province, and the type of post-office service. For instance, in the postal code '10300', the first digit ('1') refers to the central region, the second digit ('0') refers to the Bangkok area, and the third to the fifth digits ('300') refer to the type of post-office service. The fifth digit itself indicates post-office services, i.e., the '0' shows both depositing and delivering service and the '1' offers only deposit service [20]. Addresses for locations in urban areas may contain a house number, alley, street, sub-district, district, province, and postal code. In rural areas, the address format may differ and contain a house number, village number, sub-district, district, province, and postal code. To complicate matters further, a house number may be mostly numbers or numbers plus letters, and hyphens and slashes are also commonly included. Complex examples of house addresses are: '169/7-12 Th. Surawong T. Suriyawong A. Bang Rak Ch. Bangkok 10500', '636 S. Charansanitwong 42 Intersection 19 Th. Charansanitwong T. Bang Yi Khan A. Bang Phlat Ch. Bangkok 10700', or '334/29 S. Samukkee M.9 T. Arunyik A. Muang Phitsanulok Ch. Phitsanulok 65000'.

Geocoding is a process for returning geographic coordinates for a location from a text address or a place name. This is achieved by comparing the text address and the place descriptions in a reference database. Two important elements of this are therefore the reference database and the algorithm that is used to process the text [1, 21-22]. The geocoding algorithm will include translation or interpretation of the input text. This is a process for searching the reference database for the details of the input text, matching the input text to entries found in the reference database, selecting the best answer and returning this as the preferred solution. The reference database consists of data descriptions and the actual coordinates for the algorithm to use to interpolate the location on a map. However, the reference database may be based on an address data model that is structured in a manner that does not closely parallel the input text. For example, the reference database might be built on a street network, land parcel boundary, or address point data model [22]. The selection of an appropriate algorithm for managing this process and the degree of completeness of the reference data are thus important factors that directly influence the quality of the geocoding results.

Online geocoding services are available on the Internet in the form of web services. With these, the service provider selects an algorithm to process the input text and to search the appropriate reference database. Then, the provider will return a location based on this information. Users specify the text and the results are obtained from the service provider. So, users of these services need not have even basic knowledge of the geocoding process. Rather, the service provider controls, develops, and maintains the entire geocoding process and consequently, online geocoding services can geocode or search for addresses or text and return a location as a result to users in real-time.

Previous research that investigates and compares the quality of online geocoding services has primarily been done in the field of public health and epidemiology, due to the latter's need for accurate coordinates for the location of patients and disease outbreaks to reliably perform analysis [23]. Comparative research has investigated areas ranging from the choice of algorithms or tools used to transform the input text and to choose matching entries in the reference databases, to the selection of geocoding services from various commercial service providers [24-28]. After online geocoding services became more widely available, studies [2, 29-30] were conducted comparing the quality of geocoded results from online service providers. It was found that Google, Bing, Yahoo! and MapQuest had the same level of quality. However, it is difficult to generalize the results of these studies to other regions, specifically to developing countries such as Thailand.

3. Methods

In this study five online geocoding service providers were tested, Google, MapQuest, Bing, Yahoo! , and OpenCage. These providers can geocode Thai text and points of interest within Thailand free of charge, although commercial services are also available. Geocoding services may be accessed through application programming interfaces (APIs) and the geocoding results are output with confidence levels. The characteristics of the five free online geocoding services are summarized in Table 2.

In the current study, all inputs for each geocoding service are addresses or points of interest (POI), entered in the Thai language and all test locations were within the borders of Thailand. To test the quality of the services, two experiments were run, each with a different input format. The tests were conducted between August 2016 and August 2017, and the environment settings of each experiment were as follows.

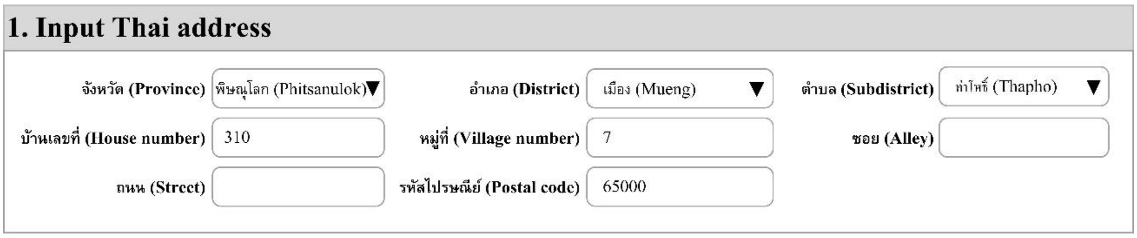

Figure 1 An input form for entering Thai addresses

Figure 2 An online map for drawing the boundary of a geo-referenced area

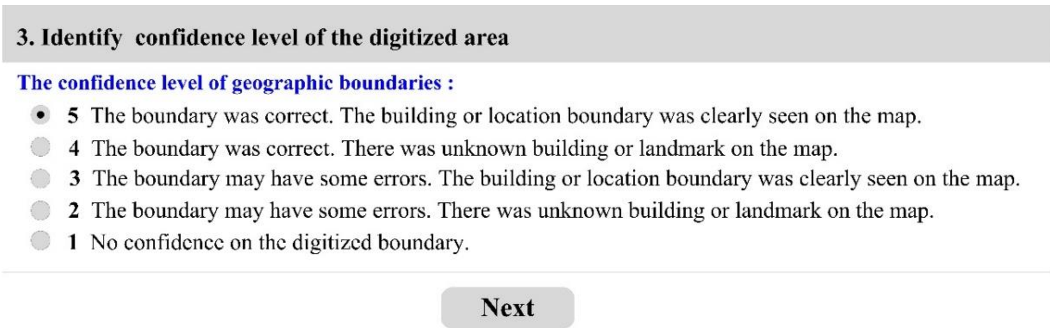

Figure 3 Confidence level assigned to a digitized area

3.1 Experiment I: Text addresses

The textual address inputs for Experiment I were in two input formats. The first format, referenced hereafter as Test-1, consisted of a house number, village number, alley, street, sub-district, district, province, and postal code. The second format, referenced henceforth as Test-2, had the same data structure, but with no postal code. This was to allow the researchers to study the impact of the inclusion of the postal code on the results, since the Thai postal code is associated with a geographical boundary that is only visible at the provincial level. Due to the small volume of reliable information relating addresses to map locations available in Thailand, the researchers developed a web application to collect text addresses with accompanying map-based information. This data was gathered with the cooperation of participants from across the country.

Screenshots of the web interface are shown in Figures 1-3, which illustrate the interface presented to users. Figure 1 shows that the user enters the required address information and is then asked to search for the location of the entered address on a map. The user draws the boundary of the geo-referenced area (Figure 2) on the map, and s/he specifies a confidence level to reflect how well s/he thinks the digitized area reflects the reliability of the data. The confidence level is a numerical value from 1 to 5 and Figure 3 contains a key describing these individual values. When the user submits the form, the application sends the input data in the format of house number, village number,

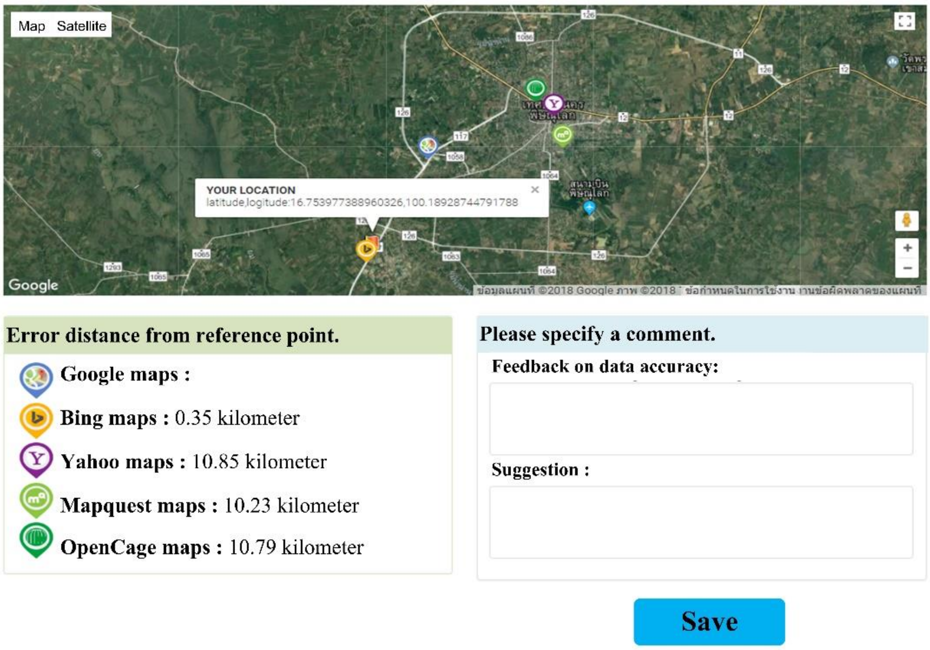

Figure 4 The geocoded results of the web application returned to the user

alley, street, sub-district, district, province and postal code to each geocoding service. The user's input and the responses returned from each service were stored in a database and the centroid of the digitized boundary was used as a reference. The web application then returned all the geocoded results and the centroid of the user's digitized area to the user, as shown in Figure 4. At this final stage, it was possible for the user to leave a comment, should he or she wish to do so.

In our first set of experiments (Experiment I), we expected the sample size of the address input to be at least 1,000 addresses at a confidence level of 5.

3.2 Experiment II: Point of Interest

The inputs for Experiment II were the Thai names of points of interest (POIs). The inputs were in two formats: (a) only the names of the POIs (referred to as Experiment IIa) and (b) the names and the administrative data of the POIs (referred to as Experiment IIb). The reason for using the POI names with their accompanying administrative data is that the researchers wished to test the impact of this on match rates and positional accuracy. The administrative data in Experiment IIb consisted of the sub-district, district, province, and postal code. The impact of the postal code data on the success of the geocoding process was studied by dividing the test into Experiment IIb-1, which was conducted with postal codes, and Experiment IIb-2, which was done without them.

The POIs used in Experiment II comprised a dataset of schools and academic institutes acquired from Thai national data sources [31]. This dataset contains attributes such as school ID, school name, sub-district, district, province, postal code, school type, and latitude and longitude. The dataset was cleaned to remove records with incomplete data and the dataset was therefore reduced to 34,334 records. The POIs were randomly sampled using proportional stratified sampling for the 76 provinces of Thailand and this yielded a sample of 5,000 POIs.

Data relating to each POI were prepared according to the defined input formats, then submitted to the five geolocation services. The responses from each service were stored in a database for later analysis. Summaries and examples of the input formats of each experiment are presented in Table 3.

3.3 Analysis (Match rate/Positional accuracy)

The geocoding quality of each service was analyzed and compared using the match rate and positional accuracy in both experiments. Here, the match rate is the number of geocoded addresses returned divided by the total number of addresses submitted. The quality of the matched results was graded as 'good', 'ambiguous', or 'unmatched'. Since there is no clear consensus on how to report the quality of geocoded results, this study sorted the codes and messages returned by each service into groups, as presented in Table 4. Positional accuracy is the average distance by which the geocoded points deviated from its actual location. This distance is a Euclidian distance between the reference point and the geocoded point. The coordinates of the reference points and geocoded points were in latitude/ longitude. However, they were transformed into the projection reference, WGS 84 / UTM zone 47N (EPSG:32647), before calculating the distance. Thus, the distances were in meters. Since these errors are typically not normally distributed [29, 32], this study uses nonparametric statistics for testing and comparison of the various services. Thus, the Friedman test, a one-way ANOVA test, is used here to assess differences between positional accuracy among several related samples. This research also uses the Wilcoxon signed rank test, a non-parametric statistical hypothesis test, which is used to compare two related samples for each pair of services.

4. Results and analysis

4.1 Experiment I

A total of 1,511 addresses were collected from the research participants. Their locations are shown in Figure 5.

Table 3 The input format of each experiment

Experiment	Input format	Description	Example
Experiment I–Test-1	Text address with postal code	Input consists of house number, village number, alley, street, sub-district, district, province and postal code	397, 9, ท่าโพธิ์, เมืองพิษณุโลก , พิษณุโลก, 65000 (397, 9, Thapho, Mueng Phitsanulok, Phitsanulok, 65000)
Experiment I– Test-2	Text address without postal code	Input consists of house number, village number, alley, street, sub-district, district and province	397, 9, ท่าโพธิ์,เมืองพิษณุโลก, พิษณุโลก (397,9,Thapho,Mueng Phitsanulok, Phitsanulok)
Experiment IIa	POI names	Thai name of point of interest (POI)	โรงเรียนลาซาล (La Salle School)
Experiment IIb –1	POI names with administrative data and postal code	Input consists of POI name, sub-district, district, province and postal code	โรงเรียนลาซาล, บางนา, เขตบางนา, กรุงเทพมหานคร, 10260 (La Salle School, Bangna, Bangna district, Bangkok, 10260)
Experiment IIb –2	POI names with administrative data but without postal code	Input consists of POI name, sub-district, district and province	โรงเรียนลาซาล, บางนา, เขตบางนา, กรุงเทพมหานคร (La Salle School, Bangna, Bangna district, Bangkok)

Table 4 Matched types of geocoded results returned from the five geocoding services

Type of match	Google Location type	MapQuest Geocode Quality	Bing Confidence	Yahoo! Place content	OpenCage Confidence
Good	Rooftop	Point	High	Point of Interest	10 (distance < 0.25 km)
Ambiguous	Range interpolated	Street	Medium	Town	9 (distance < 0.5 km)
	Geometric center	Zip	Low	Province	8 (distance < 1 km)
	Approximated	City		Postal code	7 (distance < 5 km)
		Country			6 (distance < 7.5 km)
					5 (distance < 10 km)
					4 (distance < 15 km)
					3 (distance < 20 km)
					2 (distance < 25 km)
					1 (distance ≥ 25 km)
Unmatched	No geocoded result	No geocoded result	No geocoded result	No geocoded result	No geocoded result

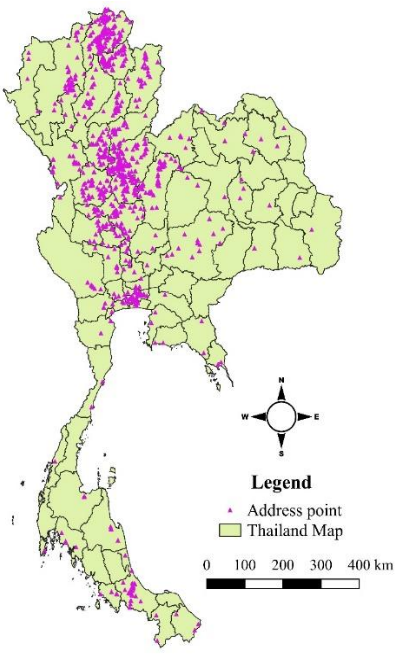

Figure 5 The location of the address points entered by the 1,511 participants

In terms of the confidence levels, there were 1,100, 206, 140, 53, and 12 addresses at confidence levels of 5, 4, 3, 2, and 1, respectively. Only the addresses at confidence level 5 were selected for further analysis and the match rates of these are summarized in Table 5. The results of Test-1, which used addresses with their postal code, show that Bing returned the highest rate of 'good' matches (17.82%), but it also had the worst unmatched rate (68.45%). The other services returned many more 'ambiguous' matches. Rates for these were: MapQuest (98.55%), Yahoo! (97.45%), OpenCage (88.18%), and Google (61.45%). Alternatively, when tested on inputs without postal codes (Test-2), the rate of 'good' matches increased for MapQuest (21.64%), Bing (17.91%), and Google (17.27%). It can be concluded from this that when dealing with Thai addresses, inclusion of the postal code in the address affects the matching rate of all the tested geocoding services. Addresses input without postal code information returned geocoded data of a larger number of matches. In Test-1, other address elements (apart from the postal code) did not appear to have an effect on the results.

The descriptive statistics of the error distances for all geocoded points (that is, only the address inputs categorized at confidence level 5) are summarized in Table 6. In Test-1 (with postal codes), Bing produced, on average, the highest accuracy with the smallest error distance

Table 5 Match rate of the addresses with a confidence level of 5 (N = 1,100)

	Match rate of addresses at confidence level 5									
	Google		MapQuest		Bing		Yahoo!		OpenCage	
	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2
Count (N)	717	1,016	1,100	896	347	1,088	1,088	1,058	973	1,028
Total match rates (%)	65.18	92.36	100	81.45	31.55	98.91	98.91	96.18	88.45	93.45
Match level										
Good (%)	3.73	17.27	1.45	21.64	17.82	17.91	1.45	1.45	0.27	-
Ambiguous (%)	61.45	76.09	98.55	60.82	13.73	82.09	97.45	95.82	88.18	94.45
Unmatched (%)	34.82	6.64	-	17.55	68.45	-	1.09	2.73	11.55	5.55

Table 6 Positional accuracy (error distance) of addresses at confidence level 5

Service	Count		Min (m)		Max (m)		Median (m)		Mean (m)	
name	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2
Google	717	1,016	0.003	2.26	438.16	434,395.35	2.66	1,988.75	4.77	4,652.22
MapQuest	1,100	896	0.038	1.33	16,645.20	662,085.34	3.80	2,309.08	1,190.77	24,280.22
Bing	347	1,088	0.017	4.52	96.26	436,589.77	1.93	2,149.96	5.92	5,644.72
Yahoo!	1,088	1,058	0.015	15.20	15,372.22	1,015,563.03	3.57	3,496.60	111.72	25,052.77
OpenCage	973	1,028	0.159	158.62	425.05	425,402.75	13.26	13,206.25	21.68	21,473.71

Table 7 Positional accuracy (error distance) of the addresses at confidence level 5 that were geocoded by all services

Service	Count		Min (m)		Max (m)		Median (m)		Mean (m)	
name	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2	Test-1	Test-2
Google	228	821	0.003	2.26	56.08	434,395.35	1.52	2,131.24	2.91	4,912.55
MapQuest	228	821	0.041	1.33	14,477.55	662,085.34	3.20	2,383.76	2,377.11	23,447.34
Bing	228	821	0.020	6.36	96.26	436,589.77	1.89	2,222.01	5.30	4,670.88
Yahoo!	228	821	0.015	71.83	8,057.11	1,015,563.03	4.10	3,495.41	205.80	26,944.47
OpenCage	228	821	0.159	158.62	78.10	425,402.75	4.73	14,713.14	12.41	22,624.68

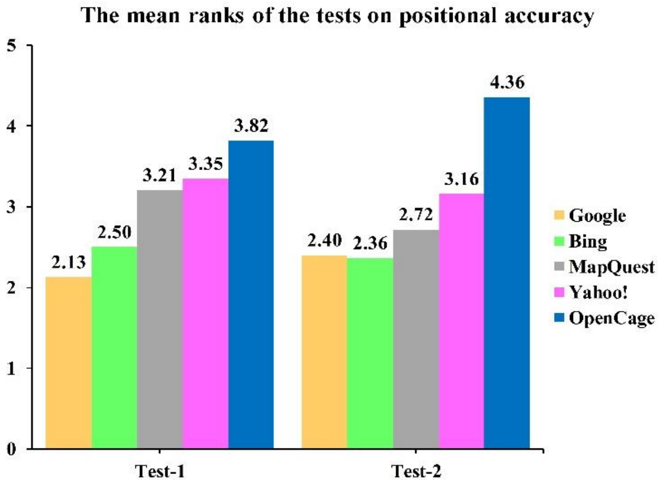

Figure 6 The mean ranks of the Friedman tests on positional accuracy for the address inputs at a confidence level of 5 that were geocoded by all services

(median 1.93 m) followed by Google (2.66 m), Yahoo! (3.57 m), MapQuest (3.80 m), and OpenCage (13.26 m). In Test-2 (without postal codes), although the match rates were higher, all services returned lower positional accuracy than in Test-1 (conducted with postal codes).

To statistically compare the geocoding quality of the services, only addresses that could be geocoded by all five services were used. Thus, the number of addresses used in the analysis was reduced to 228 samples for Test-1 and 821 samples for Test-2. Table 7 summarizes the descriptive statistics for the distance errors of these tests.

The positional accuracy of all services was compared using the Friedman test. The mean ranks of the analysis of the two tests are reported in Figure 6. In both tests, the results reveal that the error distances for the five services were significantly different (Test-1: $X^2 = 168.449$, $p < 0.05$ and Test-2: $X^2 = 890.638$, $p < 0.05$). Comparisons between each pair of the five services were made using the Wilcoxon signed rank test. In Test-1 (with postal codes), we found that the services were divided into two groups. The first group, Google and Bing, returned statistically better positional accuracy than did the second group of MapQuest, Yahoo!, and OpenCage at $p < 0.05$. In Test-2 (without postal codes), we found that the positional accuracy of the services was significantly different ($p < 0.05$), except for Google and Bing. Rated from high to low accuracy, the services

Table 8 Match rates of the geocoded POI names

	Match rates of all services (N = 5,000)				
	Google	**MapQuest**	**Bing**	**Yahoo!**	**OpenCage**
Count (N)	197	598	100	3,214	724
Match rates (%)	3.94	11.96	2.00	62.48	14.48
Match level					
Good (%)	0.14	-	2.00	62.08	14.02
Ambiguous (%)	3.80	11.96	-	0.40	0.46
Unmatched (%)	96.06	88.04	98.00	37.52	85.52

Table 9 Positional accuracy (error distance) of the geocoded POI names (N = 5,000)

Service name	Count	Min (m)	Max (m)	Median (m)	Mean (m)	STD (m)
Google	197	0.63	1,401,152.40	1,538.63	127,941.13	234,348.18
Bing	100	1.34	889,666.56	1,607.75	121,394.74	221,696.41
MapQuest	598	6.86	1,150,163.00	66,575.73	148,052.73	205,396.02
Yahoo!	3,124	3.48	1,476,697.00	188.26	87,316.77	191,642.00
OpenCage	724	2.59	1,504,287.90	11,643.03	143,280.57	227,750.33

Table 10 Positional accuracy of the POIs with name inputs that could be geocoded by all services (N = 11)

Services name	Min (m)	Max (m)	Median (m)	Mean (m)	STD (m)
MapQuest	14.37	177,830.38	162.75	16,317.93	53,567.78
Bing	97.40	824,928.00	438.97	122,580.11	259,016.37
Yahoo!	32.33	8,262.68	157.83	1,476.01	2,894.00
OpenCage	50.74	177,830.31	162.79	16,321.97	53,566.41

were ordered as follows: 1) Google and Bing, 2) MapQuest, 3) Yahoo!, and 4) OpenCage.

With regard to Experiment I, the geocoding quality of most services was sensitive to the inclusion of postal code data in Thai text addresses. Most services produced higher match rates when the addresses had no postal code. This was with the exception of MapQuest, which had a higher level of both ‘good’ matched rates and unmatched rates, and Yahoo!, which had a higher level of unmatched returns. This finding differs from the results of geocoding tests conducted in other countries, where it is usual for the inclusion of the postal code to increase match rates [33-35]. In these tests, MapQuest misidentified 143 addresses by returning addresses in the wrong country. This is due to an error misidentifying the Thai five-digit numeric postal code as the postal code of another country. For example, the 65000 postal code of Mueng District, Phitsanulok Province, Thailand was matched to both Odesa in Ukraine and Tarbes, Arrondissement de Tarbes, Hautes- Pyrénées, Midi-Pyrénées, France. However, it should be noted that the Thai postal code geocodes at the provincial level and has no geospatial relationship to lower administrative levels. This may be the reason that some geocoding services do not take the postal code into account. In contrast to match rates, the positional accuracy of all services using addresses with postal codes was on average (median values less than 5 m) higher than the results for address without postal codes (median values more than 2 km).

4.2 Experiment IIa

The match rates for the five services when submitting only the names of 5,000 sample POIs are reported in Table 8. Here, Yahoo! had the highest match rate at 62.48%, while Bing produced the lowest (2.00%). When considering match rates according to match types, Yahoo! produced the highest proportion of ‘good’ match rates (62.08%), while MapQuest did not return any responses at all with a ‘good’ match. With unmatched rates higher than 85%, Google, MapQuest, Bing and OpenCage mostly returned no geocoded results. One reason for this high failure is that these service providers do not have many academic institutes recorded as POIs in Thailand. Additionally, the names of many academic institutes in Thailand are not unique. For example, there are 101 Thairathvithaya schools nationwide (funded by the Thairath Newspaper Foundation) and each school is uniquely identified by a branch number and a name that makes reference to a local landmark, such as the name of a village or temple.

The results of testing for the positional accuracy of 5,000 POIs geocoded by all five services using name inputs is presented in Table 9. Yahoo! produced the highest positional accuracy (median 188.26 m), followed by Google and Bing (median > 1,500 m), while MapQuest and OpenCage returned results with the lowest level of accuracy (median > 11,500 m).

To make a fair comparison between the services, only the 11 POIs that could be geocoded by four the services (Bing, MapQuest, Yahoo!, and OpenCage) were tested. The results of these tests are presented in Table 10. The positional accuracies of these four services were compared using the Friedman test. The results show that the services were significantly different ($X^2 = 10.485$, $p < 0.05$). As shown in Figure 7, the mean rankings were MapQuest 1.91, OpenCage 2.00, Yahoo! 2.68, and Bing 3.41. The comparisons between each pair of services (6 pairs in total) were made using the Wilcoxon signed rank test. It was revealed that the error distances of the pairings of MapQuest and Bing ($z = -2.490$, $p = 0.0128$) and Bing and OpenCage ($z = -2.490$, $p = 0.0128$) were significantly different ($p < 0.05$), but the pairings of MapQuest and Yahoo! ($z = -0.978$, $p = 0.3281$), MapQuest and OpenCage ($z = -0.535$, $p = 0.5930$), Bing and Yahoo! ($z = -0.968$, $p = 0.3329$), and Yahoo! and OpenCage ($z = -0.978$, $p = 0.3281$) were not significantly different.

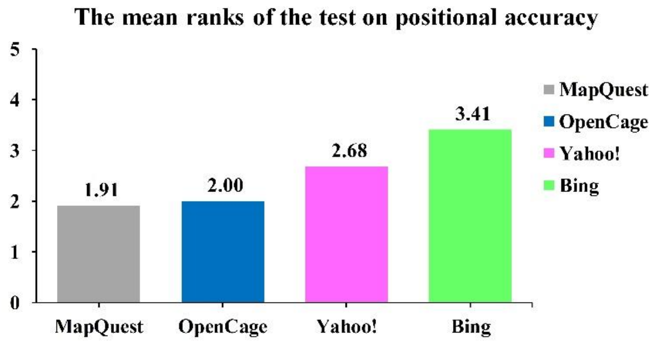

Figure 7 The mean ranks of the Friedman test on positional accuracy for POI names that could be geocoded by all services

Table 11 Match rates for Experiment IIb-1 (POI names with administrative information, including postal codes) and Experiment IIb-2 (POI names with administrative information but without postal codes)

	Google		MapQuest		Bing		Yahoo!		OpenCage	
	Exp. IIb-1	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**
Count	1,139	4,616	4,879	4,997	8	4,751	5,000	4,997	4,809	5,000
Match rates (%)	22.78	92.32	97.58	99.94	0.16	95.02	100	99.94	96.18	100
Match level										
Good (%)	0.14	12.24	-	-	0.16	0.02	-	15.90	75.04	-
Ambiguous (%)	22.66	80.10	0.06	33.90	-	95.00	94.40	78.44	-	75.68
Incorrect (outside Thailand) (%)	-	-	97.52	66.10	-	-	5.60	5.60	21.14	24.32
Unmatched (%)	77.20	7.66	2.42	-	99.84	4.98	-	0.06	3.82	-

Table 12 Positional accuracy of the geocoded results of POI names provided with administrative information (Experiment IIb-1 and Experiment IIb-2) (N = 5,000)

Service name	Count		Min (m)		Max (m)		Median (m)		Mean (m)	
	Exp. IIb-1	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**	**Exp. IIb-1**	**Exp. IIb-2**
Google	1,139	4,616	0.42	0.11	229,258.42	589,943.81	129.37	130.04	2,511.82	2,489.20
MapQuest	3	1,692	2,465.24	30.30	9,602.23	540,476.06	7,197.58	3,949.44	6,421.68	7,474.80
Bing	8	4,751	4,256.50	29.53	21,682.36	1,284,847.40	15,834.29	6,962.88	15,137.95	61,089.55
Yahoo!	4,720	4,717	1.34	1.34	296,581.75	411,715.66	6,600.90	6,600.90	9,123.08	9,202.52
OpenCage	3,752	3,784	82.19	82.19	254,612.28	411,590.28	11,197.59	11,250.47	20,462.46	20,588.94

To summarize the results of Experiment II, when testing with points of interest (POIs), Yahoo! produced the highest number of 'good' matches and these results also had better positional accuracy than did the other geocoding services. Although OpenCage produced quite a high level of 'good' matches using POI names as the input, the average error distance was 16,162.18 meters, which is more than the estimated distance (< 250 meters) defined with a confidence score of 10, referred to Table 4. Google and Bing produced a lower level of matches than MapQuest and OpenCage, but their positional accuracy was better. One reason that these services returned a low level of matches is that some POIs are located in rural or border areas, which may not be included in their reference databases.

4.3 Experiment IIb

The match rates of Experiment IIb-1 (using as inputs the POI names along with administrative information, including postal codes) and Experiment IIb-2 (using the POI names with administrative information but without postal codes) are summarized in Table 11. In Experiment IIb-1, Yahoo! produced the highest match rate (94.40%) followed by MapQuest (97.58%), OpenCage (96.18%), Google (22.78%), and Bing (0.16%). Yahoo!, MapQuest, and OpenCage appear to have high match rates, but 280 addresses returned by Yahoo!, 1,057 by OpenCage and 4,876 by MapQuest were outside Thailand. These were thus marked as incorrect, yielding incorrect match rates of 5.60%, 21.14%, and 97.52%, respectively. Most of the OpenCage responses (75.04%) returned incorrect postal codes, causing the coordinates to be located in another city. For example, a correct address of St. Louis Suksa School, 23 Sathorntai Road, Yanawa Sub-district, Sathorn District, Bangkok Province 10120 was geocoded by OpenCage as Changwat Uttaradit 10120, Thailand, an error of some 300 km. Thus, OpenCage yielded very low positional accuracy. Bing, however, produced the lowest match rates.

In Experiment IIb-2, we tested the impact of the postal code on the returns by excluding it from the submitted administrative information. In this condition, the match rates of all services were higher than 92% and those of Google, MapQuest, Bing and OpenCage increased. However, Yahoo! and OpenCage showed only small differences in their match rates compared to Experiment IIb-1. Considering only the 'good' match rates in Experiment IIb-2, Yahoo! returned the highest rate (15.90%), although at 12.24%, Google also had a higher rate of 'good' matches than in Experiment IIb-1.

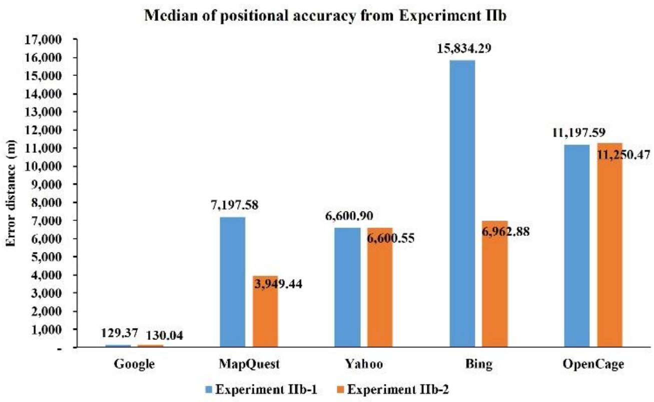

Figure 8 The positional accuracy of each geocoding service in Experiment IIb-1 and Experiment IIb-2

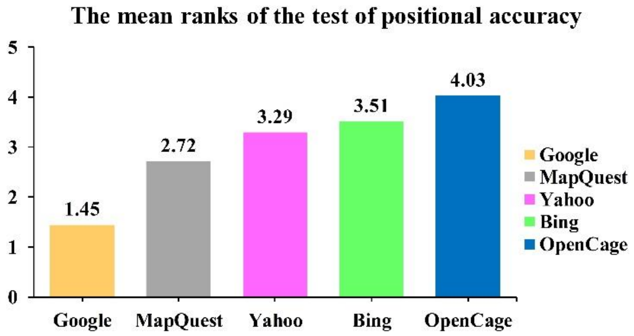

Figure 9 Mean ranks computed by the Friedman test of positional accuracy for the POI names (provided with administrative information but without postal codes) geocoded by all services

Alternatively, MapQuest and OpenCage did not return any 'good' responses. The ambiguous matched rates of all services except Yahoo! increased. Although Yahoo! returned more 'good' matches than Google, the positional accuracy of the Yahoo! geocoded results was lower than Google's. It is noteworthy that Bing's geocoding services provided a way to allow the user to specify the input region, i. e. , "userRegion=TH", and the culture, i.e., "culture=th" and thus it was possible to restrict Bing to returning geocoded responses within this predefined territory.

Descriptive statistics for the positional accuracy of the geocoded results obtained in Experiment IIb-1 (using the POI name along with administrative information including postal code as inputs) and Experiment IIb-2 (as before but without postal codes) are reported in Figure 8 and Table 12. In Experiment IIb-1, Google produced the highest positional accuracy (median 129.37 m) while Bing had the lowest (median 15,834.29 m). In Experiment IIb-2, Google still produced the highest positional accuracy (median 130.04 m), close to that of Experiment IIb-1, while MapQuest (median 3,949.44 m) and Bing (median 6,962.88 m) increased in accuracy. However, the positional accuracy of Yahoo! and OpenCage showed only minor differences, with medians of 6,600.55 m and 11,250 m, respectively.

The size of the errors (i.e. the distance by which they missed their target of all services were statistically compared using the Friedman test. It should be noted that the Friedman test could not be used with Experiment IIb-1 because there were too few matching results from MapQuest (3 addresses) and Bing (8 addresses). Experiment IIb-2 yielded a total of 1,034 inputs that could be geocoded by all services. In this case, it was possible to compare these using the Friedman test. It was thus shown that the error distances of the five services were significantly different (X^2 = 1,614.53, $p < 0.05$). The mean ranks of the five services are reported in Figure 9.

Comparisons between each pair of services (10 pairs in total) were again made using the Wilcoxon signed rank test. It was revealed that for Experiment IIb-2, the error distances for pairs of services were significantly different ($p < 0.05$), except for the pairing of Bing and OpenCage ($z = -0.467$, $p = 0.640$).

5. Discussion

Overall, Google returned lower match rates than did the other geocoding services, but on average, its matches also showed the highest positional accuracy. In our test of the sensitivity to postal code information (i.e., when inputs included Thai postal codes), Google also performed better than did other services. MapQuest was also able to geocode text addresses without postal codes with a rather high level of matches. However, MapQuest returned geocoded results with locations outside Thailand when submitting POIs with names and administrative information. Yahoo! produced a high level of 'ambiguous' matches with matches being made at lower levels (land features, suburbs, towns, and provinces) for both the text address format and when POIs were geocoded by name and administrative information. Bing, in turn, returned better geocoded results when submitting text addresses with the postal codes excluded. It is notable that, when geocoding input for Bing, users should not enter special characters, such as periods (.), commas (,), colons (:), or plus signs (+) as part of the parameter value in structured

URLs, although a hyphen (-) is acceptable as a placeholder. In the case of OpenCage, although this service returns a high level of 'good' matches with the correct input name, this service may also return results in an area located some distance from the true position (as far as 11 km away). In addition to this, when Thai addresses are written, they often include labels which mean 'village number', 'alley', 'street', 'sub-district', 'district', or 'province'. For example, a typical address might be '888/69 หมู่ 14 ซอย 2 ถนนพระพิฆเนตร ตำบลวัดไทร อำเภอเมืองนครสวรรค์ จังหวัดนครสวรรค์ 60000' which could be translated literally as 'House number 888/69, Village number 14, Alley 2, Street Prapikanate, Sub-district Watchai, District Mueng Nakhon Sawan, Province Nakhon Sawan'. However, most services returned geocoded results with higher confidence and greater accuracy when no such labels were included in the address input. For example, when submitting '397, หมู่ 9, ตำบลท่าโพธิ์, อำเภอเมืองพิษณุโลก, จังหวัดพิษณุโลก, 65000' to MapQuest and Yahoo!, the services returned geocoded results in the U.S. and in Japan. So this study did not include labels in the input submitted to any geocoding services. Continuing the example given above, this would then be submitted as '397, 9, ท่าโพธิ์, เมืองพิษณุโลก, พิษณุโลก, 65000' and indeed when submitted without these labels, all services correctly returned accurate geocoded results in Phitsanulok Province, Thailand.

In summary, there are two main problems in geocoding Thai text addresses. The first is the combined complexity of city layouts and the assignment of house numbers in Thailand. The mixture of old and new city layouts prevents application of common data models, such as street network or block coding. The current pattern of house numbering cannot be matched against street lines due to mixing of house locations, especially in private areas and on housing estates. The second problem is a lack of a spatial reference database. Although Thailand has developed a Fundamental Geographic Data Set (FGDS) with the aim of publishing common spatial datasets, none of these datasets can be used as a reference database for geocoding. The current datasets of Thailand's FGDS contain insufficient data to develop a complete and accurate reference database for geocoding at a rooftop level. Only rudimentary geospatial data (13 data layers such as administrative boundaries, roads, land use, DEM, etc.) are published. It is therefore urgent that a suitable data model be developed for use as a geocoding reference database in Thailand. The address point model may be a potential solution to this problem, but further analysis of data sources and the quality of geocoding results is required.

6. Conclusions

This study presents an analysis of online geocoding services offered by Google, MapQuest, Bing, Yahoo!, and OpenCage. In the experiments, the inputs comprised text addresses and the names of points of interest (POIs), written in Thai. The geocoded results were compared against data obtained from maps completed by participants showing the actual locations of their addresses, while the coordinates of the POIs were acquired from Thai national data sources. Geocoding quality was measured using match rates and positional accuracy, and the overall experimental results revealed that Google, Bing, Yahoo!, and OpenCage had match rates higher than 90%, but that the rate for MapQuest was slightly lower at 82%. Considering only 'good' matches, MapQuest, Google and Bing returned these at a rate of 20.32%, 19.72%, and 17.01%, respectively. In contrast, Yahoo! and OpenCage returned 'ambiguous' matches at a rate that was in excess of 90%. MapQuest, OpenCage, and Yahoo! also produced incorrect results that showed locations outside Thailand at rates of 66%, 21%, and 5%, respectively, while Bing produced 70% unmatched rates when using text addresses and POIs with postal codes. In terms of accuracy, Google produced errors with the lowest average distance (between 2.66 to 1,988 m), whereas MapQuest returned the greatest error distance, which was on average 16,008 m. According to the statistical analysis, the quality of the five service providers, with regards to both the match rates and the positional accuracy of the text addresses and POI names (provided with administrative information), was significantly different ($p < 0.05$), and these were ranked in order of best to worst. Google and Bing were jointly best, followed by MapQuest, Yahoo!, and OpenCage. The quality of the four service providers when geocoding POI names was also significantly different. MapQuest, OpenCage, Yahoo!, and Bing gave results in order of highest to lowest quality.

This study provides quantitatively tested information on the quality of geocoding services that may be used as a guideline for users when attempting to choose an appropriate geocoding service. The information also provides valuable insights into the issues one must face when geocoding data in Thailand. It is hoped that this will be of use in future research and in the development of geocoding tools suitable for Thai addresses. The authors also are of the opinion that the results show that current address and location naming standards in Thailand are detrimental to effective use of geospatial software systems, which in the future are predicted to become increasingly essential to geographical applications and land planning. A major policy initiative by the Thai government is suggested to better normalize and standardize Thai address structures.

7. Acknowledgements

The research reported in this publication was supported by the National Science and Technology Development Agency, a part of the Ministry of Science and Technology, under contract SCH-NR2016-854.

8. References

[1] Goldberg DW, Wilson JP, Knoblock CA. From text to geographic coordinates: the current state of geocoding. URISA J. 2007;19:33-46.

[2] Roongpiboonsopit D, Karimi HA. Comparative evaluation and analysis of online geocoding services. Int J Geogr Inform Sci. 2010;24:1081-100.

[3] Karen KK. Encyclopedia of geographic information science. Thousand Oaks, California: SAGE Publications Inc.; 2008.

[4] Department of Industry, Innovation and Science. PSMA Geocoded National Address File (G-NAF) [Internet]. Australia: PSMA Australia; 2018 [cited 2018 Jun 8]. Available from: https://www.psma.com.au/sites/default/files/g-naf_product_description_1.pdf.

[5] Universal Postal Union. Addressing the world - An address for everyone. 1st ed. [Internet]. Switzerland: International Bureau Universal Postal Union; 2012 [cited 2012 Jun 5]. Available from: http://www.upu.int/fileadmin/documentsFiles/activities/addressingAssistance/whitePaperAddressingTheWorldEn.pdf.

[6] Seok S, Lee J. Development of geocoding and reverse geocoding method implemented for street-based addresses in Korea. J Korean Soc Surv Geodesy Photogramm Cartogr. 2016;34:33-42.

[7] Babu TR, Chatterjee A, Khandeparker S, Subhash AV, Gupta S. Geographical address classification without using geolocation coordinates. Paris: ACM Press; 2015.

[8] Chatterjee A, Anjaria J, Roy S, Ganguli A, Seal K. SAGEL: smart address geocoding engine for supply-chain logistics. Burlingame, California: ACM Press; 2016.

[9] Davis Jr. CA, Fonseca F. Assessing the certainty of locations produced by an address geocoding system. GeoInformatica. 2007;11:103-29.

[10] Informatica LLC 1993. Address verification best practices for Japan addresses [Internet]. 2017 [cited 2018 May 3]. Available from: https://kb.informatica.com/h2l/HowTo%20Library/1/0893-AddressVerificationBestPracticesforJapanAddresses-H2L.pdf.

[11] Chang CH, Huang CY, Su YS. On Chinese postal address and associated information extraction [Internet]. 2012 [cited 2018 May 4]. Available from: https://kaigi.org/jsai/webprogram/2012/pdf/726.pdf.

[12] Pan Y, Chen B, Lu Z, Li S, Zhang J, & Zhou Y. An address geocoding method for improving rural spatial information infrastructure. Proceedings of the Sixth International Symposium on Digital Earth: models, algorithms, and virtual reality; 2009 Sep 9-12; Beijing, China. USA: SPIE; 2010. p. 7840-06.

[13] Tian Q, Ren F, Hu T, Liu J, Li R, Du Q. Using an optimized Chinese address matching method to develop a geocoding service: a case study of Shenzhen, China. Int J Geo-Inf. 2016;5(5):1-17.

[14] Li L, Wang W, He B, Zhang Y. A hybrid method for Chinese address segmentation. Int J Geogr Inform Sci. 2018;32:30-48.

[15] Eckman S, English N. Creating housing unit frames from address databases: geocoding precision and net coverage rates. Field Meth. 2012;24:399-408.

[16] Geographical Names Board (NSW). NSW addressing user manual / geographical names board of New South Wales [Internet]. 2016 [cited 2018 Dec 7]. Available from: http://www.gnb.nsw.gov.au/__data/assets/pdf_file/0007/199411/NSW_AUM_July2018_Final.pdf.

[17] National Land Information Division. Block level location reference information maintenance method [Internet]. 2007 [cited 2018 May 3]. Available from: http://nlftp.mlit.go.jp/isj/method.html.

[18] Ministry of the Interior and Safety. Introduction Road Name Address [Internet]. 2011 [cited 2018 Jun 6]. Available from: http://www.juso.go.kr/CommonPageLink.do?link=/eng/about/GuideBook.

[19] Sun Z, Qiu AG., Zhao J, Zhang F, Zhao Y, Wang L. Technology of fuzzy Chinese-geocoding method. 2013 International Conference on Information Science and Cloud Computing (ISCC); 2013 Dec 7-8; Guangzhou, China. USA: IEEE; 2013. p. 7-12.

[20] Thailand Post Co Ltd. The content of postal code Thailand [Internet]. 2015 [cited 2015 Oct 7]. Available from: postalcare@thailandpost.com.

[21] Karimi HA, Durcik M, Rasdorf W. Evaluation of uncertainties associated with geocoding techniques. Comput Aided Civ Infrastruct Eng. 2004;19:170-85.

[22] Zandbergen PA. A comparison of address point, parcel and street geocoding techniques. Comput Environ Urban Syst. 2008;32:214-32.

[23] Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. 2003;14:408-12.

[24] Krieger N, Waterman P, Lemieux K, Zierler S, Hogan JW. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. Am J Public Health. 2001;91:1114-6.

[25] Yang D- H, Bilaver LM, Hayes O, Goerge R. Improving geocoding practices: evaluation of geocoding tools. J Med Syst. 2004;28:361-70.

[26] Whitsel EA, Quibrera PM, Smith RL, Catellier DJ, Liao D, Henley AC, et al. Accuracy of commercial geocoding: assessment and implications. Epidem Perspect Innov. 2006;3:1-12.

[27] Lovasi GS, Weiss JC, Hoskins R, Whitsel EA, Rice K, Erickson CE, et al. Comparing a single- stage geocoding method to a multi-stage geocoding method: how much and where do they disagree?. International Journal of Health Geographics. 2007;6:1-11.

[28] Cetl V, Kliment T, Jogun T. A comparison of address geocoding techniques – case study of the city of Zagreb, Croatia. Survey Review. 2018;50:97-106.

[29] Roongpiboonsopit D, Karimi HA. Quality assessment of online street and rooftop geocoding services. Cartography and Geographic Information Science. 2010;37:301-18.

[30] Pietro GD, Rinnone F. Online geocoding services: a benchmarking analysis to some European cities. 2017 Baltic Geodetic Congress (BGC Geomatics); 2017 Jun 22-25; Gdansk, Poland. USA: IEEE; 2017. p. 273-81.

[31] DGA Open Government License [Internet]. Thailand: School and Academic Institute Dataset 2014; 2014. [updated 2016 Jan 4; cited 2017 Jun 9]. Available from: https://data.go.th/DatasetDetail.aspxd=8548e3ab-00bf-4eae-b29a-156a4aa52c0d.

[32] Zandbergen PA. Geocoding quality and implications for spatial analysis. Geography Compass. 2009;3:647-680.

[33] M. Beyer KM, Schultz AF, Rushton G. Using zip codes as geocodes in cancer research. In: Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL, editors. The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice. USA: CRC Press Taylor & Francis Group; 2008. p. 37-67.

[34] Wey CL, Griesse J, Kightlinger L, Wimberly MC. Geographic variability in geocoding success for west Nile virus cases in South Dakota. Health Place. 2009; 15:1108-14.

[35] Rosu A, Chen D. An improved approach for geocoding Canadian postal code- based data in health- related studies. The Canadian Geographer / Le Géographe Canadien. 2016;60(2):270-81.