The Public Health Disparities Geocoding Project Monograph

Geocoding and Monitoring US Socioeconomic Inequalities in Health: An introduction to using area-based socioeconomic measures

GEOCODING

Geocoding vs. GIS

GIS and Geocoding are two terms that you've probably been hearing a lot about recently.
What are they exactly?

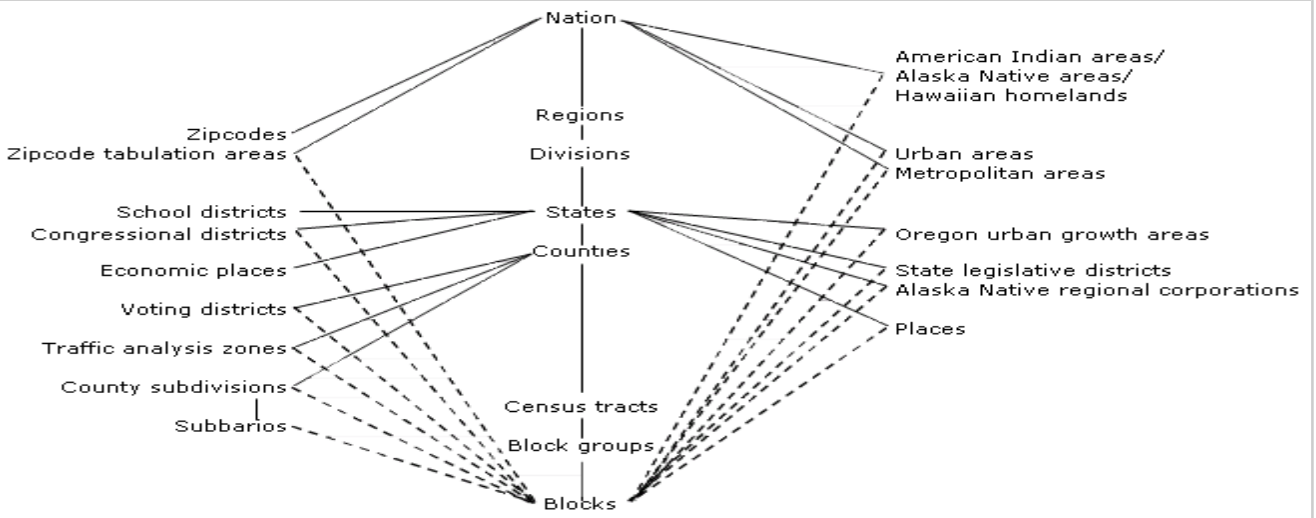
GIS – Geographical Information Systems – are technology based systems that combine layers of geographic data to give you a better understanding of a particular place[1]. For example, you might combine a layer of cholera outbreaks with a layer of water sources to be able to display graphically the relationship between the two. For more examples of GIS technology at work, visit www.esri.com.

Geocoding is the assignment of a code – usually numeric -- to a geographic location. (So, one geocode that you're probably already familiar with is your ZIPcode.) Usually however, when someone talks about geocoding, they are talking about geocodes that are a bit more specific, i. e., affixing to an individual address its latitude and longitude – which is, very simply, the vertical and horizontal distance of a point relative to the equator[2]. Once the latitude and longitude are known, you can then figure out all sorts of other geocodes to affix by determining what geographic regions the specified point lies in, e.g., what ZIPcode does this point lie in? What census tract? What census blockgroup? What police precinct? Appending any of these codes to a specific street address is considered geocoding.

Census Geography

Our project utilized primarily geocoding technology. Before continuing to discuss geocoding, it's important that you know a bit about census geography, since the geographic code that is typically affixed to an address during geocoding is either the U.S. Census Bureau defined census tract or blockgroup.

Figure 1. **Geographic Hierarchy for the 1990 Decennial Census**



The above figure displays the hierarchy of census geography[3]. As you may have already read in other sections of this monograph, we strongly recommend the census tract as the geographical unit of analyses. Census tracts, census blockgroups, and the “new to 2000” ZIP Code Tabulation Areas are U.S. Census Bureau defined, standardized, and relatively permanent geographical units. Census tracts are constructed specifically to include on average 4,000 people of fairly homogeneous population characteristics, economic position, and living conditions. Federal, state, and local governments routinely use census tracts as administrative units. For example, the Federal government uses census tracts to define urban empowerment zones and decide who's eligible for low-income housing tax credits. Census tracts are sub-divided into blockgroups -- which have an average population size of about 1,000. See Figure 2 below. Notice that ZIPcodes are off to the side, in a category all by themselves, and not linkable to anything else. In contrast to the census tracts and blockgroups, ZIP codes are U.S. Postal Service administrative units that are subject to change at any time, thus making the linking of ZIPcode level data to other datasets, e.g., the decennial U.S. Census data a bit questionable. They are far from standardized – a ZIPcode can designate a single office building or entire state county. For a more thorough discussion of the problem of using ZIPcodes in area-based analyses, please refer to our article “Zip Code caveat: bias due to spatiotemporal mismatches between ZIP Codes and US census-defined areas—the Public Health Disparities Geocoding Project”[4].

Figure 2. **Census Tracts, Block Groups, and Blocks**[5]

Programs vs. Services

For this project, we geocoded Massachusetts Department of Public Health and Rhode Island Department of Health data to the blockgroup level. Before we eventually used a commercial geocoding firm to geocode our data, we considered three things: accuracy, cost, and turnaround time. In 1999, when we explored our geocoding options, there were a handful of commercial services and two stand-alone geocoding programs available. Now, there are many more commercial geocoding services, and quite a few stand-alone programs to choose from. However, not all companies or programs are the same. So, first determine what makes the most sense for your project: using a geocoding service or using a program to do the geocoding yourself.

Considerations include time – are you working on a tight schedule, or do you have enough time for you, or someone on your staff, to become proficient with a geocoding program? Keep in mind that some of the programs have very steep learning curves. Becoming proficient at geocoding will take months, and becoming an expert may take years. The benefit to having a trained in-house geocoding specialist is that, over time, depending on the volume of your data, it may be cheaper to geocode in-house, and you have the additional benefit of having more control over the geocoding process. (Click here for more on Geocoding Programs vs. Geocoding Services.)

Testing Accuracy

If you decide to use a geocoding service, we recommend that you do a bit of testing to make sure you get the most accurate results. Many companies advertise high completion rates, that is, the percentage of addresses that they geocode, but completeness and accuracy are two different things. How do you know if they've geocoded the addresses to the right place? To test the accuracy of geocoding services, we recommend the following plan6.

First, generate a test file. This you'll do by performing some "old-fashioned" geocoding. Pull together a list of 50-75 addresses that you're familiar with. They should be spread across as large a geographic area as possible, but concentrated in the area that the majority of your data (the addresses you are eventually planning to geocode) will be from. On a street map (or more than one street map if your addresses cover a large enough area), locate and mark the exact locations of the addresses. Take this map to your regional Census Bureau office. Using the official Census Bureau blockgroup maps available there, identify the blockgroup that each address falls in. To create the full blockgroup geocode, use the following scheme:

Digits 1-2 = State code
Digits 3-5 = County code
Digits 6-11 = Census Tract code (often used with a decimal point: xxxx.xx)
Digit 12 = Blockgroup code

U.S. Census FIPS Areakey

250131402013

(Scroll your mouse over this figure for more detail.)

You'll be able to get all of the components that make up the areakey from the blockgroup maps at the census bureau. For more information about blockgroups and other units of census geography, check out The Census Geographic Areas Reference Manual3. (The U.S. Census Bureau Website is a great place to familiarize yourself with a lot of subjects that we'll be focusing on in this monograph, e.g., Census data, Census geography, area-based measures, geocoding, GIS, and mapping.)

Congratulations! You've just (a) successfully geocoded your data to the blockgroup level; and (b) created a test file. You can now use this file to test commercial geocoding firms and geocoding programs alike.

Send a file containing only the addresses to the prospective geocoding companies and then compare the results sent back from the company to the correct geocodes you ascertained at the Census Bureau office. As an external check of both you and the geocoding companies, submit your addresses to the Census Bureau Census Tract Locator on the American FactFinder website. If you opted to use a geocoding program, you can also use this test file to test your own results.

Address Cleaning

Now that you have your geocoding plan of action ready – whether it's using a geocoding program yourself or sending your data out to a geocoding service -- the next step is to clean your addresses. Geocoding follows the time tested theorem "garbage in, garbage out". If your addresses are not clean, then you are significantly increasing the probability that they will not be geocoded correctly.

Cleaning addresses means:
• retaining only the key address elements in one field: house/ building number; street name; street type; e.g., 100 Main St
• getting rid of all extraneous characters, e.g., "BSMT" "REAR" "APT 1" "UNIT 3", etc.
• standardizing spelling, e.g., converting all incidences of "Mass Ave" to "Massachusetts Ave"

Some examples:

Record #	Original Address	"Cleaned"
1	677 Huntington, #304	677 Huntington Ave
2	46 Burr REAR	46 Burr St
3	Unit B, 1200 Comm Ave.	1200 Commonwealth Ave
4	423 Allston St., 4th Floor, Suite 100	423 Allston St
5	The Landmark Building, 401 Park Drive	401 Park Drive
6	99 ½ Chauncey St	99 Chauncy St

What about those pesky P.O. Box addresses and "Rural Route" addresses with no house numbers?

• The geocoding program will look at "P.O. Box" as if it's a street name, so if there's a "Postbox St." in your neighborhood, you may get false matches.

• The individual who has this P.O.Box as a mailing address may not necessarily live in the blockgroup, census tract, or even the ZIPcode that the post office is in.

• Check a map. Does the entire rural route lie in a single census tract? Or in a single blockgroup? If so, the geocodes may be accurate since ALL structures on that route fall in the same census tract.

• Decide ahead of time on a method of dealing with P.O. Boxes and Rural

Route addresses in your analyses. Keep in mind that these addresses are often not geocodable anyway.

For more detail about cleaning and formatting addresses, speak with the Customer Service representative at the geocoding service, or check to see what format your geocoding program requires. Also note that there are a number of products on the market that will clean addresses for you. We have not evaluated them however, and so can not advise you regarding their efficacy or accuracy.

Address Formatting

The typical format of a file to be sent to a geocoding service (Excel or dbf format):

Record #	Street Address	City	State	ZIPcode
1	677 Huntington Ave	Boston	MA	02115
2	46 Burr St	Jamaica Plain	MA	02130
3	1200 Commonwealth Ave	Boston	MA	02215
4	423 Allston St	Cambridge	MA	02139
5	401 Park Drive	Boston	MA	02215
6	99 Chauncy Street	Boston	MA	02111

The typical format of a file returned from a geocoding service (Excel or dbf format):

Record #	Street Address	City	State	ZIPcode	Latitude	Longitude	Areakey	Match Code
1	677 Huntington Ave	Boston	MA	02115	-71.10	42.34	25025081000	AS0
2	46 Burr St	Jamaica Plain	MA	02130	-71.11	42.32	25025120600	AS1
3	1200 Commonwealth Ave	Boston	MA	02215	-71.12	42.35	25025000801	AS7
4	423 Allston St	Cambridge	MA	02139	-71.11	42.36	25017353200	ZB7I
5	401 Park Drive	Boston	MA	02215	-71.10	42.34	25025010200	AS0
6	99 Chauncy Street	Boston	MA	02111	-71.06	42.35	25025070100	ZB7L

The MatchCode variable (also called “georesult” by some companies) is an indicator of which address elements determined the geocode, and how certain the geocoding program is about the accuracy of the geocode. For example, the MatchCode of AS0 indicates that the geocode was derived based on the street address and matched exactly to a street segment in the program; the program is certain of blockgroup level accuracy. A MatchCode of ZC5Y indicates that the geocode assigned is based upon the location of the post office that delivers mail to that address, and the geocoding program is only comfortable claiming county level accuracy. (This is typically the MatchCode assigned to a P.O.Box address.)

A full explication of the MatchCodes will be provided to you by the geocoding service you employ, or in the technical notes of the program that you use.

Once you have your geocoded file, you should check for any discrepancies in the geocoding. Use SAS, or some other data analyses program, to look for differences in match rates by your variables of interest. At the very least, check for differences in geocoding rates by age, gender, race/ ethnicity, socioeconomic data (if available). What are possible explanations for these differences – and how will they affect your analyses?

REFERENCES

1	GIS.COM. www.gis.com/whatisgis/index.html Accessed February 5, 2004.
2	Stern, DP. From Stargazers to Starships. http://www-istp.gsfc.nasa.gov/stargaze/Slatlong.htm Accessed February 5, 2004.
3	Bureau of the Census, U.S. Department of Commerce. Geographic Areas Reference Manual. Washington, DC: Bureau of the Census, 1994. http://www.census.gov/geo/www/garm.html Accessed February 5, 2004.
4	Krieger N, Waterman P, Chen JT, Soobader M-J, Subramanian SV, Carson R. Zip Code caveat: bias due to spatiotemporal mismatches between ZIP Codes and US census-defined areas—the Public Health Disparities Geocoding Project”. American Journal of Public Health 2002; 92:1100-1102. American Journal of Public Health 2002; 92:1100-1102.
5	U.S. Department of Commerce, Census ’90 Basics. Washington, D.C.: U.S. Government Printing Office, 1990.
6	Krieger N, Waterman PD, Lemieux K, Zierler S, Hogan JW. “On the Wrong Side of the Tracts? Evaluating accuracy of geocoding for public health research.” American Journal of Public Health 2001;91:1114-16.).

This work was funded by the National Institutes of Health (1RO1HD36865-01) via the National Institute of Child Health & Human Development (NICHD) and the Office of Behavioral & Social Science Research (OBSSR).