ARTICLES

Geocoding to Create Survey Frames

Stephanie Eckman¹, Ned English²

¹ Institute for Employment Research, ² NORC at the University of Chicago

Keywords: geocoding, gis, face-to-face surveys, household surveys, sampling, address based sampling, frames

<https://doi.org/10.29115/SP-2012-0026>

Survey Practice

Vol. 5, Issue 4, 2012

Geocoding to Create Survey Frames

With the Delivery Sequence File (DSF), from the United States Post Office, surveys can cheaply and easily create address frames and samples. Many studies have examined the coverage of these frames (see for example Dohrmann, Han, and Mohadjer 2006; Iannacchione, Staab, and Redden 2003; O'Muircheartaigh, Eckman, and Weiss 2002). However, these studies do not discuss geocoding.

Geocoding is a key step in turning the DSF into a survey frame. Survey researchers who use these frames should understand the role geocoding plays, whether they do this work themselves or buy already-geocoded frames or samples.

Geocoding is necessary because there is a mismatch between the geographies on the DSF and those used in most surveys. The DSF contains only street address, city, state, zip code, and other fields related to mail delivery. Household samples, however, are often based on census geographies such as counties, tracts, and blocks. Geocoding translates the address data into census blocks.

We have learned a lot about the geocoding process over the past ten years of work with the DSF. In this article, we share what we have learned. We explain what geocoding is and how it works. We also discuss what can go wrong.

Geocoding is a two-step process. First an address is assigned a geographic coordinate (usually latitude and longitude). Then the coordinate is mapped to census geography. All addresses placed in tracts or blocks selected for the survey are part of the frame.¹

¹ There are two common geocoding software programs: ArcGIS, from ESRI, and MapMarker Plus, from Pitney Bowes Business Insight (formerly MapInfo). This article focuses on MapMarker Plus, but the two programs work similarly.

Figure 1 Example of street-level geocoding.

Step 1: Coordinate Assignment

To assign coordinates, the software compares each address to a database of street segments and house-number ranges. The database contains the location of the centerlines of street segments and the even/odd house number patterns. The program finds the street segment that matches the address, and interpolates the location of the address within the segment (Zandbergen 2008).

Consider the example address:

7422 Baltimore Avenue
College Park, MD 20740-3208

The address is first matched to the 7400–7499 street segment of Baltimore Avenue inside the 20740 zip code. The software then places the address 22 percent of the way down the block, on the even-numbered side of the street. See Figure 1 for an example.

When several units have the same street address, such as the units in an apartment building, all receive the same coordinate (Pitney Bowes MapInfo 2008). Note that the software will also geocode addresses that do not exist, if they fall in a valid address range.

This coordinate assignment method is called street-level geocoding. Eckman and English (2012) show that 83.3 percent of all residential addresses on the DSF geocode at this level. It is the most precise nationally available method of geocoding in the United States (Zandbergen 2008).

Sometimes the software is not able to find the street segment in the zip code. In these cases, it will use a less precise method. Postal-level geocoding assigns a coordinate based on the zip code. The software will attempt to geocode to the zip + 4 level.² If found in the database, the address is assigned to the centroid of that zip + 4.

² The zip + 4 is the full nine digit zip code assigned by the United States Postal Service, 20740-3208 in the example. For more information on how United States' zip codes are structured, see <http://www.usps.com/faqs/ziplookup-faqs.htm>.

If the zip + 4 code is not present on the address or is not found in the database, the software next tries geocoding to zip + 2. If that method also fails, the software will use the five digit zip code centroid. (The three-digit zip code and city-level geocoding are also possible, if all else fails. We find these are not needed when geocoding the DSF.)

Ideally, the assigned coordinates are very close to the true location of the address. Studies of the distances between the two points found errors of around 50 to 200 meters, and larger in rural areas than in urban areas (Bonner et al. 2003; Cayo and Talbot 2003; Morton et al. 2007; Schootman et al. 2007; Strickland et al. 2007; Ward et al. 2005; Whitsel et al. 2004, 2006).

Coordinates assigned by street-level geocoding are more likely to be close to their true location than those assigned by postal geocoding. zip + 4 codes refer to small areas, and geocoded coordinates assigned at this level may be accurate. In urban areas, zip + 4 codes are often one side of a census block, or floors of a large building. zip + 2 and zip geocodes are less accurate.

However, even street-level geocoding can be far off the mark. Sometimes the database is wrong about which side contains the even numbers and which the odd (O’Muircheartaigh, Eckman, and Weiss 2002; Schilp 2005). This error can affect block assignment in step 2, which is crucial for making a high quality survey frame.

Step 2: Block Assignment

The second step of the geocoding process translates the address’s coordinate into a census block code. MapMarker Plus lays a block layer over the coordinates. Each address is assigned to the block that it falls into. If the address is assigned to a block selected for the survey, it becomes part of the frame.

See Figure 2 for an illustration. In the first panel, addresses are geocoded (the stars). They are assigned to the block where the star lies. Blocks 2002, 2003, and 2004 are selected for the survey. In the second panel, only addresses assigned to these blocks are on the frame.

Ideally, this process places addresses into the block where they really are. Investigations of the accuracy of assignment to census geographies have reported that 35% of addresses are placed in the wrong blocks, and five percent in the wrong tract (Krieger et al. 2001; Morton et al. 2007; Ratcliffe 2001; Schootman et al. 2007; Strickland et al. 2007).

When addresses are placed into the wrong block, the frame may have problems of undercoverage or overcoverage. Undercoverage happens when the frame excludes addresses that are inside the selected area. Undercoverage in face-to-face surveys is hard to detect and can lead to bias. Although missed unit techniques have been proposed to fix undercoverage, their performance is not promising (Eckman and O’Muircheartaigh 2011; McMichael et al. 2008).

Figure 2 Geocoded addresses become a survey frame.

Overcoverage happens when the frame includes addresses that are not valid housing units or are outside the area. Some overcoverage is easy to fix. Interviewers can identify non-residential addresses. However, when an address in Block 2001 is incorrectly placed into Block 2003, that unit is overcovered. Such errors can be hard for interviewers to notice, or may cause them confusion. This type of overcoverage can also lead to bias.

Correct block placement depends on how the coordinate is assigned in step 1. When postal-level geocoding is used, block assignment is likely to be wrong. Addresses that geocode to the ZIP or ZIP+2 centroid are assigned to the block that contains the centroid. This block will be correct only by chance.

Block assignment is more likely to be correct when an address geocodes at the street-level. However, even these coordinates can be placed in the wrong block, especially if there are side-of-street errors. The odd and even sides of a street are often in different census blocks. There are no estimates of how often side-of-street errors occur, or of undercoverage and overcoverage due to such errors.

There is another type of error that can lead to incorrect block assignment. Sometimes the database used to assign coordinates and the block layer used to assign block codes do not line up. We call this the layer-offset problem.

Figure 3 Layer-offset problem.

Figure 3 shows an example. This map is made up of two layers. One is the street database used in geocoding. The other is the block layer used to assign coordinates to census geographies. The shaded area in Figure 3(a) indicates the blocks selected for inclusion in the survey. The survey frame will be made up of all addresses whose coordinates fall inside this shaded area.

Zooming in to the northwest corner of the selected area shows that the block (shaded) layer does not line up with the street layer. In Figure 3(b), we can see that while the boundary of the shaded area is meant to be Matau Way, the block layer contains a kink in the street that is not present in the street layer. This issue can lead to undercoverage and overcoverage, but there are no estimates of how often this error occurs in the map layers.

Most commercially-available map data are derived from Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER) data. We hope that the Census Bureau's project to improve TIGER data for the 2010 Census will reduce the side-of-street and layer-offset issues.

We want researchers to be knowledgeable consumers of geocoded data. Researchers who purchase frames or samples from the DSF should know where the data come from and how they are geocoded.

Armed with this information, different surveys will make different choices. For example, a survey which plans to merge in the distance from each selected address to the nearest hospital may decide to use only addresses that geocode at the street-level. This approach has a net coverage rate of 86.7 percent nationally.

Another survey may decide not to worry about geocoding accuracy but can use only addresses that an interviewer can visit (no post office of similar addresses). This approach has a net coverage rate of 92.3 percent nationally. Both of these coverage rates vary considerably by state, which raises concerns for regional surveys (reanalysis of data in Eckman and English 2012).

The role of geocoding in surveys is sure to increase in the next decade. This article has provided some background about the geocoding process in the context of frame creation. However, geocoding is used not only to make frames, but also in data collection and analysis. English and Pedlow (2005) discuss using geocoding to assign interviewers to cases. Nusser (2007) reviews other uses of geocoding and other types of geographic information systems (GIS) in surveys. We hope this article inspires survey researchers to learn more about GIS tools – how they can improve survey data and the errors they can introduce.

REFERENCES

- Bonner, M.R., D. Han, J. Nie, P. Rogerson, J.E. Vena, and J.L. Freudenheim. 2003. "Positional Accuracy of Geocoded Addresses in Epidemiologic Research." *Epidemiology* 14 (4): 408–12.
- Cayo, M.R., and T.O. Talbot. 2003. "Positional Error in Automated Geocoding of Residential Addresses." *International Journal of Health Geographics* 2 (10): 10.
- Dohrmann, S., D. Han, and L. Mohadjer. 2006. "Residential Address Lists vs. Traditional Listing: Enumerating Households and Group Quarters." In *Proceedings of the Section on Survey Research Methods*, 2959–64. American Statistical Association.
- Eckman, S., and N. English. 2012. "Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates." *Field Methods*. Forthcoming.
- Eckman, S., and C. O'Muircheartaigh. 2011. "Performance of the Half-Open Interval Issued Housing Unit Procedure." *Survey Research Methods* 5 (3): 125–31.
- English, N., and S. Pedlow. 2005. "Using GIS to Improve Field Interviewing Efficiency: Enhanced Interviewer Selection and Sample Allocation." In *Proceedings of the Section on Survey Research Methods*, 2981–86. American Statistical Association.
- Iannacchione, V.G., J.M. Staab, and D.T. Redden. 2003. "Evaluating the Use of Residential Mailing Addresses in a Metropolitan Household Survey." *Public Opinion Quarterly* 67 (2): 202–10.
- Krieger, N., P. Waterman, K. Lemieux, S. Zierler, and J. Hogan. 2001. "On the Wrong Side of the Tracts? Evaluating the Accuracy of Geocoding in Public Health Research." *American Journal of Public Health* 91 (7): 1114–16.
- McMichael, J.P., J.L. Ridenhour, S. Michell, K. Fahrney, and W. Stephenson. 2008. "Evaluating the Use and Effectiveness of the Half-Open Interval Procedure for Sampling Frames Based on Mailing Address Lists in Urban Areas." In *Proceedings of the Section on Survey Research Methods*, 4251–57. American Statistical Association.
- Morton, K., V. Iannacchione, J. McMichael, J. Cajka, R. Curry, and D. Cunningham. 2007. "Linking Mailing Addresses to a Household Sampling Frame Based on Census Geographies." In *Proceedings of the Section on Survey Research Methods*, 3971–74. American Statistical Association.
- Nusser, S.M. 2007. "Discussion: Using Geospatial Information Resources in Sample Surveys." *Journal of Official Statistics* 23 (3): 285–89.
- O'Muircheartaigh, C., S. Eckman, and C. Weiss. 2002. "Traditional and Enhancedfield Listing for Probability Sampling." In *Proceedings of the Section on Survey Research Methods*, 2563–67. American Statistical Association.
- Pitney Bowes MapInfo. 2008. *MapMarker Version 14 Developer's Guide*. Troy, NY: Pitney Bowes Software, Inc.
- Ratcliffe, J.H. 2001. "On the Accuracy of TIGER-Type Geocoded Address Data in Relation to Cadastral and Census Areal Units." *International Journal of Geographical Information Science* 15 (5): 473–85.
- Schilp, J. 2005. "Geocoding Procedure to Find Geographic Identifiers in the Housing Component of the Consumer Price Index." In *Proceedings of the Section on Government Statistics*, 1436–38. American Statistical Association.
- Schootman, M., D.A. Sterling, J. Struthers, Y. Yan, T. Laboube, B. Emo, and G. Higgs. 2007. "Positional Accuracy and Geographic Bias of Four Methods of Geocoding in Epidemiologic Research." *Annals of Epidemiology* 17 (6): 464–70.
- Strickland, M.J., C. Siffel, B.R. Gardner, A.K. Berzen, and A. Correa. 2007. "Quantifying Geocode Location Error Using GIS Methods." *Environmental Health* 6: 10.

- Ward, M.H., J.R. Nuckols, J. Giglierano, M.R. Bonner, C. Wolter, M. Airola, W. Mix, J.S. Colt, and P. Hartge. 2005. "Positional Accuracy of Two Methods of Geocoding." *Epidemiology* 16 (4): 542–47.
- Whitsel, E.A., P.M. Quibrera, R.L. Smith, D.J. Catellier, D. Liao, A.C. Henley, and G. Heiss. 2006. "Accuracy of Commercial Geocoding: Assessment and Implications." *Epidemiological Perspectives and Innovations* 3: 8.
- Whitsel, E.A., K.M. Rose, J.L. Wood, A.C. Henley, and G. Heiss. 2004. "Accuracy and Repeatability of Commercial Geocoding." *American Journal of Epidemiology* 160 (10): 1023–29.
- Zandbergen, P.A. 2008. "A Comparison of Address Point, Parcel and Street Geocoding Techniques." *Computers, Environment and Urban Systems* 32: 214–32.