Central Cancer Registry Geocoding Needs

John P. Wilson, Daniel W. Goldberg, and Jennifer N. Swift

GIS research laboratory

Technical Report No. 13

Ecological Consequences of Artificial Night Lighting
A Geocoding Best Practices Guide
The Handbook of Geographic Information Science
Geographic Information Systems and Public Health
Eliminating Perinatal Disparity
Visioning M

Table of Contents

Executive Summary

This technical report delineates the core requirements for a Central Cancer Registry geocoding solution based on the results of a series of geocoding user surveys and our own experience working on geocoding applications with central cancer registries over the past 2-3 years. These requirements are grouped under seven headings for ease of description and because many of the elements of the geocoding process are connected to one another and collectively affect the character and quality of the computed outputs. The bottom line is that geocoding is a complicated process and the health researchers and personnel performing geocoding and/or using geocoding results will need better software tools, geospatial datasets, and training to improve this aspect of their work over the next 5-10 years.

1. Introduction

The geocoding process is critical in many scientific arenas as it is typically one of the first steps used to create the spatial data employed in subsequent spatial analyses. Accordingly, the accuracy, granularity, and reliability of geocoded data are of paramount importance in health-related research projects and activities that use address data as their underlying spatial data source. To this end, the USC GIS Research Laboratory has conducted several surveys of geocoding best practices and needs in health-related research and developed a scalable, reliable, accurate and extensible geocoding platform for use in the academic and larger scientific communities during the past 2-3 years.This information will be used to document the core requirements for a Central Cancer Registry (CCR) geocoding solution in the following sections of this report.

That said, this is the fourth in a series of four technical reports on one or more aspects of geocoding commissioned by Northrop Grumman (NG) at the end of 2008. The original scope imagined that NG and the Division for Cancer Protection and Control (DCPC) would conduct a CCR Needs Assessment and that we would work collaboratively with both of those organizations to: (1) synthesize the results of the Geocoding Requirements Analysis that we conducted in 2008 [see Goldberg et al. (2008) for additional details] and the CCR Needs Assessment; and (2) use the findings from these two projects to prepare a report delineating recommendations for the types of software tools and protocols that will be required to support the geocoding work of central cancer registries over the next 5-10 years. This scope was subsequently modified because NG and the DCPC chose not to conduct the aforementioned CCR Needs Assessment.

What follows then is a report delineating the requirements for a CCR geocoding software solution that draws on our past work – this includes the surveys we have been involved with, the experiences we have accumulated providing geocoding services to multiple cancer registries, and our previous publications (e.g. Goldberg et al. 2007). The surveys used include the Geographic Information Systems (GIS) Survey conducted by the North American Association of Central Cancer Registries (NAACCR) GIS Committee in 2005 (NAACCR 2008), the Geocoding Best Practices Survey conducted by the University of Southern California (USC) GIS Research Laboratory in 2006 (Goldberg 2008a, b), the Geocoding Capacity Survey conducted by the USC GIS Research Laboratory in 2008 (Goldberg et al. 2008a, b), and our initial attempt to analyze and describe the geocoding user requirements for central cancer registry researchers and staff (Goldberg et al. 2009). The 2008 survey provides a useful introduction because it: (1) lists the organizations that were surveyed; (2) describes the types of questions that were asked in all three of the aforementioned surveys; and (3) presented a rich account of current practices and as a consequence, documented the large variety of geocoding approaches, references datasets, etc. that are currently deployed by the cancer registries and the large variations in the knowledge of geocoding challenges, tradeoffs, etc. among the scientists and staff to whom the geocoding tasks are assigned.

The remainder of this report is organized as follows. The next section (Section 2) offers a high level description of the geocoding process. This is important because there are many components as well as many inputs and decisions that must (should) be made when turning addresses into geographic coordinates (i.e. the process of geocoding). Section 3 documents what the geocoding solutions will need to do to better serve the geocoding requirements of central cancer registries. These ideas are grouped under a series of headings (seven in all) that build on the results compiled by Goldberg et al. (2009) and speak to the complexity of the process and the interdependencies among the many components. The fourth and final section offers some general observations and draws some conclusions in terms of what is likely to happen in the immediate future.

2. The Geocoding Process

Geocoding is most commonly considered to be the process of converting a location description such as a street address into some form of geographic representation such as geographic coordinates (latitude and longitude). The main components of a postal address geocoding system include address parsing, reference data set definition and storage, feature matching, and feature interpolation as illustrated in Figure 1. Hence, such a system will usually accept input data supplied by a user in the form of an unparsed street address and a city and/or USPS ZIP code combination. The input street address is first parsed and normalized to identify standard values for each of the postal address components. After normalization, the system attempts to find one or more reference features that match the input address from within each of the reference data layers that it maintains. If the system is able to obtain a matching reference feature, feature interpolation is performed to determine an appropriate output location within or along the reference feature based on the input address.

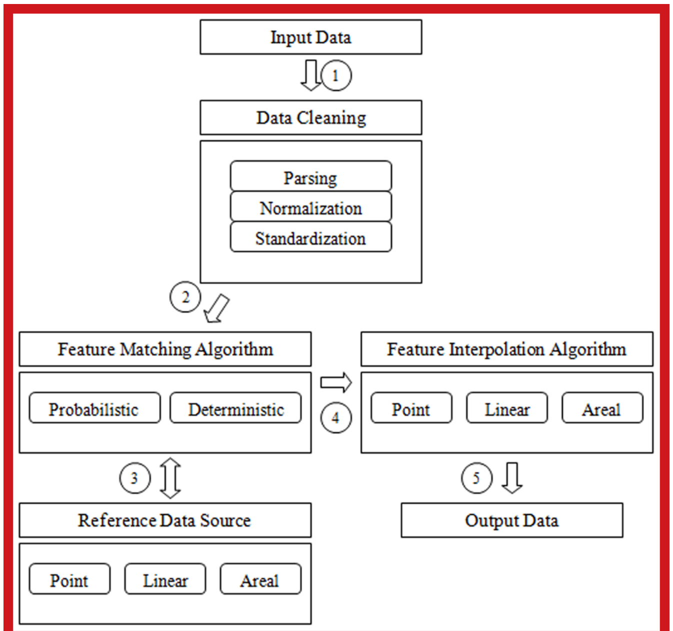

Figure 1 – Main components of a typical geocoding system

3. Central Cancer Registry Geocoding Needs

The geocoding process incorporates three main processing components (address processing, feature matching, and feature interpolation) and one or more sets of underlying geographic reference datasets. The current geocoding solutions that are available use a wide variety of approaches and options for each of these components and the subsections which follow describe what is needed to build a robust and yet versatile geocoding platform that serves the needs of central cancer registries under seven headings.

3.1 Input Data

The best geocoding solutions will support many different types and formats of input data such as the USPS Publication 28 (2009) and Urban and Regional Information Systems Association (2009) address standards and maintain sets of rules for automatically identifying types of input data. By doing this, the best geocoding solutions will be able to take advantage of diverse address formats by processing them differently using specialized approaches. This is a key capability given the large variety of systems used by health agencies and providers to collect patient data and lack of interest in utilizing address validation software to help clean the address data at source.

3.2 Address Cleaning

To ensure compatibility geocoding solutions will also need to be able to standardize the address data to the Urban and Regional Information Systems Association (2009) address standard. In addition, to ensure reliability, the standardization algorithm should be CASS certified and support a base set of deterministic operations (rules), and support probabilistic feature matching approaches as documented in the following subsection. Finally, the best geocoding solution will provide some guidance on which supplemental sources of data to use for address clarification, and how to incorporate them into the address record while noting their inclusion in the metadata that is produced as part of the output of the geocoding process.

3.3 Feature Matching Algorithm(s)

The ideal geocoding solutions will support a large range of geocoding options in terms of the types of geocoding that can be performed (manual, interactive, interactive with prompting, single address, batch mode, etc.) and the types of components that are supported. The latter includes the types of data sources that can be used (linear and areal data features) and/or the forms of matching and interpolation that are supported (feature matching only, feature interpolation, etc.). Ideally, the user should be able to control the geocoding process in terms of the types of components that are employed (data sources, interpolation methods, etc.) and the order in which they are applied so they can control the hierarchy that is used.

There are lots of choices and subtleties at play here. Both deterministic and probabilistic feature matching algorithms must be supported and the user must have the ability to decide which algorithm to use when and to change algorithms from one record to the next. The user must be able to turn on and off the capability of manually breaking feature matching ties, and the criteria may include some prescribed as well as user-defined rules. The ability to use attribute relaxation approaches must be included, and the user must be able to specify which attributes should be relaxed in which order, if at all. The user must be able to choose whether or not to use phonetic algorithms such as SOUNDEX, and the uncertainty cutoffs for probabilistic matching must be user-definable.

Looking beyond these requirements, the best geocoding solutions will also support manual, linear- and areal-based interpolation methods. These software solutions will support a consistent standardized protocol for performing manual geocoding and user customization in terms of the reference data sources that can be used. The linear-based interpolation algorithm(s) must be able to support the inclusion of additional information to overcome the assumptions present, such as the number and sizes of parcels along a road segment, and user-defined and modifiable offsets (i.e. dropback distances from the road centerlines). The areal-based interpolation algorithms should support centroid calculations for deriving the most likely geographic coordinates and sub-parcel matching to parcels and building footprints. This final class of feature matching is likely to grow rapidly in importance as more and more geospatial datasets and remotely sensed imagery are provided on the web for little or no cost over the next few years.

Last but by no means least, a strong argument can be made that all central cancer registries use: (1) the same attribute relaxation hierarchy; (2) the same components (address parsing, reference data set definition and storage, feature matching, and feature interpolation); and (3) a single consistent match score for probabilistic matching to guarantee the standardization of feature matching methods.

3.4 Reference Data Sources

A flexible geocoding solution will support multiple data sources (i.e. different types of reference data) simultaneously and allow the user to switch between them, so different sources can be used as deemed appropriate during the geocoding process. The software should be able to switch between sources on a per-record basis following one or more sets of predetermined criteria, or automatically based on the spatial extent and/or some preferred or user-defined accuracy requirements. The best geocoding solution will also offer users, such as cities and counties, the opportunity to utilize their own reference data, and the output should include metadata that specifies the rationale for using one or more reference datasets to geocode individual addresses.

3.5 Output

It is critical that geocoding solutions be able to produce outputs that serve a wide variety of consumers. These solutions should be able to output geographic points with appropriate metadata that describes the geocoding process as well as the accuracy of the computed coordinates, and they should be capable of generating output in the form of text files, ESRI shapefiles, ESRI geodatabases, and various forms of non-spatial databases including GML and KML. The best solutions will report match rates along with the geographic points (i.e. latitude/longitude coordinates).

The best geocoding solutions will also be able to derive accuracy metrics from both the reference features and the other geocodes by utilizing measures of completeness and the spatial accuracy of the attributes in the reference datasets, as well as accuracy measures for individual regions and information as to how accuracy varies across regions. These solutions must also be able to utilize multiple feature matching hierarchies that can be user-defined and selected, and they must be able to report the feature match type and the hierarchy used in feature matching as a part of the metadata.

Finally, a geocoding to meet the need of CCRs must report accuracy metrics for the whole process, for each component, and for each individual geocoding result. The accuracy metrics should express the probability of a correct match based on the supports for the match (e.g. percentage of attributes matched, percentage of attributes relaxed, etc.) and be able to derive and report estimates of the spatial uncertainty or error based the metadata summarizing the key components of the geocoding process (probability that a feature matched correctly, area of the matched feature, etc.).

3.6 Software Usability

A CCR geocoding solution must be flexible and customizable and built in such a way that multiple geocoding strategies can be supported simultaneously. In addition, the authors and vendors need provide detailed metadata about the internal workings of their geocoding processes to clarify what components are used and the lineage, vintage and spatial accuracy of the reference datasets that are employed.

3.7 Confidentiality Concerns

The confidentiality and privacy of the information contained in the geocoded data must be protected, and a variety of methods should be provided to accomplish these outcomes. Hence, both individual level and geographically masked geocodes must be provided to consumers, and the forms of geographic masking supported should include randomization, aggregation and resolution lowering. These capabilities, taken as a whole, must be capable of ensuring both the physical and logical security of the geocoded data at all times.

4. Discussion and Conclusions

The preceding section has pointed to the most important characteristics for geocoding solutions that would serve the present-day needs of the central cancer registries. No systems currently provide these capabilities and there would be additional needs and challenges even if they did. The central cancer registries would need to document their geocoding practices and develop a series of training programs and support documents (user guides, online help systems, etc.) that would allow them to build and sustain a sophisticated workforce that could use these new tools and make the appropriate decisions based on the methodologies and data sources at hand. To this end, the provision of state-of-the-art geocoding software would constitute a necessary first step to improving the geocoding solutions currently employed by the various central cancer registries scattered across the United States and Canada.

5. References

Goldberg D W (2008a) Geocoding Best Practices Survey. WWW document, http://ahf410-pc4.usc.edu:8080/NAACCR/survey.jsp

Goldberg D W (2008b) A Geocoding Best Practices Guide. Springfield, IL, North American Association of Central Cancer Registries

Goldberg D W, Knoblock C A, and Wilson J P (2007) From text to geographic coordinates: The current state of geocoding. Journal of the Urban and Regional Information Systems Association 19(1): 33-46

Goldberg D W, Swift J N, and Wilson J P (2008a) Geocoding Capacity Survey. WWW document, http://webgis.usc.edu/Surveys/CDC/

Goldberg D W, Swift J N, and Wilson J P (2008b) Geocoding Best Practices: Analysis of Geocoding User Requirements. Los Angeles, CA, USC GIS Research Laboratory Technical Report No. 9

NAACCR (2008) Geographic Information Systems Survey. WWW document, http://www.naaccr.org/filesystem/word/GIS%20survey_Final.doc

Urban and Regional Information Systems Association (2009) Draft Street Address Data Standard. Park Ridge, IL: Urban and Regional Information Systems Association

U.S. Postal Service (2009) Publication 28: Postal Addressing Standards. Washington, D.C., United States Postal Service

The University of Southern California GIS Research Laboratory seeks to develop cutting edge geographic analysis tools and to apply those tools in ways that increase our knowledge of the built and natural environments while training the next generation of geographic information scientists and promoting the utilization of geographic information science concepts and technologies throughout the academy.

To learn more about our research and teaching programs, contact Leilani Banks, GIS Research Laboratory, University of Southern California, 3620 South Vermont Avenue, Los Angeles, CA 90089-0255

GIS research laboratory http://gislab.usc.edu