

A multifaceted comparison of ArcGIS and MapMarker for automated geocoding

Sanjaya Kumar, Ming Liu, Syni-An Hwang

Center for Environmental Health, New York State Department of Health, Troy, NY, USA

Abstract. Geocoding is increasingly being used for public health surveillance and spatial epidemiology studies. Public health departments in the United States of America (USA) often use this approach to investigate disease outbreaks and clusters or assign health records to appropriate geographic units. We evaluated two commonly used geocoding software packages, ArcGIS and MapMarker, for automated geocoding of a large number of residential addresses from health administrative data in New York State, USA to better understand their features, performance and limitations. The comparison was based on three metrics of evaluation: completeness (or match rate), geocode similarity and positional accuracy. Of the 551,798 input addresses, 318,302 (57.7%) were geocoded by MapMarker and 420,813 (76.3%) by the ArcGIS composite address locator. High similarity between the geocodes assigned by the two methods was found, especially in suburban and urban areas. Among addresses with a distance of greater than 100 m between the geocodes assigned by the two packages, the point assigned by ArcGIS was closer to the associated parcel centroid (“true” location) compared with that assigned by MapMarker. In addition, the composite address locator in ArcGIS allows users to fully utilise available reference data, which consequently results in better geocoding results. However, the positional differences found were minimal, and a large majority of addresses were placed on the same locations by both geocoding packages. Using both methods and combining the results can maximise match rates and save the time needed for manual geocoding.

Keywords: geocoding, ArcGIS, comparison of geocodes, MapMarker, match rate, positional accuracy, United States of America.

Introduction

Geocoding is widely used by researchers who conduct public health surveillance and spatial epidemiology studies, and the importance of it has been recognised by leading experts in the field. For example, geocoding was included in the objectives of “Healthy People 2010”, i.e. increase the proportion of major national health data systems that use geocoding to promote nationwide use of geographic information systems” (<http://www.healthypeople.gov/2010/>). Geocoding is also considered critical to achieving the various goals of “Healthy People 2020”, such as reducing health disparities and protecting children from air pollution (Swift et al., 2008; Healthy People 2020 Summary of Objectives, 2010; Krieger et al., 2012).

A number of studies have evaluated various geocoding methods, including in-house software, commercial firms, and web-based geocoding services in the United

States of America (USA). Researchers in the University of Southern California (USC) GIS Research Laboratory conducted a comprehensive evaluation of eight frequently used PC-based or online geocoding software packages, including Centrus, Geolytics, ESRI ArcGIS, Geocoder.us, Google Earth, Google Maps API, Yahoo API, and open source USC Geocoding Platforms (Swift et al., 2008). The results indicate that each of these platforms has strengths and weaknesses and, in general, no one performed significantly better or worse than the others. One limitation of this evaluation is that the input data only covered 50 addresses in California and therefore additional studies using data from other parts of the USA are needed. The researchers also suggested further research to evaluate other commonly used geocoding software, including Pitney Bowes’ MapMarker.

The goal of the present project was to evaluate and compare two commonly utilised geocoding software packages, ArcGIS 10 (Environmental System Research Institute (ESRI), Redlands, CA, USA) and MapMarker 22 (Pitney Bowes Business Insight, Troy, NY, USA), based on their features and performance in geocoding a large administrative health care dataset in New York State (NYS). There are no studies to our knowledge comparing the features and geocoding results between these two geocoding methods. In this project, we focused on the automated geocoding. Three metrics of

Corresponding author:
Sanjaya Kumar
Center for Environmental Health
New York State Department of Health, Corning tower Room 1203
Albany, NY 12237-0684, USA
Tel. +1 518 402 7960; Fax +1 518 402 7959
E-mail: sxk10@health.state.ny.us

evaluation were used for the comparison: (i) completeness (or match rate); (ii) similarity of geocodes; and (iii) positional accuracy. Completeness and positional accuracy are two common metrics that have been used for evaluating geocoding quality (Zandbergen, 2007, 2008). Similarity of geocodes, defined as the distance between assigned coordinates for each address by different methods, is a relatively new metric (Lovasi et al., 2007; Roongpiboonspoit and Karimi, 2010), and reported data on it has been very limited. Measurement of similarity can be used as an alternative and probably more efficient way to confirm positional accuracy of geocoding results (Roongpiboonspoit and Karimi, 2010). The MapInfo Spider Graph tool was used to measure both similarity and positional accuracy in this project.

Various methods have been used in previous research to estimate positional accuracy of geocoded addresses. For example, studies used a global positioning systems (GPS) unit that obtains precise geodetic locations of addresses from satellites (Ward et al., 2005; Zhan et al., 2006; Schootman et al., 2007). This method is time-consuming and may not be feasible in the absence of a field survey. One can also utilise aerial photography as the “gold” standard, which has been less frequently used than GPS measurements (Schootman et al., 2007). For example, Cayo and Talbot (2003) defined the “true” location of each address as the point that was the visual centre of the house using 1 m resolution digitally enhanced aerial orthoimagery with a horizontal accuracy of 10 m. The determined points were manually placed in the centre of the structure. The researchers reported some problems in identifying the visual centre for closely spaced homes, houses with dark rooftops, houses surrounded by dense canopy cover from trees and houses with a large outbuilding.

A third method is the utilization of the centroid or boundary of the land parcel. Parcel boundaries are traditionally considered as the most accurate spatial data with address information. Geocoding against parcels has now become more prevalent thanks to the development of parcel-level databases in the USA (Zandbergen, 2008). In order to study the effect of positional accuracy of street geocoding on traffic pollution exposure, Zandbergen (2007) measured the distance between the street-level geocoded point and the centroid of the associated parcel, which was thought to be more accurate than using the property boundary to determine positional error. This method was adopted in the present project. The accuracy of parcel data in the Capital Region of NYS was confirmed in the

study by Cayo and Talbot (2003). In general, position errors measured using different approaches ranged from 38 to 75 m in previous studies, and the errors were found to be higher in rural areas than in urban areas (Zandbergen, 2007).

This article provides a multifaceted comparison of two commonly used geocoding methods, with the objective of better understanding their features, performance and limitations. The methods and results in this paper can help researchers to decide which geocoding method best suits their geocoding needs and resources and also which approach assists them in creating a customised geocoding procedure. In addition, this paper provides practical information such as software settings associated with each geocoding method. One of the advantages of this research is that we used two sets of reference data for each geocoding method, which greatly improved the geocoding results.

Materials and methods

Address data

The address data used in this project consisted of 551,798 addresses randomly selected from the 1999 to 2006 Statewide Planning and Research Cooperative System (SPARCS) database in NYS. SPARCS is a comprehensive data reporting system that collects data on billing and medical record information for patients from all acute care hospitals in NYS. Because of the nature of the system, postal addresses rather than physical street addresses were collected from patients, resulting in problematic addresses in the sample data such as postal box numbers, rural route addresses, or nearest street intersection. The database contained fields for street address, city, state and ZIP code.

Reference data for geocoding

For MapMarker, both the default Address Dictionary (data vintage July 2009, MapMarker Version 22.0.0.13) and the 2010 NAVTEQ Address Points data set were used as the reference data in this study. The latter was obtained from the NYS Geographic Information Systems (GIS) Clearinghouse (<http://www.gis.ny.gov>) that is operated by the NYS Office of Cyber Security. Geocoding using address points reference data results in excellent positional accuracy, match rates only slightly lower than those for street geocoding, and a low number of ties (Zandbergen, 2008).

To use the geocoding tools in ArcGIS, reference data were first prepared by creating address locators in ArcCatalog. Two sets of reference data were prepared and used for geocoding with ArcGIS in the study. One was also the 2010 NAVTEQ Address Points data set, and the other was the 2010 NYS Street Segment data that were also provided by the NYS GIS Clearinghouse.

Geocoding process

Geocoding with MapMarker was based on an established geocoding protocol in the NYS Department of Health Center for Environmental Health. Two geocoding passes were completed for this study. In pass one, only the NAVTEQ Address Points data were used as the reference data. The geocoding precision was set to be at the street level (the highest), with exact match on house number, street name and ZIP code being required. Other geocoding parameters were set as follows: offset from road of 15 m, offset from corner of 50 m, no fallback to ZIP centroid or geographic centroid, no match on a multiple match. In the second pass, both the MapMarker Address Dictionary and the NAVTEQ Address Points data were used as the reference data, and the geocoding parameters were the same as in the first pass. Those addresses not successfully geocoded in pass one or receiving a result code between S1 and S4 were geocoded in pass two. Result codes are output by MapMarker during each geocoding process and provide information on the accuracy of the assigned points. The code generally has a range between S1 and S8; a higher number indicates a higher positional accuracy.

Before using the geocoding tools in ArcGIS, two individual address locators were first created in ArcCatalog using the NAVTEQ Address Points data and the NYS Street Segment data. In order to take advantage of both reference data sets in one geocoding process, a composite address locator was also created. With a composite address locator, addresses are matched against each of the composite address locator's individual addresses. The composite address locator attempts to match the first individual address and, if no match is found, it attempts to match against the second address. The same geocoding parameters were set for both individual address locators: spelling sensitivity of 80 and minimum match score of 85 (out of 100, i.e. the default settings of ArcGIS), side offset of 15 m, end offset of 50 m, and no match if there were multiple matches.

Data analysis

For each geocoding method, descriptive information on the geocoding process and results was recorded (e.g. reference data used, match rate). To estimate the similarity of assigned geocodes by the two methods, we first used SAS version 9.2 (SAS Institute Inc.; Cary, NC, USA) to randomly select 5% of the addresses that were geocoded by both methods from Albany, Niagara, Jefferson and New York counties. These counties were selected to represent suburban, rural, and urban counties in NYS. For each sampled address, we then used the Spider Graph tool in MapInfo to calculate the distance between the points assigned by the two methods. Furthermore, in order to determine which geocodes were more accurate, for addresses with a calculated distance of ≥ 100 m, we also used the Spider Graph tool to calculate the distances between the assigned points and the "true" location of the address, defined as the centroid of the parcel associated with the address.

Results

Match rate

Of the 551,798 input addresses, 318,302 (57.7%) addresses were geocoded by MapMarker, and 420,813 (76.3%) addresses were geocoded with the ArcGIS composite address locator. A total of 298,288 (54.1%) of the addresses were geocoded by both methods, and 440,827 (79.9%) addresses were geocoded by at least one method (Table 1).

In the first pass of geocoding with MapMarker, 203,449 (36.9%) addresses were matched, and 114,853 (20.8%) addresses were geocoded in the second pass. Of the addresses geocoded by MapMarker, 307,009 (96.5%) of them had a result code between S5 and S8, indicating that the assigned locations were

Table 1. Match rates and result scores.

Geocoding software	Match rate (%)	Result code/score
MapMarker	57.7	96.5% had a result code between S5 and S8
ArcGIS	76.3	Mean match scores of 95.6
By both methods	54.1	
By at least one method	79.9	

accurate. ArcGIS provides numeric match scores (up to 100) for geocoded addresses. The mean of the match scores in this study was 95.6, with a standard deviation of 6.0. The minimum match score was 85, which had been decided by one of the geocoding settings. A total of 264,246 (62.8%) of the addresses matched by ArcGIS had a match score of 100. In addition, the ArcGIS output was found to include the name of the individual address locator that the address was geocoded against; 72.5% of the geocoded addresses were matched against the NAVTEQ Address Points data, and the rest of them were geocoded against the NYS Street Segment data.

Similarity of geocodes

We found that the two geocoding methods assigned highly consistent geocodes. Among the 160 addresses reviewed for Albany county, 133 (83%) were assigned an identical position by both methods. For 13 (8%) of these addresses, the positional difference was between 1 and 5 meters, and 2 (1%) of the addresses had a distance greater than 100 m (Fig. 1A). When a total of 146 addresses were compared for Niagara county, 133 (91%) were assigned an identical position and 3 (2%) of the addresses had a distance of assigned locations greater than 100 m (Fig. 1B). A hundred and eight

addresses were compared for Jefferson county, a rural county in NYS. Among these, 91 (84%) were assigned the same location by the two methods. However, 5 (5%) of the compared addresses had a positional difference greater than 100 m (Fig. 1C). New York county had a much larger sample due to its high population density. Among the 1,115 compared addresses, 694 (62%) were assigned identical locations and 295 (26%) of them had a small difference between 1 and 5 m. Only 10 (1%) of the New York county addresses had distances between the two sets of assigned coordinates greater than 100 m (Fig. 1D).

Positional accuracy

For addresses with a calculated distance of greater than 100 m between the geocodes assigned by the two methods, we further investigated which assigned location was the more accurate, or closer to the parcel centroid ("true" location). Fig. 2 illustrates the distance calculated between the assigned points by the two methods (yellow line) and the distances between each of them and the associated parcel centroid (red lines). The length of the yellow line was calculated using the Spider Graph tool for estimating the similarity of assigned points, and the length of the two red lines were compared to determine which method generated

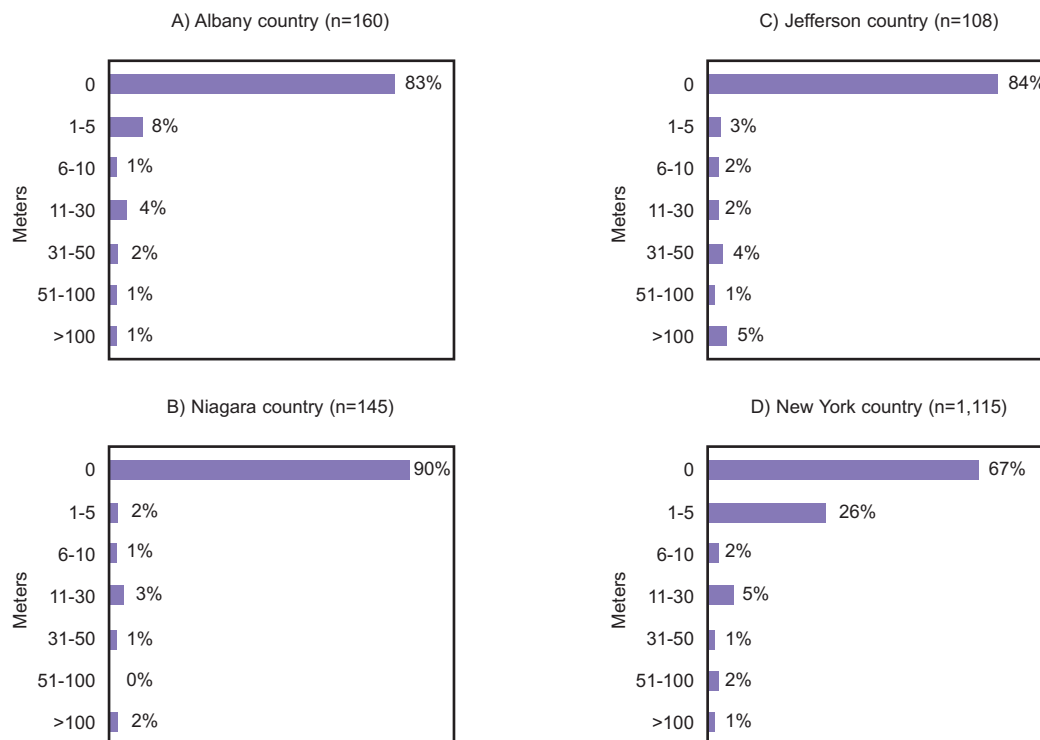


Fig. 1. Distances between the coordinates assigned by MapMarker and ArcGIS in four New York State counties.



Fig. 2. Distance between the assigned points and the distances between each assigned point and the associated parcel centroid.

the most accurate results for the address.

Because parcel data were not available, we could not compare addresses in New York county. We were able to investigate 8 of the 10 eligible addresses with parcel data available from the other three counties using the Spider Graph (Table 2). The third column in Table 2 lists the calculated distances between a MapMarker point and an ArcGIS assigned point. The next column shows the distances between a MapMarker point and an associated parcel centroid. Similarly, in the next column are the distances between an ArcGIS point and an associated parcel centroid. The result codes output for the address by each geocoding method, MapMarker georeference and ArcGIS match score, are also shown in Table 2. The results showed a consistent

pattern for all the examined addresses. The point assigned by ArcGIS was much closer to the associated parcel centroid when compared with that assigned by MapMarker. Most of the ArcGIS assigned points were located extremely close to the parcel centroid with a distance of only 2 m or less.

Discussion

Health researchers are increasingly using geocoding to display health-related information on maps and for spatial analyses. Public health departments in the USA often use geocoding to investigate disease outbreaks and clusters, or assign health records to appropriate geographic units such as ZIP code and county for dis-

Table 2. Comparison of geocoded points with the associated parcel centroid.

County	Address	MapMarker <i>vs.</i> ArcGIS	ArcGIS <i>vs.</i> parcel	ArcGIS <i>vs.</i> parcel	MapMarker Georeference	ArcGIS match score
Albany	Sample address 1	112	105	14	S5	100
Albany	Sample address 2	1,689	1,830	149	S2	87
Niagara	Sample address 3	177	176	2	S5	100
Niagara	Sample address 4	706	707	2	S1	87
Niagara	Sample address 5	121	121	1	S5	100
Jefferson	Sample address 6	7,852	7,889	50	S5	100
Jefferson	Sample address 7	245	245	0.3	S5	89
Jefferson	Sample address 8	109	110	1	S5	89

play (Washington State Department of Health, 2008). In this study, we evaluated match rates, similarity and positional accuracy of two commonly used geocoding software packages, ArcGIS 10 and MapMarker 22, in geocoding a large number of residential addresses from health administrative data in NYS. This study compares the features and performance of these two geocoding methods in an effort to help researchers decide which geocoding method best suits their geocoding needs and resources.

Our findings suggest that ArcGIS provides both a higher match rate and better positional accuracy of geocoding compared with MapMarker. In this study, geocoding with ArcGIS took approximately one third of time required by the MapMarker package. The results showed a high similarity between the geocodes assigned by the two methods, especially in suburban and urban areas. Both methods can be used together to maximise match rates and also to check the validity of certain geocoded locations. Attention should be focused on those addresses with large positional difference between assigned points (Duncan et al., 2011). In practice, however, many researchers would not have both software packages and would have to choose one that suits the nature of the project and available resources. We discuss the advantages and disadvantages of using ArcGIS and MapMarker for automated geocoding below.

MapMarker is a stand-alone geocoding programme that is licensed for an annual fee. A default reference data set is included in the package, which is critical to users that do not have access to better reference data. To display the addresses geocoded by MapMarker on a map, however, one also needs to own MapInfo and complete additional steps to create the points. The geocoding tools in ArcGIS are an integrated part of the software package. Therefore, there will be no additional cost for using those geocoding tools for those who already own ArcGIS and use it for other GIS purposes. Geocoded addresses are automatically shown on a map after the end of a geocoding process in ArcGIS. Nonprofit organisations can also take advantage of the discount prices offered by ESRI. On the other hand, ArcGIS does not provide default reference data and therefore users need to have access to or purchase necessary reference data sets in order to use ArcGIS for geocoding. ArcGIS also provides users with free online geocoding services that utilise reference data from TomTom (<http://www.tomtom.com>), but limits users to a maximum of 1,000 stored batch geocodes (ESRI: World Geocoding Task Service).

In addition to the default Address Dictionary, MapMarker allows users to add their own user dictionary and to utilise both reference data sets in one geocoding process. The composite address locator in ArcGIS, however, allows multiple sets of user defined reference data sets to be included in one geocoding process. ArcGIS allows users to fully utilise their data resources which consequently improve both match rates and the positional accuracy of geocoding. Our pilot tests confirmed that the composite address locator resulted in the highest match rate (76.3%) compared with the other two individual address locators using only the NAVTEQ Address Points (55.3%) data or the NYS Streets Segment (67.2%) data. However, creating an address locator in ArcGIS requires more user knowledge and preparation time compared with using the Address Dictionary or User Dictionary in MapMarker.

The two software packages provide users with different options to customise the geocoding process. With MapMarker, one can require exact match on certain address information including house number, street name, city name, and ZIP code. Many epidemiologic studies require highly accurate geocoding results for further analyses, and therefore it would be necessary to set up exact matching on most components of the address. In ArcGIS, the minimum match score represents the degree of agreement between the address being geocoded and its actual location in the reference data, which can be adjusted according to the required accuracy of geocodes. Pilot studies can be conducted prior to the main study to determine an adequate minimum match score (Zandbergen, 2008; Duncan et al., 2011). The setting of spelling sensitivity in ArcGIS affects the number of address candidates considered by ArcGIS by controlling the amount of variation allowed in searching for potential candidates in the reference data. It does not, however, affect the match scores of address candidates (Zhan et al., 2006). Generalizability of results in this comparison to other scenarios with different parameter settings of the two methods requires further research.

One limitation of this study is that the findings are based on NYS addresses only and the evaluation of similarity and positional accuracy are limited to four counties in NYS. Future testing should include representative data from other parts of the USA to confirm the generalizability of our findings to other jurisdictions. Another limitation is that a substantial portion of the input data could not be geocoded using either of the two methods. The low quality of the input addresses was probably a major contributing factor.

This problem may have introduced selection bias in comparing the two geocoding methods because the results may be different when more standardised addresses were used for comparison. A final limitation of this study is that only eight addresses were investigated to determine the accuracy of assigned geocodes by the two methods. We considered 100 m as a meaningful threshold after reviewing previous literature in the field (Ward et al., 2005; Zandbergen, 2007). Future research could use smaller thresholds and/or larger sample sizes for a more thorough comparison of positional accuracy between the two geocoding methods.

Conclusions

ArcGIS generated better geocoding results when compared with MapMarker. Locations assigned by ArcGIS are probably more accurate than those from MapMarker, but overall the positional differences found with the two geocoding methods were minimal and a large majority of addresses were placed at the same locations by the two methods. For projects involving a large number of addresses, using both methods and combining the results should maximise match rates and limit the amount of time needed for manual geocoding. When reference data of high quality are available, ArcGIS is the better choice and the composite address locator is especially useful to improve the quality of geocoding results.

Acknowledgements

We thank Mr. Thomas Talbot for his guidance during this research and Ms. Gwen Babcock for assistance in obtaining the reference data. This research is supported by the US Centers for Disease Control and Prevention Environmental Public Health Tracking Grant # 5U38EH000184. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

References

Cayo M, Talbot T, 2003. Positional error in automated geocoding of residential addresses. *Int J Health Geogr* 2, 1-12.
Duncan D, Castro M, Blossom J, Bennett G, Gortmaker S,

2011. Evaluation of the positional difference between two common geocoding methods. *Geospat Health* 5, 265-273.
ESRI: World Geocoding Task Service. Available at: <http://www.esri.com/software/arcgis/arcgis-online-map-and-geoservices/geoservices> (accessed on January 2012).
Healthy People 2020 Summary of Objectives, 2010. Available at: <http://healthypeople.gov/2020/topicsobjectives2020/pdfs/EnvironmentalHealth.pdf> (accessed on January 2012).
Krieger N, Waterman P, Chen J, Rehkopf D, 2012. The public health disparities geocoding project monograph. Available at: <http://www.hsph.harvard.edu/thegeocodingproject> (accessed on January 2012).
Lovasi G, Weiss J, Hoskins R, Whitsel E, Kenneth R, Erickson C, Psaty B, 2007. Comparing a single-stage geocoding method to a multi-stage geocoding method: how much and where do they disagree? *Int J Health Geogr* 6, 6-12.
NYSGIS Clearinghouse. Available at: www.gis.ny.gov (accessed on January 2012).
Roongpiboonsopit D, Karimi H, 2010. Comparative evaluation and analysis of online geocoding services. *Int J Geogr Inf Sci* 24, 1081-1100.
Schootman M, Sterling D, Struthers J, Yan Y, Laboube T, Emo B, 2007. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Ann Epidemiol* 17, 464-470.
Swift J, Goldberg D, Wilson J, 2008. Geocoding best practices, review of eight commonly used geocoding systems. Available at: http://spatial.usc.edu/Users/dan/gislabtr10_Eight-Commonly-Used-Geocoding-Systems.pdf (accessed on January 2012).
Ward M, Nuckols J, Giglierano J, Bonner M, Wolter C, Airola M, Mix W, Colt J, Hartge P, 2005. Positional accuracy of two methods of geocoding. *Epidemiology* 16, 542-547.
Washington State Department of Health, 2008. Guidelines for address matching and geocoding. Available at: http://ww4.doh.wa.gov/gis/geocoding_guideline.htm (accessed on January 2012).
Zandbergen P, 2007. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health* 7, 7-37.
Zandbergen P, 2008. A comparison of address point, parcel and street geocoding techniques. *Comput Environ Urban* 32, 214-232.
Zhan F, Brender J, DE Lima I, Suarez L, Langlois P, 2006. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Ann Epidemiol* 16, 842-849.