Geocoding Best Practices: Analysis of Geocoding User Requirements

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

Prepared for: Division of Cancer Prevention and Control
National Center For Chronic Disease Prevention and Health Promotion
Center for Disease Control and Prevention

Cover Photos:
Maps and addresses (University of Southern California GIS Research Laboratory, Department of Geography, 2008).

Preferred Citation:
Goldberg DW, Swift JN and Wilson JP 2008 Geocoding Best Practices: Analysis of Geocoding Requirements. Los Angeles, CA, University of Southern California GIS Research Laboratory Technical Report No 9.

Table of Contents

List of Figures

List of Tables

Executive Summary

The purpose of this report is to provide a detailed list of geocoder user requirements based on a review of past, recent and emerging geocoding technologies to the Division of Cancer Prevention and Control (DCPC), Centers for Disease Control and Prevention (CDC), and to examine trends in these requirements such that the best possible recommendations are made for the future. To accomplish this, three geocoding surveys of the cancer registry and research communities are analyzed and synthesized to tease out the common needs of the cancer registries. This is the second in a series of three reports which documents geocoding best practices for the DCPC and CDC.

The surveys used include a *Geographic Information Systems (GIS) Survey* conducted by the North American Association of Central Cancer Registries (NAACCR) GIS Committee in 2005 (NAACCR 2008b), a *Geocoding Best Practices Survey* conducted by the University of Southern California (USC) GIS Research Laboratory in 2006 (Goldberg 2008a), and a follow up *Geocoding Capacity Survey* also conducted by the USC GIS Research Laboratory on behalf of NGC and the CDC in 2008 (Goldberg et al. 2008b). While each of these surveys and respondent sets are unique, taken together they serve to capture distinct snapshots of the current geocoding practices at moments in time throughout the past several years. Each of the surveys also surveyed (for the most part) a separate set of cancer registries and cancer-related organizations. In combination, these distinct user groups provide a comprehensive view of the many different opinions and needs present throughout the diverse cancer community.

A set of minimal geocoding needs is derived from the results of these three surveys in this particular report. This list of needs is by no means complete with regard to the specific needs of any specific registry. These needs represent what should be minimally included in a geocoder to serve the largest possible audience in the best possible manner. These are based upon what has been successful for those using geocoding processes in the past, and trends that have been identified as the most probable pathways the cancer registries will want to pursue in the future.

1 Introduction

The member registries of the Centers for Disease Control and Prevention (CDC) National Program of Cancer Registries (NPCR) and North American Association of Central Cancer Registries (NAACCR) continuously collect and utilize spatial information in epidemiological research. To carry out these research activities that require geospatial mapping and/or analysis, many cancer registries perform the process of geocoding or utilize third party services (vendors) to perform the geocoding for them. This process of geocoding typically requires input data, reference data, and a methodology for generating geocoded output data (Goldberg et al. 2008a).

In general, most of the input data required for geocoding is reported in the form of postal addresses which are usually collected at cancer diagnosis or treatment facilities. These data are subsequently submitted to and processed by individual cancer registries. Typical address information includes the street address, city, and province or state of a patient at the diagnosis of their disease (dxAddress, dxCity, dxState). Geocoding is typically performed using software systems called geocoders, the primary purpose of which is to convert textual descriptions of postal street addresses into valid, computer-usable geospatial data. The end result is that information which originally had no geographically computer-compatible reference can subsequently be used for spatial analyses in epidemiological research.

Presently, different registries have vastly different priorities in terms of geocoding services, dependent upon the individual goals or intended uses for the geocoded data, staffing, available resources, and the type of registry (e.g. level of government, size of registry). In order to best address current user needs and the most likely future requirements with respect to geocoding activities, this document assembles a list of minimized user requirements based on a review of all available surveys related to geocoding practice, both academic and anecdotal. Currently this information includes the *Geographic Information Systems (GIS) Survey* conducted in 2005 by NAACCR (NAACCR 2008b and c), the *Geocoding Best Practice Survey* conducted by the USC GIS Research Laboratory in 2006 (Goldberg 2008a) to support the preparation of the *Geocoding Best Practices Guide* (Goldberg 2008b), and the *Geocoding Capacity Survey* conducted in 2008 by the USC GIS Research Laboratory on behalf of Northrop Grumman Corporation (NGC) and the CDC during the creation of this document (Goldberg et al. 2008b).

By identifying what is most important to cancer registries in regards to geocoding services, this document will assist the CDC in their progress toward the development of a standardized and centralized geocoder, freely available to the cancer research community. Thus, the primary purpose of this research report is to present the results of past and recent geocoding related surveys of local, state and national cancer registries and cancer-related organizations. The results of these surveys are summarized as sets of priorities geared toward the level of geocoding expertise of the various registries that participated in the user requirements studies. The priorities, in turn, are summarized according to what geocoding services are actually used, what processes the services perform, who in fact carries out geocoding activities, and how the results are utilized. It is important to recognize that trends in user needs may be influenced in the near future by improvements in the quality and accessibility of reference data, advancements in geocoding methodologies, and increased utilization of geocoding across all levels of governmental, academia, and non-profit agency research and practice.

The geocoding user requirements developed and presented in this document identify what the most important aspects of the geocoding process are to cancer registries, noting that those who utilize geocoding services vary in experience from novices to advanced users. These requirements will ultimately be utilized to enhance the design and functionality of a standardized and centralized geocoder designed specifically to serve the needs of the cancer research community.

2 Cancer Registries Surveyed

The research results presented in this report are based on three separate geocoding surveys. The first two surveys were conducted to specifically survey NAACCR member registries, the first in 2005 by the NAACCR GIS Committee (NAACCR 2008b and c), and the second in 2006 by Goldberg (2008a). The third survey was focused on the CDC NPCR member registries, some of whom are also members of NAACCR. Documentation of the 2005 study can be obtained from the NAACCR website (NAACCR 2008b and c), while the results of the 2006 and 2008 surveys are original, previously unpublished observations.

The difference between the two NAACCR surveys is that the 2005 survey questions were more general in nature, while the 2006 survey asked specific, detailed technical questions of the respondents. A list of the individual cancer registries that participated in the 2006 survey is provided in Table 1. Represented in this list are organizations ranging in scale from local to national, some of which are commercial enterprises, federal government organizations such as the CDC and the National Cancer Institute (NCI), and professional as well as academic and philanthropic organizations. Note that some of the largest state cancer registries are included, i.e. the New York State Cancer Registry, as well as some of the smallest, i.e. the Alaskan Cancer Registry, and that both rural and urban regions are represented.

Table 1 Organizations that participated in the 2006 Geocoding Best Practices Survey

Name	Level of Government
Alaska Cancer Registry	State
American Cancer Society	National
American College of Surgeons	National
Baystate Medical Center	Local
California Cancer Registry	State
Cancer Data Registry of Idaho	State
CancerCare Manitoba	Provincial
Centers for Disease Control and Prevention	National
Florida Cancer Data Systems	State
IMPAC Medical Systems, Inc	Local/State/National
Massachusetts Cancer Registry	State
National Cancer Institute	National
North Carolina Cancer Registry	State
New Jersey Cancer Registry	State
New York State Cancer Registry	State

Name	Level of Government
University of Southern California	State
Wisconsin Division of Public Health - Cancer Reporting System	State

The *Geocoding Capacity Survey* (Goldberg et al. 2008b) was specifically developed to be compatible with the *Geocoding Best Practices Survey* (Goldberg 2008a). To do so, the questions present on the *Geocoding Capacity Survey* represent a subset of those from the original *Geocoding Best Practices Survey*. However, these questions were re-arranged and re-categorized to overcome limitations in the survey instrument design that became apparent upon an analysis of the results from the original *Geocoding Best Practices Survey*.

In contrast to those who responded to the *Geocoding Best Practices Survey* (Goldberg 2008a), the respondents that participated in the *Geocoding Capacity Survey* (Goldberg et al. 2008b) form a far more cohesive cohort (Table 2); they are all state level cancer registries. Of the 17 participants in this survey, three registries also participated in the *Geocoding Best Practices Survey* (Goldberg 2008a). Again note that these registries range in size from large, e.g. the California Cancer Registry, to small, e.g. the District of Columbia Cancer Registry, and that both rural and urban regions are represented across the continental US.

Table 2 Organizations that participated in the 2008 Geocoding Capacity Survey

Name	Level of Government
California Cancer Registry	State
Cancer Data Registry of Idaho	State
Colorado Department of Public Health and Environment	State
District of Columbia Cancer Registry	State
Florida Cancer Data Systems	State
Indiana State Department of Health	State
Louisiana State University School of Public Health	State
Maryland Cancer Registry	State
Michigan Department of Community Health	State
North Dakota Cancer Registry	State
Oklahoma State Department of Health	State
South Carolina Department of Health and Environmental Control	State
South Dakota Cancer Registry	State
Vermont Department of Health	State
Virginia Cancer Registry	State
Washington State Department of Health	State
Wyoming Cancer Surveillance Program	State

To develop the geocoder user requirements compiled in Section 7 an analysis was performed comparing the results of the NAACCR (2008b), Goldberg (2008a), and Goldberg et al. (2008b) surveys.

The following sections provide the detailed results for each of these surveys. Taken together, these surveys represent all available quantitative data regarding the usage and practices of geocoding in the US cancer registry community and other related organizations.

3 2005 NAACCR Geographic Information Systems Survey

According to the NAACCR website, the NAACCR GIS Committee was formed to "address the appropriate uses of geographic information systems (GIS) in cancer registry practice" (NAACCR 2008a), including the use of geocoding services and geocoded data. Consequently, the 2005 NAACCR *Geographic Information Systems Survey* (NAACCR 2008b) was created by the NAACCR GIS Committee specifically to assess the GIS capabilities and training needs of the NAACCR membership. This survey was posted on the NAACCR website and distributed to the membership via email. The results in this section present a compressed version of the original findings available in the *Geographic Information Systems Survey* (NAACCR 2008b).

Of the 72 NAACRR registries surveyed, a total of 45 participated in the survey. Forty-one were US and four were Canadian registries, giving an overall response rate of 63%. Specifically regarding geocoding, 82% of the respondents actually geocode patient addresses at the location of diagnosis. When asked who specifically performed the geocoding tasks, the 45 registries responded:

- Central registry staff (27%)
- Private vendor (26%)
- Non-registry staff within each organization (19%)
- A combination of central registry staff and others within their organization (19%)
- A combination of central registry staff and a private vendor (7%)

In terms of address quality control or cleaning prior to geocoding, 70% of the registries reported performing this task in-house. When asked specifically what address cleaning tasks are performed, the registries responded as follows:

- When a street address cannot be matched to a feature, the match is made to a centroid of a geographic area (i.e. town or ZIP code) (78%)
- Standardization of the street format (70%)
- When a street address cannot be matched to a feature, use manual interactive geocoding for feature matching (57%)
- Separate non-geocodable addresses (such as PO Boxes) from geocodable ones based on individual project criteria (54%)
- Manually adjust the matching criteria of their geocoding software (51%)
- Use address parsing [and normalization and standardization] techniques to increase the probability of an address match (such as changing "Northgate Way" into "N Gate Way") (46%)
- Perform an automated comparison of more than one geocode source (i.e. reference data-set)for a given address (27%)

According to the NAACCR 2005 *Geographic Information Systems Survey*, the number of geocodes attempted each year is 1.7 times the annual case load, on average. In addition, 16% of cases fail batch geocoding, and 7% cannot be geocoded at all, on average per year. In addition, 18% of the registries assign exiting geocodes to repeat incoming addresses, and 26% link external address data to their registry database for updating or confirming patient addresses.

4 2006 USC Geocoding Best Practices Survey

As previously stated, a *Geocoding Best Practices Survey* (Goldberg 2008a) was conducted in 2006 by the USC GIS Research Laboratory to aid in the preparation of the *Geocoding Best Practices Guide* (Goldberg 2008a and b). This survey was organized as a series of categories of questions and administered via a website hosted at USC. The survey was sent to 46 individuals representing 32 organizations. 20 individuals responded (an individual response rate of 43%), representing 17 organizations (an organizational response rate of 85%). The results of each are presented in the following sub-sections.

4.1 Reference Data Sources

In terms of reference data used in the geocoding process, eight of 10 respondents use road (vector) reference data sources, while two use imagery, and two use parcels (Figure 1). As for having control over the reference data sources used, four out of 10 reported having control, while four out of 10 reported not having control, and two out of 10 reported not knowing. Also, one out of nine respondents are able to control which reference data source was used by spatial extent and accuracy requirements. Six out of 10 respondents use multiple sources of reference data, and three out of 10 use only a single source of reference data. When asked if they choose reference data sources based on characteristics of completes, accuracy and temporal compatibility with input data, five out of nine respondents always choose the most complete reference data sets, while four out of nine always use the most accurate. One out of nine respondents uses the reference data that is temporally closest to the input data, whereas One out of nine simply let the software decide for them. Regardless of how they reported they *would* choose their data sources, only two out of seven of the respondents indicated that they know exactly how their geocoder chooses a reference data source, while two out of seven, the majority, do not know how the software chooses reference data sets.

In terms of the specific sources of linear reference datasets most often utilized by the respondents, five out of nine use TIGER/Lines data (US Census Bureau 2008), two out of nine use “Enhanced TIGER/Lines data”, Tele Atlas (Tele Atlas Inc. 2008) or NAVTEQ (NAVTEQ 2008), and two out of nine utilize local reference data sources (Figure 2). In addition, two out of eight respondents use historical reference data because their input data are historical. While seven out of eight of respondents indicated that they change or update reference data sources over time, six out of seven answered that they change sources as soon as new sources become available, and one out of seven indicated that they change only after a new source has been used and tested by others.

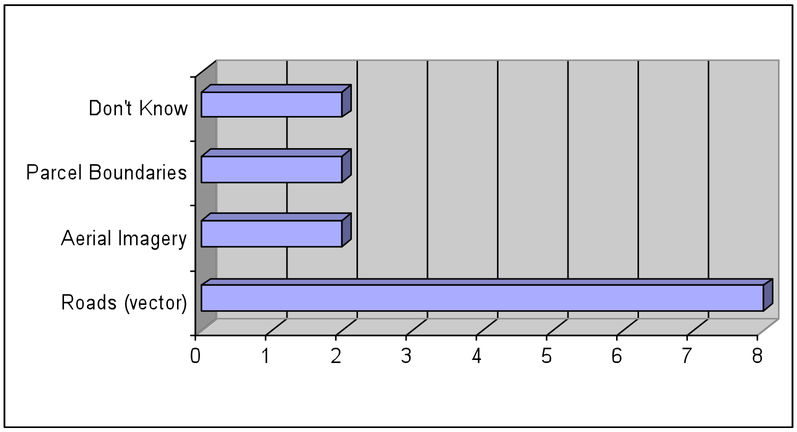

Figure 1 Different types of reference data sources used in geocoding

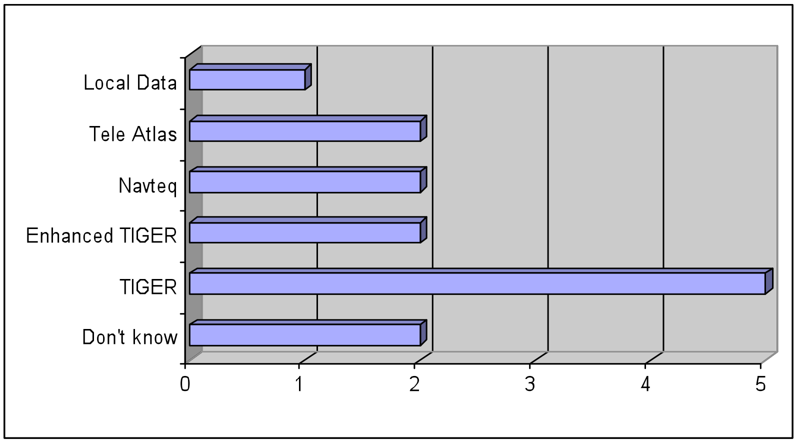

Figure 2 Different linear reference data sources used in geocoding

Subsequently, the survey asked the respondents to report what reference data sources they considered most versus least accurate, most versus least complete, and which of these metrics is more important (Figure 3). In regards to TIGER/Lines data, two out of seven respondents believe that it is the most accurate reference data source, whereas four out of seven feel it is the least accurate, and three out of seven consider it to be the most complete reference dataset, versus two out of seven that indicated it is the least complete. As for NAVTEQ, Tele Atlas and the "local data" reference sources of the participants, one out of seven respondents believes these datasets are among the most accurate and complete available. Concerning using parcels as reference data, two out of seven respondents feel that parcel data is the most accurate and one out of seven the most complete. Conversely, another two out of seven indicated that they believe it is the least complete. Lastly, multiple respondents indicated that they did not know which of these reference datasets are their most accurate (two out of seven), most complete (three out of seven), least accurate (four out of seven), or least complete (three out of seven).

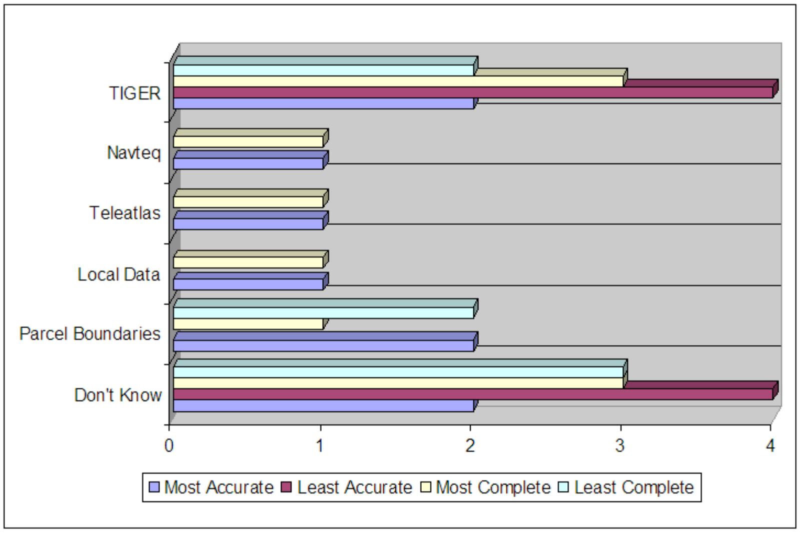

Figure 3 Perceptions of linear reference data source accuracy and completeness

4.2 Input Data Sources

Three out of 11 respondents allow relative spatial locations, postal ZIP code delineations, and census delineations as input data. In addition, four out of 11 encounter street intersections and eight out of 11 PO Boxes. Nine out of 11 respondents accept named places, rural addresses, addresses with unit indicators, and urban addresses as input data. (Figure 4). It is interesting to note that four out of eight respondents use the same geocoding method regardless of the type of input data, while three out of eight choose their geocoding methodology based on the type of input data (Figure 5). In terms of classification of input data types, one out of eight use an automated SQL technique to identify input data types, while four out of eight respondents classify input data types manually (Figure 6). And, regarding whether or not they process historical address data, four out of seven of respondents accept historical address data, and one out of four handle historical addresses differently than they do current address data (Figure 7).

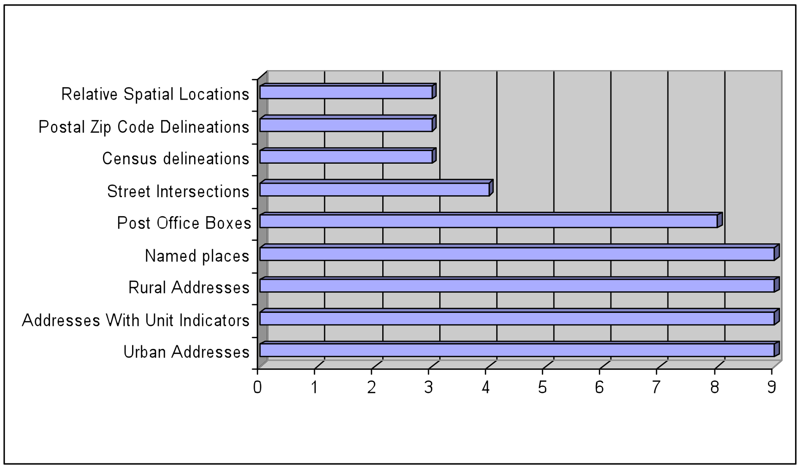

Figure 4 Different types of input data used in geocoding

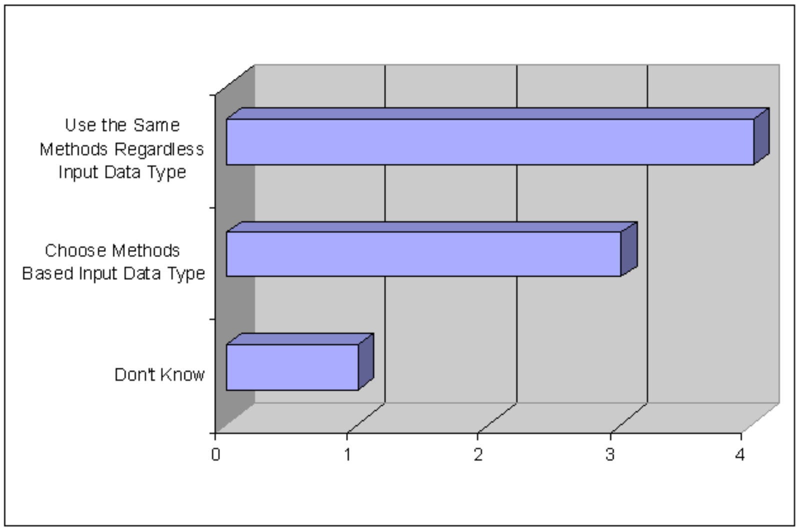

Figure 5 Influence of input data type on choice of geocoding methodology

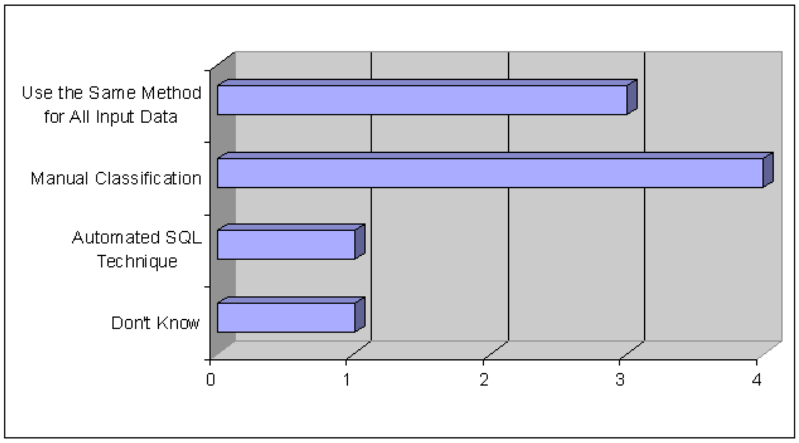

Figure 6 Methods used to classify input data

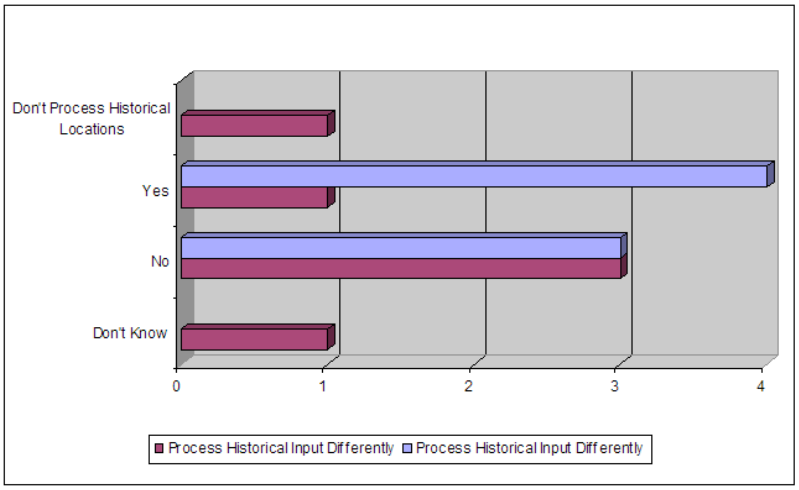

Figure 7 Utilization and processing of historical addresses

In terms of standardization of input data, six out of nine respondents indicated that they always standardize input data, and one out of nine reported doing it "sometimes". All seven of these respondents standardize to the USPS format. Of those that standardize, three out of seven use CASS certified software (US Postal Service 2008), two out of seven non-CASS certified software, two out of seven custom software, and one out of seven use manual techniques (Figure 8). Out of eight respondents which allow them, the standardization techniques include abbreviation replacement (seven), tokenization (three), attribute correction (eight), missing attribute imputation (three), and probabilistic methods (one). One out of seven respondents validates input data using reference data sources.

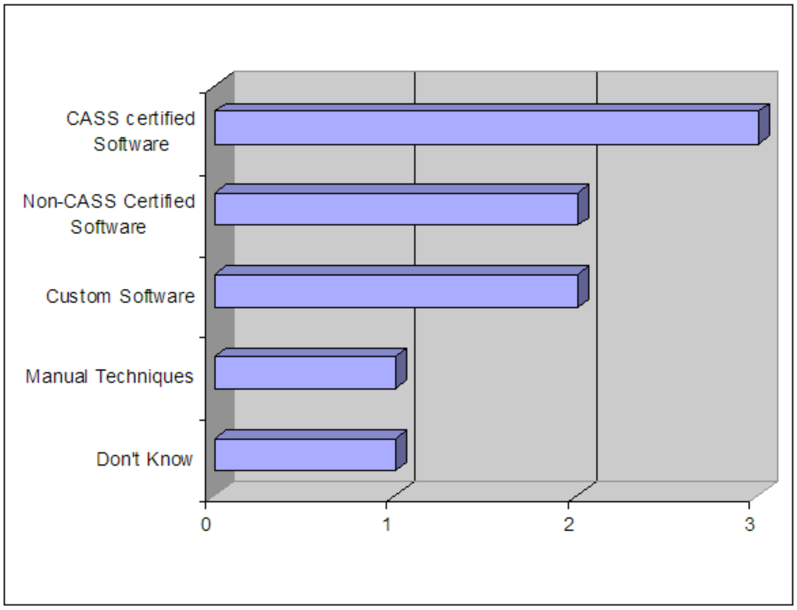

Figure 8 Input data standardization processes

Detailed data investigations using additional data sources are performed by eight out of nine respondents when input data are incomplete or ambiguous, with seven out of nine subsequently able to geocode successfully "most" (i.e. ≥85%) of the time, and one out of seven subsequently able to geocode successfully "some" (~30%) of the time. This process takes 30 minutes for four out of six respondents, while the other two respondents answered that it takes 5 minutes (Figure 9). All nine respondents reported that current input data are the easiest to investigate, and eight of nine respondents find urban addresses easier to geocode than rural addresses. The attributes most commonly altered after an investigation include address name and number, and directional prefix and suffix (Figure 10). The survey included a question about whether or not the respondents revert to lower resolution address when they are unable to track down the required missing information that prevents an address from being correctly geocoded. Four out of six respondents indicated that they do resort to a lower resolution if a match fails, while two out of six do not.

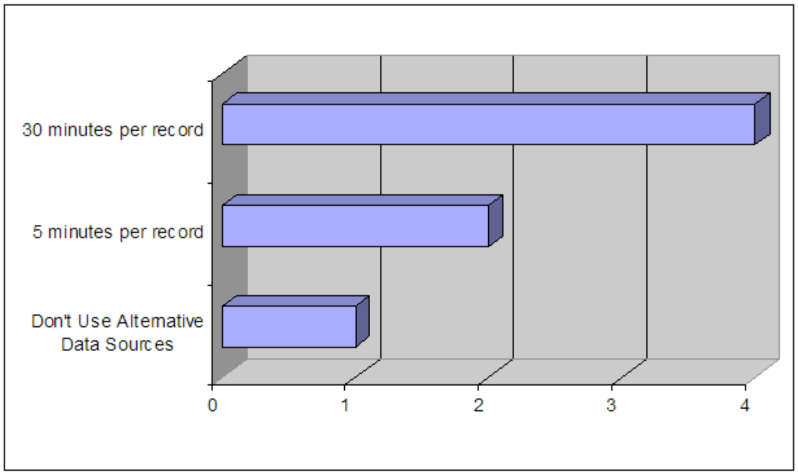

Figure 9 Input data investigation time estimates

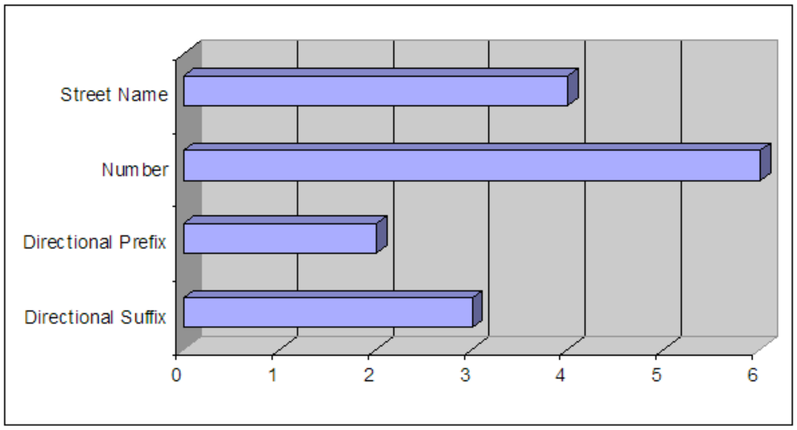

Figure 10 Address input data attributes most commonly updated after investigation

4.3 Geocode Output

All six respondents produce geographic points as their geocoding output, with two linking geographic metadata with the output, with one recording the geodetic datum and coordinate system that was used. When asked if "non-geographic" data is reported as opposed to specifically "geographic metadata", only one out of six respondents gave an affirmative response. However, four out of five responded that "geocode process information" is associated with the output, and three out of five answered that "accuracy information" is associated with the output. No respondents reported returning demographic data with geocodes, but five out of five stated that if they did, they would use a point-in-polygon method to obtain it, while one out of five would also use text-based linkages to other data sources. In addition, four out of seven respondents store their output as spatial data, and four out of seven store the output as text. In terms of specific formats, six out of eight respondents maintain their geocoded data in non-spatial databases, three out of eight in ESRI shapefiles, and one out of eight in ESRI geodatabases. Also, one out of seven respondents transfer data between different formats or storage types.

The last question in the survey related to determining who actually used the output. Eight of the nine respondents reported having academic researchers as consumers. Government researchers, government officials, the general public, and corporate researchers were reported to be users of geocoded data by six, four, two and one of the respondents, respectively (Figure 11).

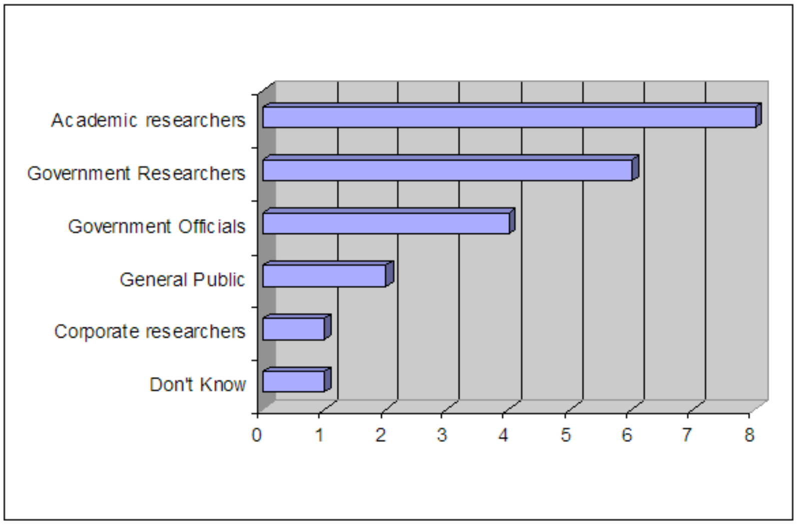

Figure 11 Types of users of geocoding results identified by the respondents to the 2006 survey

4.4 Accuracy

Five of seven respondents reported associating the term "accuracy" with the likelihood that a match is correct, one considers "accuracy" to be the percentage of address attributes which match, and the final respondent considers "accuracy" to be the spatial distance between the output geocode and the position on the ground. It is interesting to note that the percentage of attributes which matched can and should be used as a basis of support within the calculation of the likelihood that a match is correct.

With respect to actual accuracy reporting, eight of nine respondents specified that they report accuracy along with their geocode output, with five of these same respondents specifying the accuracy value for the process as a whole and two specifying it for each component of the process. However, in response to a later question about whether or not accuracy values are associated with each component of the geocoding process, only one of five respondents states they follow this procedure (for the feature matching algorithm). When asked if they report accuracies per geocode, six out of eight respondents stated that they do while the other two respondents stated instead that they report accuracies for the geocoding process as a whole (in contrast to the four out of seven just cited).

Of the eight (of nine) respondents who produce an accuracy estimate, seven report the level of geography matched to (the match type), and one reports the accuracy of the reference data. Four out of seven of those describing accuracy know what metrics the accuracy is reported in. Of these, two out of four specify accuracy as spatial distances, one out of four report probabilities, and one out of four report spatial areas. Five out of eight respondents indicated that the accuracy they associate with a geocode is computed during the geocoding process. In more detail, two out of eight use the accuracy values reported from the components of the geocoding process, and one out of eight use the accuracy values associated with the reference data alone. In terms of accuracy guarantees, four out of seven responded that their geocoding method does not provide any guarantees. But, in response to a separate question, six out of seven respondents report that their geocoding platform is capable of providing feature match certainty as a guarantee, while only one out of seven respondents thought their method provided any type of spatial accuracy guarantee for their geocode output. It is believed that this discrepancy is a reflection of inconsistency in the respondent's knowledge or understanding regarding accuracy in geocoding.

When asked for minimum acceptable levels of reference data sources spatial accuracy, one out of four respondents each responded 5, 10, 1,000 m, and that "it will depend on rural versus urban classifications". Six out of eight of the respondents do not know the accuracy of their reference data, while two out of eight claim to know, but of the seven who responded to a separate question, one simply knew what the vendor tells them, and one claimed to verify their reference data by using geocodes produced from another source. Six out of eight respondents indicated that they are aware the spatial accuracy of reference data is not uniform across large areas.

Nonetheless, four out of six respondents base the accuracy of their geocode results on the accuracy of the reference dataset, although none of these reported knowing what the accuracy of their reference data were. In addition, these four respondents include the single respondent who the spatial accuracy of their geocodes was guaranteed, yet this one participant does not know the accuracy of their reference data, nor have they ever checked the spatial accuracy of the reference dataset or the resulting geocode. Two out of four respondents have checked the spatial accuracy of their non-guaranteed data, with one respondent using GPS measurements and the second of these respondents comparing with different geocodes, while five out of six respondents stated that they are concerned about the temporal accuracy (age) of their reference data, six out of six reported acceptable ages for their reference data. These were 5, 10 and 20 years reported by three out of six, one out of six, and two out of six respondents, respectively.

Regarding what most affects the accuracy of the geocoding process, eight out of nine respondents reported that reference data is the component which most significantly affects the accuracy, and seven out of nine, two out of nine, and one out of nine believe that input data, methodology, and standardization also affect the output accuracy, respectively. As for what has the largest effect on the accuracy of the results, five out of eight respondents think that reference dataset completeness has the greatest effect. In addition, two out of eight respondents specified reference source spatial accuracy as having the greatest effect on its own, while one out of eight think that attribute accuracy in conjunction with spatial accuracy most affects their outcomes. At the same time, six out of eight respondents indicated that input data completeness alone has the largest effect on the accuracy of the results. One other respondent indicated that input data ambiguity has the most effect, while one final respondent thought that input data completeness and ambiguity together most affect the geocode output.

The minimal spatial accuracy that respondents reported as being acceptable ranged drastically, with 5, 100, 200, and 1,000 m give as a response by one, one, two and three of the respondents, respectively (Figure 12), while the final respondent indicated that "a minimum acceptable value would be dependant on the application for which the geocode was to be used". Two out of six respondents reported that minimal acceptable accuracy is useful for large scale (national) spatial analysis, while four out of six think it suitable for aggregate analysis, and one out of six thinking it suitable for rural areas because census tracts are large. Interestingly, the respondents that gave the 5 m and 1,000 m responses to the question about minimally acceptable accuracy both thought the data would be suitable for aggregate analysis, clearly indicating the widely varying opinions on the data quality requirement.

In turn, the maximum required spatial accuracies were reported as 0.5, 2, 5, 50, and 1,000 m by single respondents (Figure 13). Again one out of eight stated it would depend on the application. Four out of six think maximum level accuracy geocodes are suitable for individual scale spatial analysis, whereas one respondent each believes they are suitable for micro-scale (few meters) and urban spatial analysis. When asked specifically "What level of accuracy do you need to be able to geocode to", three respondents reported 0.5, 2, 1,000 m, respectively (Figure 14). This time two out of five respondents stated the need to be able to generate any user-defined level of accuracy.

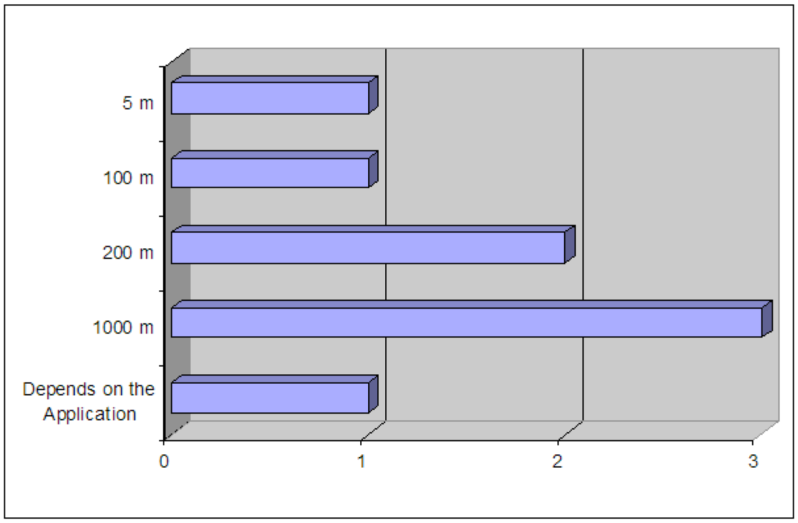

Figure 12 Minimum acceptable geocode spatial accuracy (distance)

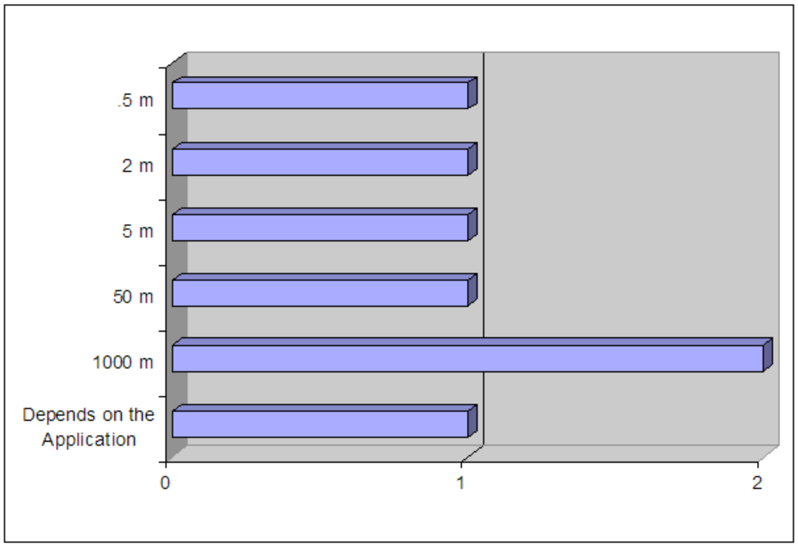

Figure 13 Maximum acceptable geocode spatial accuracy (distance)

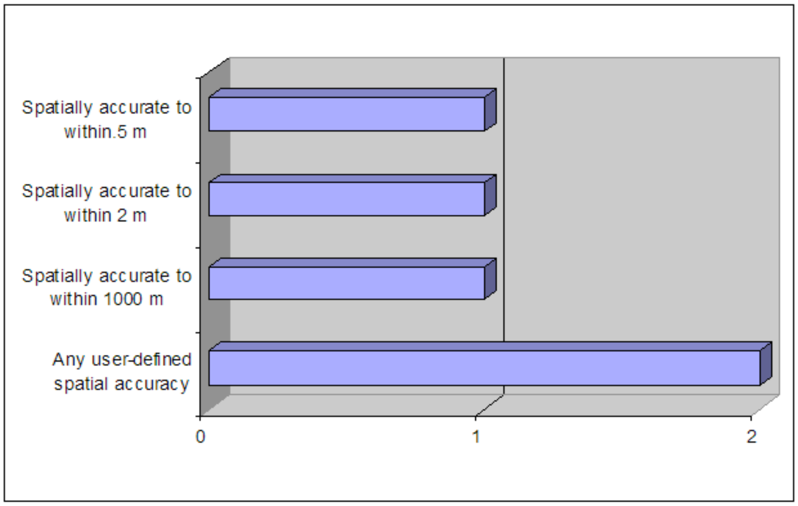

Figure 14 Acceptable geocode spatial accuracy (distance)

Turning our attention next to the actual accuracy levels that are currently achievable, two out of six respondents reported being able to obtain census tract level accuracy, while one out of six each thought the capability of reaching census block and census block group accuracy levels was necessary (Figure 15). Two of six respondents reported being able to achieve these rates 70% of the time, while three out of six are able to achieve them 85% of the time. Three of four respondents do not know the level of spatial accuracy they are able to achieve, while one out of four reported they are able to achieve 250 m spatial accuracy 85% of the time. It is important to note that this respondent is the same one that indicates earlier that they rely 100% on the accuracy of their reference data to dictate the spatial accuracy of the output, although neither of these outcomes have ever been verified by this particular respondent.

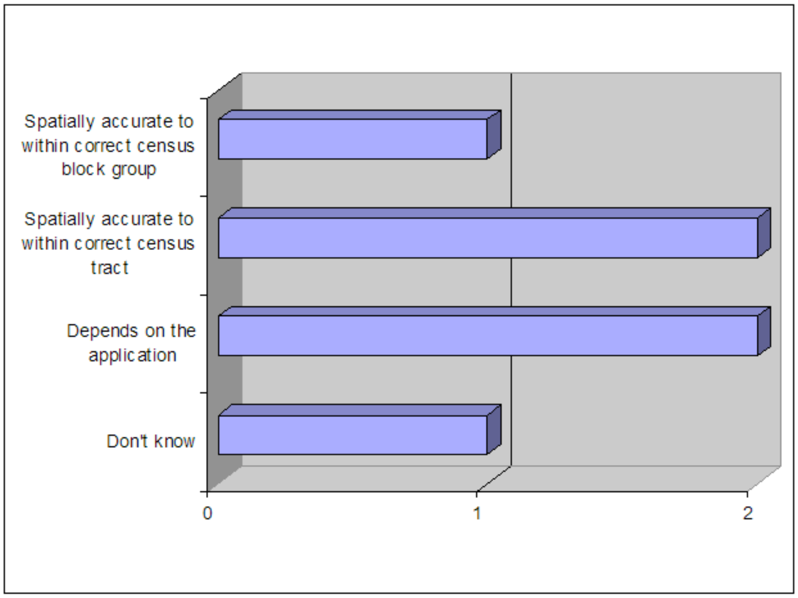

Figure 15 Geocode spatial accuracy achieved (distance)

This is a complicated picture given that four of six respondents reported that they are required to produce geocodes or varying levels of accuracy. Three respondents base this need on the availability of different data sources, two on the notion that different classifications of areas need more or less accuracy (rural needs less, urban needs more), and the final respondent thought the needs of their data consumers would drive their expectations.

4.5 Organizational Geocoding Capacity

In response to questions related to how respondents organize their staff that performs geocoding tasks, all of the respondents reported that the staff who perform their geocoding tasks were trained specifically to do geocoding tasks. Two staff members perform their tasks in four of the six organizations. The number of persons assigned to geocoding appears to be proportional to the size of the organization and its annual caseload. The annual number of cases geocoded also grows proportionally with the size of the organization, ranging from 500 cases to more than 100,000 cases for the six respondents.

Labor costs per year are likewise proportional and range from less than one person-month to greater than 2 person years. In terms of where the money is spent, four out of six respondents report that employee salaries make up most of the cost, while two out of six report the cost of commercial firms as the largest outlay. As a final point, one respondent reported that in addition to employee salaries, obtaining reference data is the most costly element of geocoding.

4.6 Software

In response to questions in this section, eight out of nine respondents reported that they are actively geocoding (in comparison to the 83% reported on the 2005 NAACCR GIS Survey). According to the results of the questions on the remainder of the survey, it is clear that the one out of nine who reported they are not presently geocoding in fact are geocoding in that they produce geographic points for address data, they simply just do not use a commercial package to do so (even though they report they plan to do so within the next year). Six out of nine respondents have been geocoding for more than 7 years, from which we can characterize the majority of the survey respondents as having a great deal of technical knowledge.

Six of eight respondents have not evaluated more than one geocoding option. Of those that knew how their geocoding process was chosen, two out of seven chose the one they use based on availability, and only one respondent chose their approach based on evaluation of specific criteria (flexibility and customization). Three out of six respondents change their geocoding software from time to time, with one doing so because of managerial decisions, one because of technical improvements (higher achievable match rates), and one because different geocoders (probabilistic vs. deterministic) work better in different circumstances. Three out of six are able to achieve higher match rates in certain situations.

Six out of seven respondents perform their geocoding in-house. Five out of seven use a commercial product, one out of seven uses a home-grown solution, and one out of seven uses a combination of the two. Five out of eight commercial product users rely on ESRI's ArcGIS products (ESRI 2008), two use MapInfo products (Pitney Bowes Software Inc. 2008), and two use Centrus products (Group 1 Software Inc. 2008). One of the four in-house geocoding respondents make their software available to other state agencies.

Just one of seven respondents uses a commercial firm (Bamberg-Hadley), and no information is known about cost in terms of bulk or per-record rate or the geocoding process used by the firm. Three out of eight respondents use free online geocoding services including geocoder.us (three out of three), maporama.com (two out of three), geocodeme.com (two out of three), mapquest.com (one out of three), and google.com (one out of three).

4.7 General Methodology Questions

Overall, the survey respondents were very knowledgeable about the theoretical aspects of the various geocoding methodological issues, but there were a number of discrepancies between what the respondents thought they understood about their own particular geocoding process and the detailed answers they were able to provide. For instance, while six out of six respondents reported that they knew what their geocoding process did, at least one respondent answered “I do not know” to 11 of 21 methodology questions. In response to specific questions about respondent’s perceived geocoding knowledge, this particular respondent self-identified themselves as possessing “medium high geocoding knowledge”, and also self-identified their knowledge on the topic as “I know quite a bit”. This inability to answer methodological questions is in fact not due to the respondents’ lack of knowledge about geocoding methodology, but instead because their geocoding was performed by vendors who do not provide these answers.

When asked about the components within their geocoders, two out of five respondents do not know the components of their geocoding process. One out of six reported that their geocoding methods were not documented. Of the four out of six who do document their processes, three out of five answered that they document the “assumptions” without specifying exactly what they are, one out of five document the data sources used, and the remaining one out of five did not know what was documented. Twelve out of fourteen respondents reported that they felt they could control (choose) portions of their geocoding process. The portions of the geocoding process which respondents could and could not control are displayed in Figure 16.

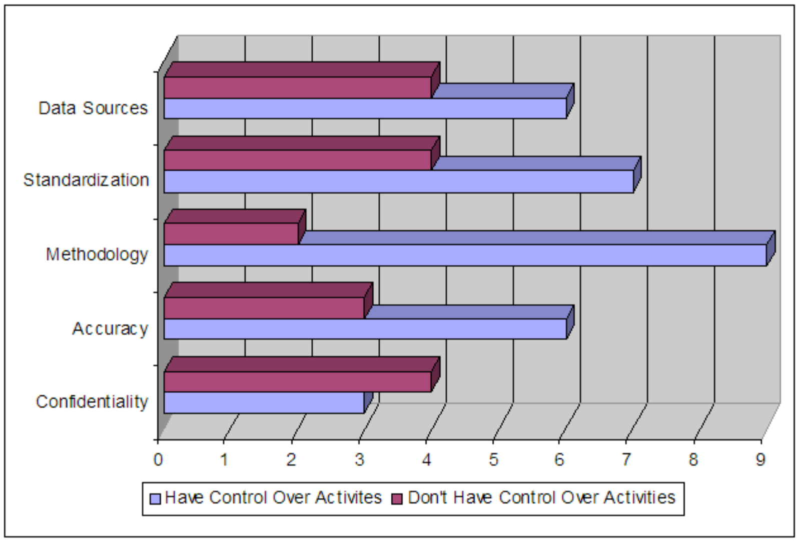

Figure 16 Control over various geocoding activities

In general, the survey respondents reported using a wide range of geocoding process types, with five out of six using manual geocoding, four out of six using linear-based interpolation, three out of six using areal unit-based interpolation, and five out of six using a simple feature matching approach (lookup lists of points only with no interpolation). Three out of five respondents are sure they use a feature matching only geocoding method, and (in response to a separate question) two out of four are sure they do not use any feature interpolation. However, later in the survey five out of five respondents reported they do not know the difference between feature matching only and feature interpolation-based, casting doubt on these answers. An analysis of the survey results confirms this; of the two out of four who are sure they do not use feature interpolation, one out of two stated conversely that they in fact do, in both the linear-based and areal unit-based detail sections (see Sections 4.8 and 4.9, respectively), while the other one out of two stated using it in only the linear-based interpolation section.

Likewise, confusion exists as to the use of multiple reference data sources. This is apparent, because four out of five respondents reported not using a second data source after a geocoding failure on a first component, though all these respondents (five out of five) reported being able to geocode different components of an input address (street address, city, state, and ZIP) which requires at least four different data sources (one for streets, one for cities, etc.). Again, this same confusion is evident when three out of five respondents reported not using a geocoding hierarchy, although they all (five out of five) also indicate that they obey the address component hierarchy just described (as discussed in Section 4.4, where seven out of nine of those reporting accuracy specified the level of geography matched). The one respondent who reported employing a hierarchy uses address points first, then parcel centroids, and finally linear interpolation as a last resort.

Regarding working in batch-mode, three out of five respondents reported that their geocoding process could work in batch-mode, one out of five reported per-record ability, and one out of five reported having both capabilities. As far as being interactive, three out of five respondents use an interactive geocoding process, with only one out of four respondents stating that a user is prompted for more information when it is required. As for sub-parcel geocoding, three out of five respondents are able to perform sub-parcel level geocoding. However, of these respondents, three out of three assign the same geocode to all sub-units within a single address (within a single parcel).

4.8 Linear-Based Interpolation Methodology

Five respondents indicated they use linear interpolation as a part of their geocoding process, and stated that they understand how the process works in great detail. Although three out of four answered that they know the sources of error, only two out of four responded that they know how to quantify them. All of the respondents “know the assumptions” used in linear-based interpolation geocoding, yet four out of five indicated that they have not heard of the assumptions listed in Goldberg (2008b). Three out of five respondents use linear interpolation on 85% of their data, with the remaining one out of five using it on 100% and one out of five using it on 30%. Four out of four respondents use linear interpolation on current data, with three out of three using it on urban data, and only one out of three using it on rural data.

Concerning the use of dropbacks in linear interpolation, three out of four respondents use a dropback, and of these, three out of three uses a constant distance value and two out of three constant direction value, with one out of two using zero degrees. However, different distance values (.5, 5, and 10 m) were reported by each of these respondents (one out of three). One of four respondents attempt to overcome the parcel existence assumption by utilizing the actual number of parcels along a street segment, and/or using parcel dimensions in their linear interpolation calculation, although none of the other respondents reported not using a corner offset.

4.9 Areal Unit-Based Interpolation Methodology

Four of five respondents reported using areal unit-based interpolation as part of their geocoding process. It should be noted that this is in contrast to the five out of six who indicated that they use areal unit-based interpolation geocoding earlier in the survey (see Section 4.7). Of the four respondents, three apply this methodology to current data, and two apply it to urban data. However, the amount of data it is used on varies from 15% for two of the three of the respondents, to 70% for te third respondent. In terms of time and dollar costs, two out of three respondents performing areal unit-based interpolation do not know what it costs, while the other one out of three report a monetary cost of $0.01 per record, with it taking on average 0.01 seconds to process each record.

Concerning the use of centroid derivation in this methodology, one of two respondents using areal unit-based interpolation reported using the centroid of the area unit as the output, with the other respondent indicating that they do not use the centroid, but not listing what they used instead. Neither of these respondents reported using any form of centroid weighting methodology to move the output point away from the geographic centroid. Lastly, one out of two reported that the only decision made in centroid placement was to disallow centroids falling outside of the areal unit.

4.10 Manual Geocoding Methodology

Five of six respondents surveyed indicated they are familiar with manual geocoding. In this detailed section of the survey, three out of four reported using manual geocoding as a part of their process, with the remaining respondent unsure if it was used or not. It should be noted that these numbers do correlate with the five out of six respondents who reported using manual geocoding earlier in the survey (see Section 4.7). One out of four respondents reported that 15% of cases require manual geocoding. One out of three and two out of three respondents reported that manual geocoding is performed on historical and current data, respectively. One out of three and two out of three respondents reported that manual geocoding is performed on rural and urban data, respectively.

None of the respondents (two out of two) were able to quantify how much the manual geocoding process costs in terms of a dollar figure, but of those performing it, one out of three reported that it takes an average of 30 minutes per record, while two out of three indicated that the process takes only 5 minutes per record (Figure 9). The sources of data used for manual geocoding varied from freely available public domain information such as phonebooks (both online and paper) and staff manually searching for websites, to official government data such as Division of Motor Vehicles (DMV) records, census tract data and rural town names.

4.11 Feature Matching Methodology

All six respondents surveyed answered that they are familiar with the concept of feature matching. With regard to the feature matching processes employed, two out of six respondents reported using strictly probabilistic matching methods, three out of six using strictly deterministic methods, and one out of six reported using both. All of those using probabilistic methods (two out of two respondents) know what their uncertainty cutoffs are (70% and 80%, respectively, in these instances). Two out of three respondents that reported while performing feature matching they do not attempt to break ties between ambiguous feature matching results, preferring instead to leave the record non-geocoded.

While four out of five respondents know what the attribute relaxation process is, only one out of four reported using it when specifically asked. However, when asked specific questions, one out of two responded that all address components (attributes) are considered candidates for relaxation, while one out of two reported relaxing only street type. With regard to the usage of phonetic algorithms, five out of five respondents knew of the SOUNDEX algorithm, two out of four were sure their process uses it, while the remaining two out of four indicated that they did not know if their geocoding processes utilize it.

All six respondents reported knowing what a match rate is, and none consider it the same as accuracy. The respondents were split between whether a higher match rate (one out of six) or higher accuracy (three out of six) is more desirable in the geocoding process, or if they are both equally important (two out of six). Of those who responded (n=5), four knew what their match rate is two reported match rates of 70%, and two reported match rates of 80%. Three out of five respondents each reported that 15% and 30% of urban and rural addresses are non-matchable, respectively. From the reported results, current (non-historical) data represent 30%, 70%, and 85% of the data that failed to match, each at one out of five respondents. Similarly, historical location data represent 15% and 30% of the data that fail to match, again each at one out of five respondents. In terms of processing current and historical data, one out of four respondents reported processing current and historical records separately.

4.12 Confidentiality

All six respondents surveyed are aware of and concerned with confidentiality, security, and privacy issues inherent in creating, storing, and disseminating geocoded health data. These concerns cover aspects of both data security, such as who can access the data, as well as information security, for instance what information can be gleaned from the data once it has been obtained. In many cases, these two distinct areas of concerns are comingled, and most respondents consider them as a single topic, evidenced by the variety of responses within each of the two categories.

Survey questions about attitudes and practices relating to data security revealed that in order to protect access to geocoded data, five out of six of respondents store their data in what they would consider "physically secure" locations, while three out of six respondents also employ some form of authentication. One out of six respondents noted that they require contractual user agreements before data can be released, while one out of six answered that simple bureaucracy was an acceptable method for ensuring limited access to these data.

All six respondents provide access to some sort of masked data (aggregate versions), with three out of six also providing access to individual level data. Three out of six respondents release geocodes directly, and two out of six allow access to the raw address data. One out of two who indicated that they release raw data also responded that their consumers do not geocode the raw data in-house themselves, so the utility of this practice is not clear. Respondents also reported employing a variety of masking techniques to ensure privacy and confidentiality once the data have been released. For instance six out of six respondents release data in an aggregated form, five out of six release lower geographic resolution versions of geocodes, and the last respondent reported using randomization.

All respondents (six out of six) allow researchers access to data, three out of six allow access to the general public, and three out of six allow access to public officials. Note that one out of six indicated they provide their data to anyone, but only at an aggregate level. Also, only one out of six respondents noted an Institutional Review Board (IRB) requirement before any data could be released. All respondents (six out of six) make the data available in a digital format, with four out of six allowing Internet downloads, one out of six allowing email transmission, and three out of six allowing data shipments via regular mail.

5 2008 USC Geocoding Needs Assessment Survey

Similar to the *Geocoding Best Practices Survey* (Goldberg 2008a), the *Geocoding Capacity Survey* (Goldberg et al 2008b) was organized as a series of categories of questions. This survey was administered through a website. It was disseminated to 44 individuals each from a different organization. Seventeen individuals responded, giving a response rate of 38%. The results are presented in the following sub-sections.

5.1 General Questions

Almost all of the survey respondents reported that their organizations presently perform some type of geocoding (15 of 16 respondents). The majority of the respondents (eight) have been geocoding their registry for five or more years (Figure 17). Geocoding is performed by a single person at roughly half of the registries (seven out of 13), while one out of 13 reported that geocoding is performed by five or more people. The same respondents who reported a single person performing the geocoding also reported that this single person comprised 15% of their registry staff. Two out of 16 respondents are not presently geocoding. One out of 16 stated they do not currently do so because they lack the time, money, and knowledge of how to, and that they do not know when they plan to start geocoding.

Seven of 15 respondents reported that their organization evaluated multiple geocoding solutions, and seven out of 15 also reported that their organization has not. Of those who had, the reason most often cited for the decision was that the organization chose the geocoding solution that gave the best match rate (three out of seven registries), followed by availability (two out of seven registries), and highest accuracy (two out of seven registries) (Figure 18). One respondent noted that they are in the process of evaluating other options, but have not yet found one that would produce comparable results at a similar price without creating more work for individuals within their organization.

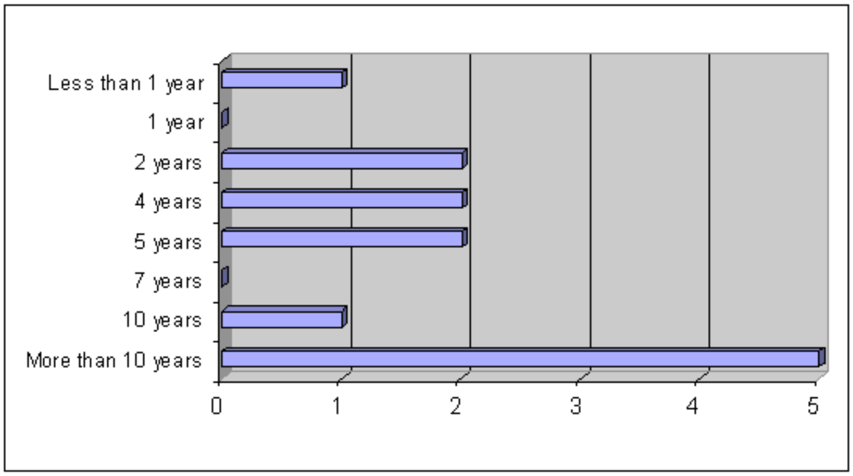

Figure 17 Length of time respondents have been geocoding

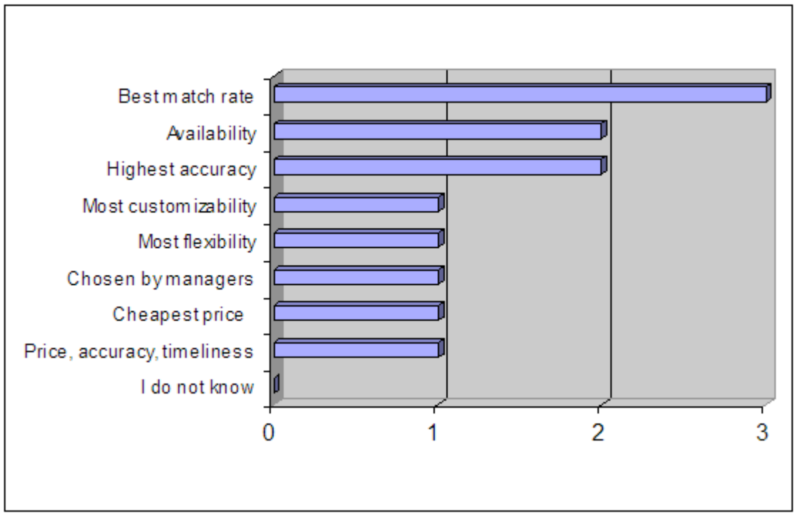

Figure 18 Reasons for choosing current geocoding process

5.2 Caseload Questions

The number of cases geocoded by each of the respondent registries varies greatly from 500,000 cases per year (five registries) to 1,000 cases per year (three registries) (Figure 19).

The scatter plots in Figure 20 and Figure 21 show the total cost in terms of dollars spent on geocoding and cost per-geocode versus the numbers of cases geocoded at each of the seven registries where the respondents knew the cost of their geocoding process. The scatter plot in Figure 22 shows the cost in terms of time (person months) spent on geocoding versus the number of cases geocoded at the four registries who reported these costs. From these figures, it is clear that there is no simple relationship between the number of cases geocoded and either the total cost or the cost per-geocode. Respondents stated that the most costly parts of the geocoding process are commercial firm charges (four out of 11), employee salaries (three out of 11), and the development of custom geocoding applications (one out of 11). The remaining three respondents do not know what the most costly aspects of their geocoding process are.

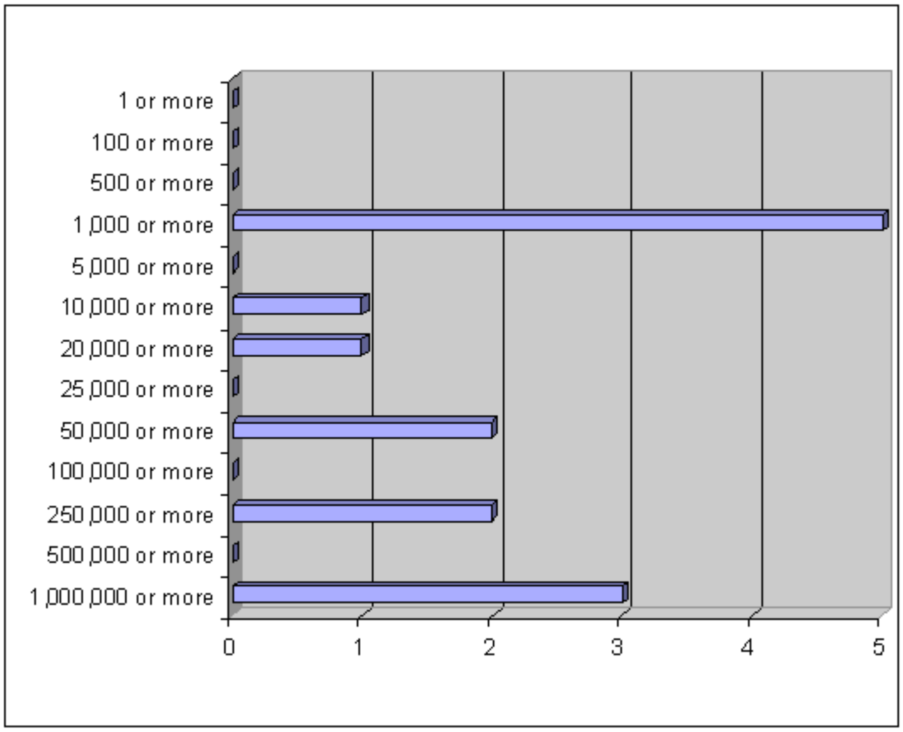

Figure 19 Number of cases geocoded per year

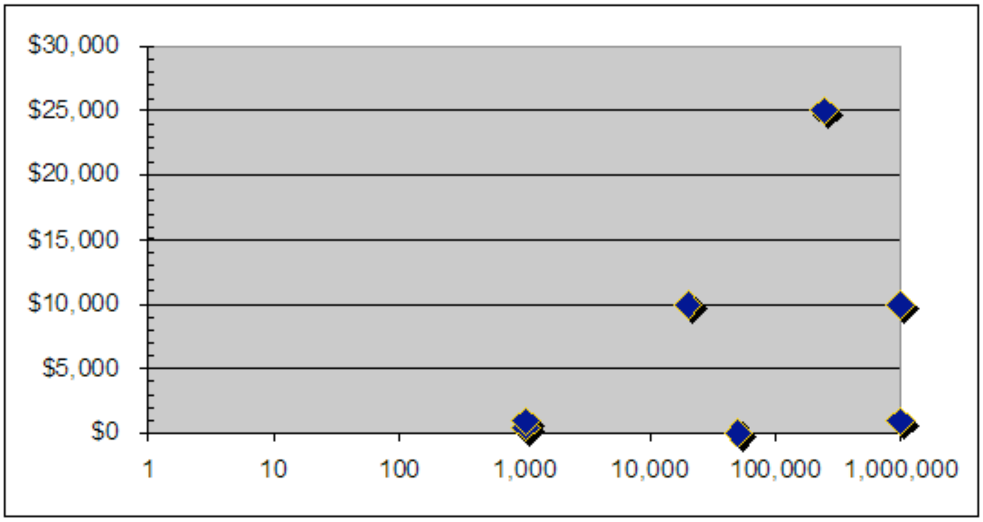

Figure 20 Total geocoding dollar cost vs. number of cases geocoded per year

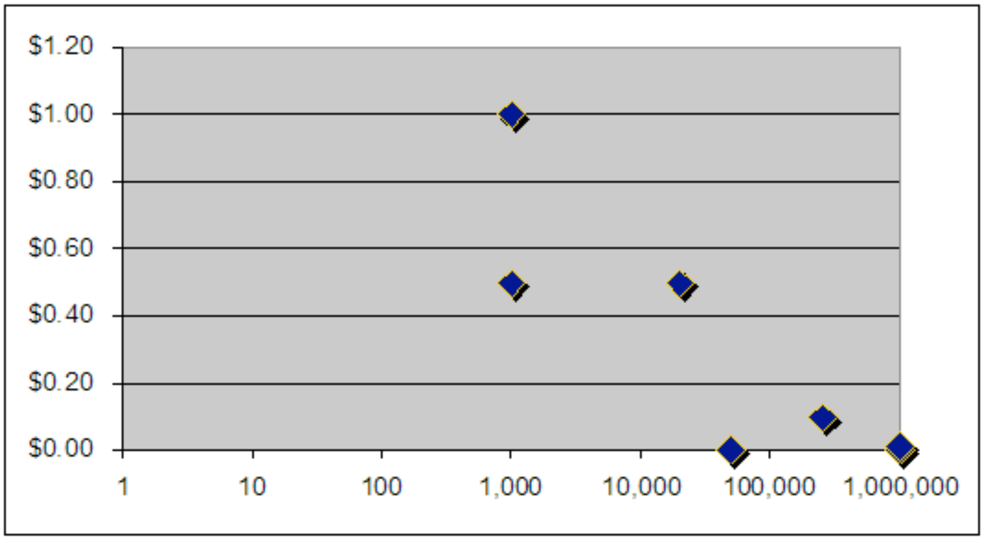

Figure 21 Dollar cost per geocode vs. number of cases geocoded per year

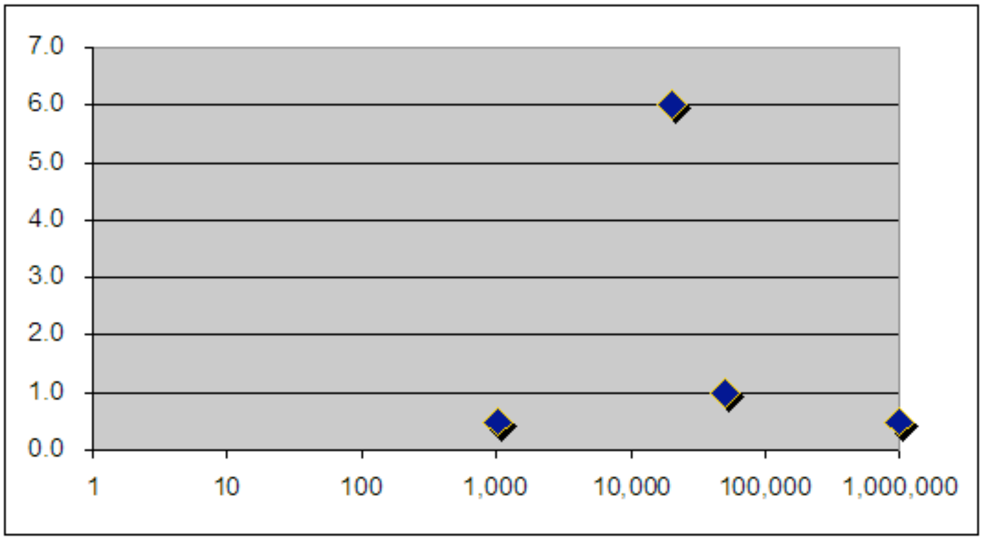

Figure 22 Total time cost per geocode (months) vs. number of cases geocoded per year

5.3 Training Questions

Nine of 15 respondents reported that members of their organization have been trained for geocoding. However, one respondent qualified their answer that even though "training" has been performed, it was "not formal". Half of the respondents (four out of eight) stated that only one person has been trained, while three respondents reported two people and one reported that ten or more people have been trained. Of those that could quantify it, six out of seven registries placed this level of training at 15% of their organizational staff, while one out of seven reported it was greater than 1%, but less than 15%.

5.4 Online Geocoding Questions

Survey respondents were almost evenly split between whether or not they use free online geocoding websites (Figure 23). The online geocoders that are utilized by the seven respondents who answer this question "yes" are listed in Figure 24.

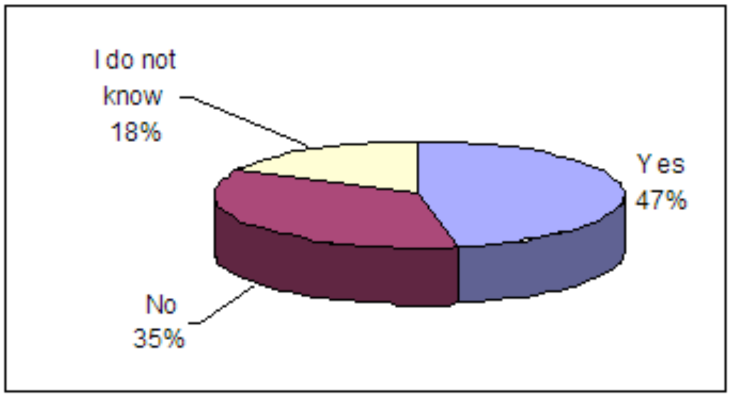

Figure 23 Usage of free online geocoding websites (n=16)

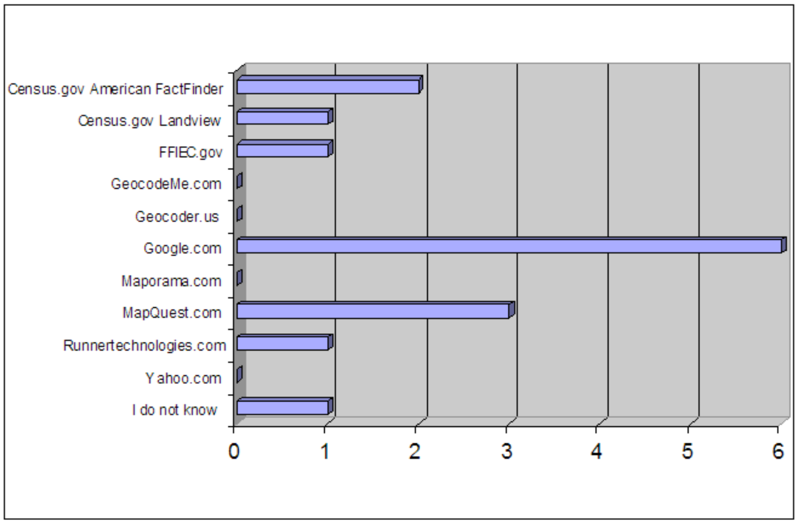

Figure 24 Free online geocoding websites utilized

5.5 In-House Geocoding Questions

Three-quarters of the registries surveyed (n= 16) responded that they perform in-house geocoding. Only three registries are sure that they do not perform geocoding in-house, while the remaining respondents did not know. Of those performing in-house geocoding, nine reported that they use commercial software to do so, reported they do not, and one respondent did not know. The commercial geocoding software utilized by the nine respondents who answered "yes" is listed in Figure 25.

Only four out of 11 registries responded that they have written their own custom geocoding software. Five out of 11 registries were positive they have not, while the other two did not know. One of the four respondents who said that they had written their own software qualified their response by adding that they perform "hybrid geocoding combining multiple commercial products and reference data", and one stated that they make their software available to other organizations.

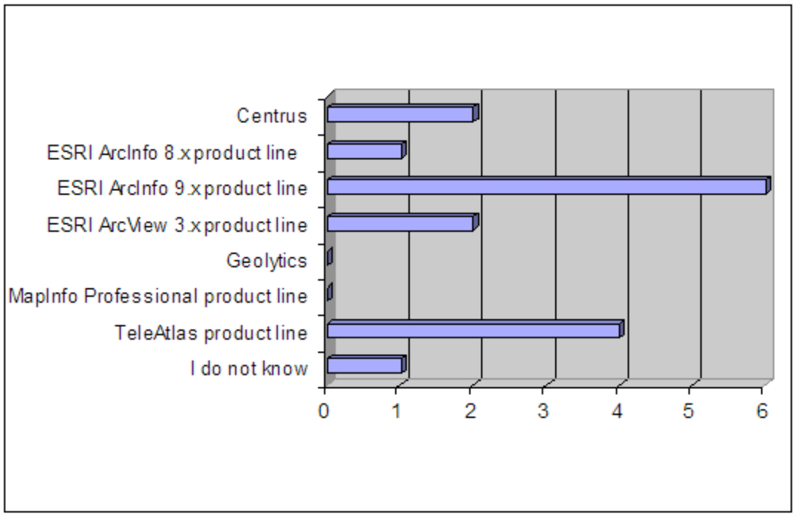

Figure 25 Commercial geocoding software utilized

5.6 Commercial Geocoding Questions

The majority of registries (nine out of 16) responded that they do not use a commercial firm to perform their geocoding, but six responded that they do, and one did not know. The six out of 16 who responded that they included one registry who reported that their geocoding process "had been performed by their database vendor in the past, but were now in the process of setting up a geocode arrangement with a commercial geocoding vendor" in responses to an earlier question asking "Does your registry perform any type of geocoding?". Three reported using the commercial firms Geocode.com (Tele Atlas), Claritas, and "Local Vendors".

Of the four respondents who knew about costs, two reported that the commercial firm they use charges per-record, and two reported that they do not. One respondent noted that "there is one charge for batch submittal [and] another for console-level match". Of the two registries that reported a value, one indicated that their commercial firm charges $20.00 per 1,000 records ($0.0two out of record), while the other indicated being charged $0.50/record.

Two of four registries reported that their commercial vendor charges them lump-sum, and two out of four indicated they do not. Of the two that are charged lump-sum, one indicated that they are charged $1000 or more and the other one out of two indicated being charged $500 or more. Note that these costs for the registries who reported paying their geocoding firm lump-sum are equivalent to the costs reported in Figure 20. Thus, these two respondents validated their responses again, and from these responses the per-geocode can be calculated to be $1.00 and $0.50 for each of these registries, respectively. Also note that of the four registries who answered both questions about per-record and lump-sum charges, the two registries who responded affirmatively to either question responded negatively to the other and vice versa (i.e. if a registry is charged lump-sum they are not charged per-record).

5.7 Methodology Questions

The majority of respondents (11 of 16) indicated that their organization has a geocoding process (methodology). Of the five who responded that they do not, two qualified their response by stating that they use a commercial firm. Those who have a geocoding process reported "knowing how their geocoding process works", although two out of 10 qualified their response by stating they know "the general idea, but not all the techniques the primary individual uses". Three out of 10 stated that they would characterize their knowledge about their geocoding process as "High – I know everything", while three out of 10 characterized it as "Medium High – I know quite a bit", and four out of 10 "Medium – I am familiar". Five out of eight reported that they know the components of their geocoding process, and seven out of nine indicated that they know the stages of their geocoding process that could introduce error/uncertainty.

Nine out of nine respondents reported using address list-lookup based geocoding (i.e. no interpolation), with the majority (seven out of nine) performing manual geocoding, and less than half (four out of 9) performing some type of interpolation-based geocoding (Figure 26).

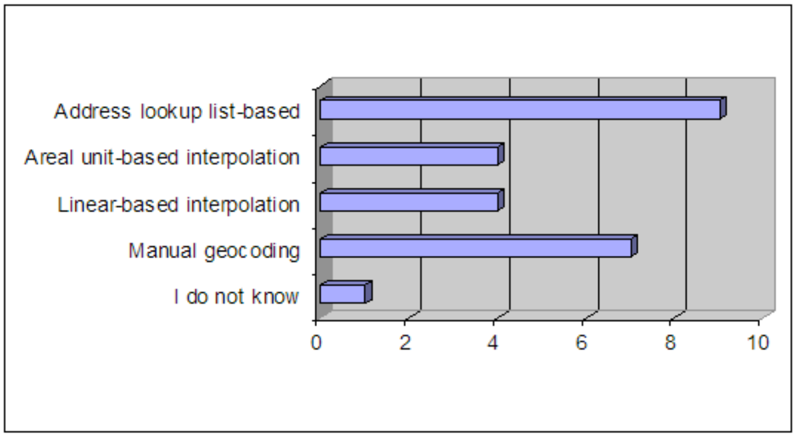

Figure 26 Geocoding methodology utilized

Eight of nine registries reported that their geocoding methodology attempts to match (i.e. find the correct reference feature in a reference dataset) with multiple reference data sources (in the case when a match is not found in the initial one). These same eight registries also reported that a set hierarchy is used in these cases where multiple reference datasets are employed. Five of these eight registries know the hierarchy (order) used for their reference sources although they all use a different order (Table 3).

All nine registries responded that their geocoding process is batch-oriented. Only one respondent reported that their process is not one-at-a-time oriented, while five respondents stated that it specifically is. However, one of these five qualified their response stating "Records failing batch match are processed one at a time". Also, two out of nine did not report yes or no, but instead that "When records do not match in the batch process they may be looked up manually", which can be considered one-at-a-time processing, thus seven of nine registries do in fact process records one-at-a-time.

Table 3 Reference dataset hierarchies used at registries

Registry	Reference Dataset Order
1	(Waiting on parcel data) Address point data Street vectors ZIP code polygons City polygons County polygons
2	Parcel data Street vectors ZIP code polygons City polygons
3	Address point data Parcel data Street vectors
4	TIGER/Lines 2006 streets Parcel data centroids County Level Streets TeleAtlas Streets street segment centroids ZIP Plus4 centroids ZIP code centroids City/Place centroids Post Office centroids
5	Address point data ZIP code polygons

While only three out of nine registries responded that their geocoding process is interactive, of the six out of nine who responded that it is not, one out of six stated that their process is in fact "interactive only for reject processing". Thus, four out of 9 registries do in fact have an interactive geocoding process. Only one respondent reported that their interactive process will prompt the user for more information if it is required.

None of the nine respondents were sure that they perform sub-parcel geocoding, but four reported that they do not know. Therefore, none of the responding registries were able to say whether or not they assign a unique geocode to distinct units at the same address, or how those distinct units were derived.

5.8 Documentation Questions

Ten of 16 respondents reported that their geocoding methods are documented, and three reported that they are not. However, one who responded negatively qualified their answer to state “API is documented. 'Internal' process not yet documented”. Also, two other respondents who declined to answer yes or no stated “Somewhat”, and “A new mandate will necessitate this but is only first being implemented now”. Therefore, the rate of registries who document their geocoding process should be 13 out of 16.

Eight of nine registries reported documenting the reference datasets used in the geocoding process. One registry reported documenting the following “Geocoding and Address Matching Primer”, “Address Standardization”, “Proper Use of Geocoded Data”, “Software and Data Used”, “Accuracy Issues”, “Output File Structure”, and “Return Code Sample Matrix”.

6 User Types and Trends

The analysis of results from the three surveys documented in this report provide some insight into three distinct classes of geocode users; high-, medium-, and low-achievers. Most of the survey respondents fit into the medium-achiever category, with just a few in both the high- and low-achiever groups.

The high-achiever group consists of registries who are completely knowledgeable about every aspect of the geocoding process. It is clear that these are the individuals who perform the geocoding and thus have vast practical knowledge, have developed a geocoder themselves, or have spent considerable amounts of time researching their options. These individuals typically are mid- to lower-level employees, and performing geocoding is usually part of their job. Members of this group (and/or their registries) typically geocode a very high number of cases each year, have been using geocoding processes for more than 4 years, and are from both rural and urban regions. From the survey responses, this group appears to be most concerned with geocode accuracy and the methodological issues inherent in different geocoding strategies.

The medium-achiever group consists of users who possess detailed knowledge of many aspects of the geocoding process, but this knowledge is, in several cases, not directly related to specifically being certain (or in some cases even correct) about the inner workings of their geocoding process, which translates to limited practical knowledge about their geocoding process. Members of this group come from registries that geocode both high and low numbers of records each year, have typically been using geocoding processes for one to four years, and are from both urban and rural areas. These individuals are typically mid- to high-level employees, and performing geocoding is typically not part of their job. The primary concern of this group appears to be the costs associated with the geocoding process, both in terms of time and money. While also concerned about accuracy, this group seems to be less knowledgeable and/or concerned about the methodological issues related to the geocoding process.

The low-achiever group consists of users who are, for the most part, not very knowledgeable about the geocoding process at their registry, and in some cases, the geocoding process at all. The registries to which these individuals belong are typically not yet geocoding or have been using geocoding processes for less than four years. Individuals in this group range from low- to high-level employees and their registries are also spread across rural and urban regions. This group of individuals seems to be most chiefly concerned with operationalizing a geocoding process. What appears to matter to them the most is the availability and cost of geocoding software and reference data.

7 Priorities of Registries

In terms of control over geocoding procedures, the responses submitted early on in the 2006 USC survey were a reflection of the respondents' original understanding of the level of control they had over geocoding activities. Once the technical details of the geocoding process were further outlined and the questions refined, a shift or change occurred where the registries responded that in actuality they have less control over geocoding activities than their responses to the first set of control questions indicated. This discrepancy indicates that many of the registries may have gained a clearer understanding of the meaning of "control" over the geocoding process as they got deeper into the survey. These results emphasize the disconnect between the beliefs people have about the geocoding process and the realities they actually face on a daily basis when performing their geocoding tasks.

Concerning the level of understanding of reference data sets used in geocoding, both the 2006 and 2008 USC survey results are indicative of the level of training and expertise of the survey respondents. Many are not clear what their existing geocoders actually do with reference datasets, or for instance how the geocoders actually choose one dataset over another. Also, though many of the registries indicate that their geocoders support all types of reference data (imagery, vector, etc.), several don't know actually which ones, and conversely respond that they don't use the reference data types they initially indicated they utilize. Such inconsistencies in the responses indicate that these survey questions need either to be more general, or much more detailed, in order for the respondents to fully understand them.

From these widely varying language related to geocoding accuracy, such as "most required", "worst acceptable", "needed", and "currently achievable", it becomes clear that more discussion and research is required into spatial accuracy related to geocoding. All of the registries report having extraordinarily high acceptability and needs requirements; however, the accuracy of any of these components is presently unknown and/or the levels or accuracy they "require" are currently unachievable by any of the registries.

From the results of the 2005, 2006, and 2008 surveys, Table 4 identifies the minimal recommended user requirements for a geocoder:

Table 4 Minimum geocoder user requirements identified from available survey data

Question Category	Priority
Level of control registries have over geocoding activities	A geocoder should enable a user to feel as if they are in control of the geocoding processing decisions
Reference Datasets and Sources	A geocoder needs to be able to support different types of reference data, from different sources The user needs to have the ability to control the sources of reference data used by the geocoder, on a per-record basis following a set of criteria or automatically, such as by spatial extent and/or accuracy requirements A geocoder needs to support multiple data sources simultaneously and allow the user to switch between them, so different sources can be used as deemed appropriate during the geocoding processes A geocoder needs to report how and why it chooses to use a given reference dataset to geocode an input address A geocoder needs to permit users the ability to utilize their own reference data A geocoder needs to support historical reference data
Input Data and Sources	A geocoder should support many different types and formats of input data A geocoder should take advantage of diverse address formats by processing them differently using specialized approaches A geocoder should maintain a set of rules for identifying types of input data A geocoder needs to be able to standardize to USPS A geocoder standardization algorithm does not need to be CASS certified, but does need to support a standard base set of deterministic operations (rules), and/or support probabilistic approaches A geocoder should be used to validate input data after reporting using a given reference dataset A geocoder should be able to process historical data, but does not need to process it differently A geocoder should provide guidance on which supplemental sources of data can be used for address investigations and how to incorporate them into the address record while notating their inclusion A geocoder should be able to revert to a lower resolution geographic feature if one fails to match

Table 4. Cont.

Question Category	Priority
Geocode Output	A geocoder should be able to output geographic points with appropriate geographic metadata
	A geocoder should report metadata with its output that describes the geocoding process info as well as accuracy info about the output
	A geocoder needs to be able to output text and geographic spatial geometries
	A geocoder should be able to generate output in the form of text files, non-spatial databases, ESRI geodatabases, and ESRI shapefiles
	A geocoder should be able to convert between it's own output format and other formats as well
	Geocode output needs to be able to meet the needs of a wide variety of consumers
Geocoder Accuracy	Investments into geocoding accuracy improvement projects should focus on reference datasets (improving the completeness, accuracy, and how it can be utilized under uncertainty) and input data (cleaning and validation)
	A geocoder should be able utilize measures of completeness and spatial accuracy of attribute accuracy of reference datasets, as well as accuracy measures for individual regions and across regions
	A geocoder needs to derive accuracy metrics from both the reference feature and other geocodes
	A geocoder should report accuracy metrics for the whole process, each component, and each individual geocode
	A geocoder needs to report feature match type and the hierarchy used in feature matching
	A geocoder must be able to support multiple feature matching hierarchies that can be user-defined and user-selectable
Geocoder Accuracy (cont.)	A geocoder should report accuracy metrics that express the probability of a correct feature match based on the supports of the match (percentage of attributes matched, percentage of attributes relaxed, etc.)
	A geocoder should be able to derive estimates of spatial error based on metadata about the components of the geocoding process (probability that a feature matched correctly, area of the matched feature, etc.)
	A geocoder should enable a user to select and guarantee minimum and maximum levels of spatial accuracy based on registry requirements and application usage
	A geocoder should be able to provide a feature matching certainty
	A geocoder needs to be able to operate with reference data sources that cover a wide range of spatial accuracies
	A geocoder needs to allow users to utilize historical reference data

Table 4. Cont.

Question Category	Priority
Software	A geocoder should to be flexible and customizable, but for the most part people will use what is available Multiple geocoding strategies need to be able to be employed simultaneously
Geocoding Methodology	Vendors and geocoding processes need to provide detailed metadata about the internal workings of their geocoding processes with regard to the components used and metadata about each such as the type of reference data with its vintage, lineage, and spatial accuracy estimates Registries need to document their geocoding processes A detailed and comprehensive list of geocoding assumptions needs to be developed and utilized within the documentation of a geocoding process A geocoding platform needs to support a wide range of geocoding options in terms of the types of geocoding that can be performed (manual, interactive, interactive with prompting, batch, single, etc.), and the types of components supported such as the types of data sources that can be use (linear and areal) or the type of matching/interpolation (feature matching only, feature interpolation) A user must be able to control the geocoding process in terms of the types of components used (data sources, interpolation methods, etc.) as well as when and in which order they are applied, i.e. control the hierarchy used Sub parcel geocoding must be supported
Linear-Based Interpolation Methodology	Linear-based interpolation must be supported by a geocoding process Dropback distance values must be used in the linear-based geocoding processes, and the distance value of the dropback must be user-definable and modifiable A linear-based interpolation algorithm 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 street segment

Table 4. Cont.

Question Category	Priority
Areal-Based Interpolation Methodology	Areal unit-based interpolation must be supported in a geocoding process Areal unit-based interpolation must be at least attemptable for all classifications of input data, e.g. rural/urban, current/historical Centroid calculations for deriving an output from an areal unit must be supported Centroid weighting does not appear to be required (although all indications in emerging literature recommend otherwise) A user must be able to turn on and off centroid acceptance criteria A user must be able to supply centroid acceptance criteria such as the constraint that the centroid has to be within the areal unit
Manual Geocoding Methodologies	Manual geocoding must be a supported geocoding option A manual geocoding process must support user customization in terms of the data sources used A consistent standardized protocol for performing manual geocoding needs to be developed and utilized
Feature Matching Methodology	Match rates must be reported along with geocoded data Both deterministic and probabilistic feature matching algorithms must be supported and available, and a user must have the ability to decide which algorithm to use when The uncertainty cutoffs for probabilistic matching must be user-definable The ability to use attribute relaxation approaches must be an option, 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 turn on and off the ability to break feature matching ties, and the criteria used must be user-definable The user must be able to choose whether or not to use SOUNDEX A user must be able to choose whether or not to treat historical data separately from temporally current data With regards to standardization of feature matching methods, all registries should use the same attribute relaxation hierarchy, and use the same components and a single consistent match score for probabilistic matching

Table 4. Cont.

Question Category	Priority
Confidentiality	Physical and logical (authentication) security of the geocoded data itself must be ensured Confidentiality and privacy of the information contained in the geocoded data must be ensured, and support must exist for a variety of methods to accomplish each Both individual level and geographically masked geocodes must be provided to consumers Multiple forms of geographic masking should be supported and available for a user to choose from including randomization, aggregation, resolution lowering Consumers must be able to securely obtain digital geocoded data via electronic transmission and standard mail services
Historical Input Data	Issues regarding utilization of historical input addresses should be further investigated in follow-up user needs assessments

8 Conclusions

This research report in the second in a series of three reports about geocoding best practices. This particular report provides details on past, present, and future geocoding needs of the cancer registry community throughout North America. The needs developed within this work have been gathered though a thorough evaluation and synthesis of both published survey on the topic and the results of a new survey. Taken together, the participants in the three surveys form a diverse group of registries in terms of knowledge, practical skill, capacity and level of sophistication, ranging from those just starting out to those with more than a decade of experience.

In regards to differences in responses between the 2006 and 2008 surveys, the respondents' understanding of the level of control they had over geocoding activities in 2006 versus 2008 indicate that though in the past they believed that had control over the geocoding process, they now have a more realistic view of the process and their own level of control over it. Although 67% of the registry staff who responded to the 2006 survey were actually performing geocoding activities, the caseload per registry as well as the number of cancer registry staff participating in these activities has risen to 75% in 2008. Also, the percentage of registries responding that any type of geocoding is performed, either within the registry or using a third party vendor, has risen from 82% in 2005 to 94% in 2008.

The availability or accessibility of good quality (reliable) reference data stands out as one of the primary needs of the geocoder user community. In the near future the level of priority of this requirement may abate due to improvements in the availability of key reference datasets. For instance, as of 2008 TIGER/Line data is being made publicly available as shapefiles, a ubiquitous GIS data format (US Census Bureau 2008). In addition, parcel data is becoming more freely available, largely due to recent changes in their legal status (e.g. see Lockyer (2005) for details about recent changes in California). These advancements coupled with the current interest in developing open source and extensible geocoding platforms will undoubtedly have a large impact on addressing the current and future needs of the cancer registry and research communities.

It is recommended that additional needs assessment activities should be considered in a future phase of this research. The activities should concentrate on obtaining more information regarding the types of metadata typically collected by cancer registries, geared toward development of methodologies for collection, formatting and interpretation of metadata. Furthermore, it is also recommended that in the future, research methods or protocols for defining the accuracy of reference datasets used in geocoding be developed for cancer registries.

9 Acknowledgements

The development of report was funded by the CDC. It draws upon the work of NAACCR (2008a) and Goldberg (2008a), which has been funded in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, and from the Centers for Disease Control and Prevention.

10 References

ESRI, 2008, *Defining the Address Locator Components. ArcGIS 9.2 Desktop Help*. WWW Document available at http://webhelp.esri.com/arcgisdesktop/9.two out of index.cfm?TopicName=Defining_the_address_locator_components (June 20, 2008)

Goldberg DW 2008a *Geocoding Best Practices Survey*. WWW Document available at http://ahf410-pc4.usc.edu:8080/NAACCR/survey.jsp (June 20, 2008)

Goldberg DW 2008b *A Geocoding Best Practices Guide*. In preparation. Springfield, Il. North American Association of Cancer Registries

Goldberg DW, Swift JN and Wilson JP 2008a *Geocoding Best Practices: Reference Data, Input Data and Feature Matching*. Los Angles, CA, University of Southern California GIS Research Laboratory Technical Report No 8

Goldberg DW, Swift JN and Wilson JP 2008b *Geocoding Capacity Survey*. WWW Document available at https://webgis.usc.edu/Surveys/CDC/ (June 20, 2008)

Group 1 Software Inc. 2008 *Centrus - Business Geographics, Data Quality, Real-Time Customer Matching*. WWW Document available at http://www.centrus.com (June 20, 2008)

Lockyer B 2005 *Office of the Attorney General of the State of California Legal Opinion 04-1105*. Available online at http://ag.ca.gov/opinions/pdfs/04-1105.pdf (June 20, 2008)

NAVTEQ 2008 *NAVSTREETS*. WWW Document available at http://developer.navteq.com /site/global/dev_resources/170_navteqproducts/navdataformats/navstreets/p_navstreets.jsp (June 20, 2008)

NAACCR 2008a *NAACCR GIS Committee*. WWW Document available at http://www.naaccr.org/committees/gis (June 20, 2008)

NAACCR 2008b *Geographic Information Systems Survey*. WWW Document available at http://www.naaccr.org/filesystem/word/GIS%20survey_Final.doc (June 20, 2008)

NAACCR 2008c *NAACCR Results of GIS Survey*. Spring 2006 NAACCR Newsletter. WWW Document available at http://www.naaccr.org/index.asp?Col_SectionKey=6&Col_Content ID=497 (June 20, 2008)

Pitney Bowes Software Inc. 2008 *Pitney Bowes MapInfo: The Leading Global Provider of Location Intelligence Solutions*. WWW Document available at http://www.mapinfo.com (May 23rd, 2008)

Tele Atlas Inc. 2008 *Dynamap Map Database*. WWW Document available at http://www.teleatlas.com/OurProducts/MapData/Dynamap/index.htm (June 20, 2008)

US Census Bureau 2008 *U.S. Census Bureau - TIGER/Line*. Washington, DC, United States Census Bureau. WWW Document available at http://www.census.gov/geo/www/tiger/ (June 20, 2008)

United States Postal Service 2008 *CASS Mailer's Guide*. Washington, DC United States Postal Service. WWW Document available at http://ribbs.usps.gov/doc/cmg.html (June 20, 2008)

11 List of Terms

Abbreviation	Description
CDC	Centers for Disease Control and Prevention
DCPC	Division of Cancer Prevention and Control
GIS	Geographic Information System
IRB	Institutional Review Board
NAACCR	North American Association of Central Cancer Registries
NCI	National Cancer Institute
NGC	Northrop Grumman
NPCR	National Program of Cancer Registries