The Implications of Geocoding Error on Address-Based Sampling

Bonnie E. Shook-Sa[1], Joseph P. McMichael[1], Jamie L. Ridenhour[1]
Vincent G. Iannacchione[2]
[1] RTI International, 3040 East Cornwallis Road, Research Triangle Park, NC 27709
[2]RTI International, 701 13th Street, N.W., Suite 750, Washington, DC 20005

Abstract
In-person surveys that use address-based sampling are often based on area segments defined by census geography rather than postal geography. Census geography enables more accurate inclusion of demographic information in the sample selection procedures and the use of frame supplementation methods to increase coverage. However, area frames based on census geography contain more frame error than frames based on postal geography because addresses must be allocated (i.e. geocoded) into area segments. When addresses are incorrectly geocoded into area segments, sampling inefficiencies occur. We examine data from the 2009 National Survey on Drug Use and Health to determine the extent of geocoding error in sampled segments and its implications on coverage and efficiency of area frame samples.

Key Words: frame coverage, in-person surveys, ABS, mailing lists

1. Introduction

Traditionally, field enumeration has been used to create sampling frames for area-based household surveys. While field enumeration has a high level of coverage, it is also very costly and must be completed months in advance of sampling. Address-based sampling (ABS), which utilizes mailing lists obtained from the United States Postal Service's Computerized Delivery Sequence File, can be used to construct sampling frames of locatable mailing addresses for a fraction of the cost and time required to create field enumerated frames.

Mailing lists can be purchased by either postal or census geography. Postal geography is the geography in which the United States Postal Service organizes and delivers mail. It consists of zip codes and carrier or postal routes (i.e., the area within a zip code where the mail is delivered by an individual delivery person). Census geography is the geography in which the United States Census Bureau collects and summarizes data. Census geography includes counties, census tracts, census block groups, and census blocks.

Creating area segments based on postal geography has minimal frame error since the mailing lists are already organized by postal geography. There is no error associated with assigning addresses to zip codes or postal routes. However, there are limited sources of data that are released based on postal geography which makes it difficult to append external demographic data for sample selection and weighting. Census Zip Code Tabulation Areas (ZCTAs) were developed for the 2000 Census by the United States Census Bureau. ZCTAs are formed by combining census blocks within five-digit zip

codes. Census data are summarized at the ZCTA level and are released for the majority of five-digit zip codes in the country. However, the most recent data available by ZCTAs comes from the 2000 Census, and data will not be released again at the ZCTA level until the 2010 Census.

Postal geography also lacks discernable boundaries that are needed for frame supplementation procedures. Census geography allows for external demographic data to be appended and contains discernable boundaries for frame supplementation. For these reasons, ABS studies typically base area segments on census geography rather than postal geography. Table 1 summarizes the advantages and disadvantages of using postal and census geography in ABS studies.

Table 1: Postal Geography vs. Census Geography for ABS Studies

	Advantages	**Disadvantages**
Postal Geography	• Minimal frame error	• Limited external demographic data available • No discernable boundaries for frame supplementation
Census Geography	• Can easily append external demographic data • Discernable boundaries allow for frame supplementation	• Frame error resulting from allocating addresses into segments

One challenge that emerges from area segments defined by census geography is that mailing addresses must be allocated, or geocoded, into area segments by assigning a latitude and longitude coordinate to each address and evaluating census boundary files to determine to which segment to allocate the address. Geocoding error can occur when addresses are incorrectly allocated across area segments.

Two types of geocoding error can occur at the segment level. Under-coverage geocoding error occurs when addresses are misallocated out of the selected segment. Addresses that are present in the segment are not included on the segment's mailing list. Over-coverage geocoding error occurs when addresses are misallocated into the selected segment. Addresses that are not present within the segment boundaries are included on the segment's mailing list. Figure 1 illustrates the two types of geocoding error.

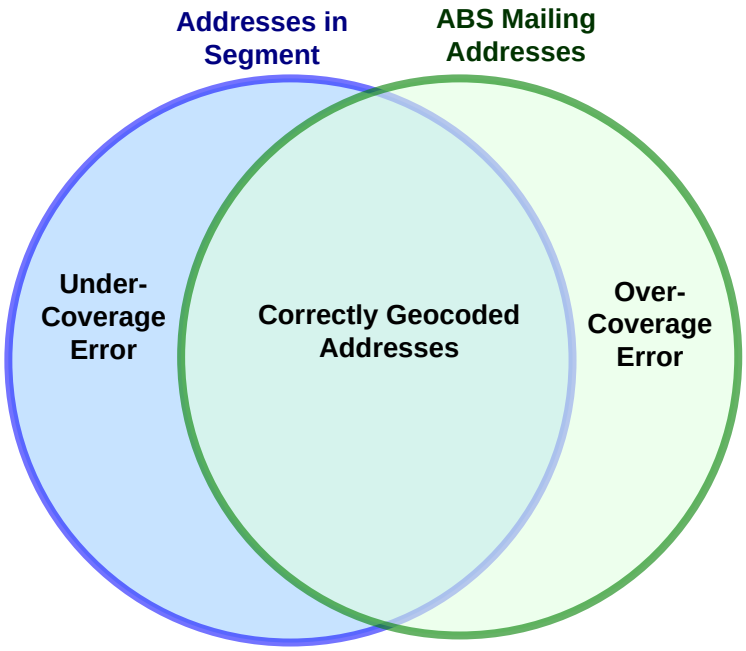

Figure 1: Under-Coverage and Over-Coverage Geocoding Error

Under-coverage error is the more serious type of geocoding error. Addresses that should be included on the sampling frame are incorrectly excluded, and the coverage of the frame will be reduced unless frame-supplementation procedures pick up these excluded addresses. Over-coverage error can result in sampling inefficiencies, because addresses are included on the frame that are not truly eligible for the study since they are not in the selected segment. This can result in higher ineligibility rates and increased field costs.

This paper estimates the prevalence of geocoding error using a nationally-representative sample and examines address and segment characteristics associated with under-coverage and over-coverage geocoding error.

2. Mailing List Field Study

Data for this study come from the National Survey on Drug Use and Health (NSDUH). The NSDUH provides national, state and substate data on substance use and mental health in the civilian, noninstitutionalized population age 12 and older. It is conducted by RTI International under contract with the Substance Abuse and Mental Health Services Administration (SAMHSA). Data are collected throughout the year. Currently the NSDUH sampling frame is created using field enumeration in area segments that are collections of census blocks[1].

The 2009 Mailing List Field Study examined the coverage, potential bias, and cost savings of an alternative sampling frame for the NSDUH. The alternative frame is a hybrid frame that utilizes mailing lists supplemented with the Check for Housing Units Missed (CHUM) frame-linking procedure in segments where mailing list coverage is

[1] Segments for the NSDUH comprise one or more adjacent census blocks that in combination meet or exceed the minimum requirement of 100 DUs in rural areas or 150 DUs in urban areas.

expected to be adequate and retains field enumeration in low coverage segments (Iannacchione et al., 2010).

We sampled 200 segments from the 2009 Quarter 1 NSDUH and purchased the mailing lists associated with these segments. We estimated the mailing list coverage in each of the 200 segments by taking the ratio of locatable mailing addresses in the segment to the estimated number of dwelling units in the segment (McMichael et al. 2010). We oversampled segments where we expected the mailing list coverage to be low and areas with a high percentage of group quarters, since mailing list coverage of group quarters was expected to be poor. Mailing addresses were geocoded into the NSDUH segments using a zip+4 to census block crosswalk. We matched the addresses associated with 3,878 screened and eligible dwelling units[2] from the field enumerated NSDUH frame to the mailing list to estimate the coverage properties of the alternative frame (i.e. the mailing list).

Since geocoding error has coverage and efficiency implications, one component of this field study was to estimate the amount of geocoding error present on the alternative frame and to determine what address and segment characteristics are related to geocoding error. To estimate the prevalence of under-coverage geocoding error, we used the match results from the matching of the 3,878 screened and eligible dwelling units to the mailing list. The locations of these addresses were known from the field enumeration process which allowed us to examine the level of geocoding accuracy. Furthermore, characteristics of these addresses and segments can be analyzed to determine what characteristics are related to under-coverage geocoding error.

To examine over-coverage geocoding error, we took a sample of 1,360 addresses from the mailing list that geocoded into blocks associated with the 200 selected NSDUH segments and matched them to the field enumerated listing. It was cost prohibitive to match all of the mailing list addresses that geocoded into the selected segments to the field enumerated listing since addresses that did not match to the field enumerated listing through the automated matching procedure were matched manually. We stratified the sample by the urban/rural classification of the segment[3] because urbanicity was expected to be highly related to geocoding error. Addresses on the field enumerated list were known to be in segment and were assumed to be a complete listing of dwelling units contained in the segment. Therefore if an address from the mailing list was not on the field enumerated list, it was assumed to have incorrectly geocoded into the segment. These addresses were analyzed to determine the characteristics related to over-coverage geocoding error. They were not used to obtain an estimate of over-coverage geocoding error since they were not investigated in the field like the 3,878 screened and eligible dwelling units were and were thus not fully resolved.

3. Geocoding Results

The results of this research allowed us to estimate and compare the prevalence of geocoding error for different levels of geography. It also allowed us to identify

[2] An eligible DU for the NSDUH is either a housing unit (HU) for a single household or a non-institutional group quarters (GQ) where at least one civilian aged 12 years or older resides for the majority of a calendar quarter.

[3] For a segment to be classified as rural, all of the census blocks in the segment have to be rural. If one or more of the segment's blocks are urban, the segment is also urban.

characteristics of addresses and segments that are related to both under-coverage and over-coverage geocoding error.

3.1 Prevalence of Geocoding Error

Of the 3,878 screened and eligible dwelling units obtained from the NSDUH field enumerated frame, 3,229 matched to the mailing list for an 89.6 percent weighted match rate. The remaining addresses either did not match to the mailing list, were unresolved, or matched to the mailing list but were incorrectly classified as business addresses.

We examined the level of geocoding accuracy for the 3,229 field enumeration to mailing list matches by comparing the true segment, census block group, census tract, and county location of each dwelling unit to its mailing address' geocoded location. The cumulative number of matches at each level of geocoding accuracy and the cumulative weighted percent of matches geocoding at each level is displayed by urbanicity in Table 2. Overall, an estimated 89.9 percent of addresses geocode into the correct segment. The proportion of addresses that geocode into the correct segment increases significantly at the census block group level, where an estimated 99.3 percent of addresses geocode into the correct census block group. This significant increase is consistent with previous research that showed higher levels of geocoding accuracy for larger geographic areas (Morton et al., 2007). This finding supports the claim that larger geographic segments should be used in ABS studies since there is minimal loss in coverage due to geocoding error beyond the census block group level.

Geocoding accuracy is much better in urban segments than rural segments with 92.5 percent of addresses in urban segments geocoding into the correct segment while only 76.6 percent of rural addresses geocode into the correct segment. Small segment sizes in ABS studies that include rural areas can result in significant under-coverage as a result of geocoding error.

Table 2: Cumulative Level of Geocoding Accuracy by Urbanicity

	Overall		Urban Segments		Rural Segments	
Level of Accuracy	Num.	Wtd. Pct.	Num.	Wtd. Pct.	Num.	Wtd. Pct.
Segment	2,689	89.88%	2,273	92.54%	416	76.61%
Census Block Group	3,186	99.27%	2,605	99.82%	581	96.51%
Census Tract	3,226	99.96%	2,619	100.00%	607	99.77%
County	3,229	100.00%	2,619	100.00%	610	100.00%

3.2 Modeling Geocoding Error

After exploring the prevalence of geocoding error at different levels of geographic specificity (i.e. segments, census block groups, census tracts, and counties), we then sought to determine what characteristics of addresses and segments are related to segment-level under-coverage and over-coverage geocoding error. We modeled each type of geocoding error using a logistic regression model in SUDAAN's proc logistic, which takes into account both the study design and the design weights (RTI, 2008). The variables considered as predictor variables and their sources are listed in Table 3.

Variables include both address-level variables obtained from the mailing list and segment-level variables obtained from either the 2000 Census or Census projections obtained from Claritas and Geolytics. The initial models contained main effects for all of the variables in Table 3. To obtain the final model, we excluded insignificant variables and collapsed levels of variables for variables that were significant in the model but not all levels of the variable were significantly different.

Table 3: Potential Geocoding Error Predictors and Sources

Address-Level Variables	Segment-Level Variables
Vacancy Status (Y/N)[1]	Rural/Urban[2]
Drop Point (Y/N)[1]	Census Division[2]
Route Type[1]	Total Number of Dwelling Units[3]
• Street	• < 150
• High Rise	• 150 to 250
Delivery Type[1]	• > 250
• Residential Curb	Area (Square Miles)[2]
• Residential Cluster Box Unit	• < 0.08
• Residential Central	• 0.08 to 1.30
• Residential Other	• > 1.30
	Median Home Value[4]
	• < $100,000
	• $100,000 to $300,000
	• > $300,000
	New Homes[4]
	• Proportion > National Average
	• Proportion ≤ National Average
	Percent Owner Occupancy[4]
	• < 35%
	• 35% to 85%
	• > 85%

[1] Mailing List Classification
[2] Obtained or derived from 2000 Census
[3] Claritas Estimate
[4] Geolytics Estimate

3.2.1 Modeling Under-Coverage Geocoding Error

To determine what characteristics of addresses and segments lead to higher levels of under-coverage geocoding error, we fit a logistic regression model. For each of the 3,229 screened and eligible dwelling units from the NSDUH field enumerated frame that matched to the mailing list, we determined whether the mailing address had geocoded into the correct segment or not by comparing the dwelling unit's true location obtained during the field enumeration process to its mailing address' geocoded location. The binary outcome *Under-Coverage Geocoding Error* indicated whether or not each address incorrectly geocoded out of the segment.

After fitting a logistic regression model with the predictors listed in Table 3, we fit a subsequent model with insignificant predictors removed. We collapsed the area of segment, median home value, and census division variables since not all levels of these variables were significant. The final model was:

Under-Coverage Geocoding Error = Route Type, Rural/Urban, Area of Segment, New Homes, Median Home Value, Census Division

The odds ratios for the predictors in the final model are displayed in Table 4. A number of characteristics were significantly related to under-coverage geocoding error. Addresses associated with high rise postal routes were more likely to incorrectly geocode out of the segment compared to addresses associated with street routes. Addresses in rural segments have over two times the odds of an under-coverage geocoding error compared to addresses in urban segments. The area of the segment also impacts the probability of incorrectly geocoding out of the segment, with larger segments having a higher probability of under-coverage error than smaller segments, although this is probably due to an interaction between the urbanicity of the segment and the size of the segment. Areas with a higher proportion of new homes also have increased odds of under-coverage error, as do areas with median home values of less than $300,000. Under-coverage geocoding error rates also vary by census division, with significantly higher levels of error occurring in the South Atlantic and Mountain divisions than the other census divisions.

Table 4: Under-Coverage Geocoding Error Odds Ratios

Variable	Wtd. Pct.	OR	95% CI
Address-Level Variables			
Route Type			
• High Rise	20.5%	2.54	(1.33, 4.88)
• Street	79.5%	1.00	(1.00, 1.00)
Segment-Level Variables			
Rural/Urban			
• Rural	16.7%	2.49	(1.48, 4.18)
• Urban	83.3%	1.00	(1.00, 1.00)
Area of Segment (SQ Miles)			
• ≤ 1.30	80.5%	1.00	(1.00, 1.00)
• > 1.30	19.5%	2.36	(1.46, 3.81)
New Homes			
• Prop. < National Average	63.0%	1.00	(1.00, 1.00)
• Prop. ≥ National Average	37.0%	1.71	(1.05, 2.80)
Median Home Value			
• < $300,000	80.9%	2.70	(1.38, 5.26)
• ≥ $300,000	19.1%	1.00	(1.00, 1.00)
Census Division			
• South Atlantic and Mountain	24.7%	2.18	(1.28, 3.73)
• All Other	75.3%	1.00	(1.00, 1.00)

3.2.2 Modeling Over-Coverage Geocoding Error

We fit another logistic regression model to determine what characteristics of addresses and segments lead to the highest levels of over-coverage geocoding error. For each of the 1,360 mailing addresses sampled from the mailing list that geocoded into selected segments, we determined whether or not the address correctly geocoded into the segment by matching it against the field enumerated list. Address from the mailing list that were not on the field enumerated list were assumed to have incorrectly geocoded into the segment since the field enumerated listing was assumed to be a complete list of dwelling units in the segment. The binary outcome *Over-Coverage Geocoding Error* indicated whether or not each address incorrectly geocoded into the segment.

After fitting a logistic regression model with the predictors listed in Table 3, we removed insignificant terms from the model. The delivery type and census division variables were collapsed since not all levels of these variables were significantly different. The final model was:

Over-Coverage Geocoding Error = Delivery Type, Rural/Urban, Census Division

The odds ratios for the predictors in the final model are displayed in Table 5. Three characteristics were significantly related to over-coverage geocoding error. Addresses where residents receive mail somewhere other than their curb or a centralized delivery point (e.g. in a mail slot) had significantly lower odds of incorrectly geocoding into a segment than addresses where residents receive mail in a centralized location or at their curb. As with under-coverage geocoding error, over-coverage geocoding error is more prevalent in rural segments compared to urban segments. This is consistent with previous research which showed that ABS over-coverage is more prevalent in rural areas (O'Muircheartaigh et al., 2009). Over-coverage geocoding error rates also vary by census division, with higher levels of error occurring in the New England and Mountain divisions than the other census divisions.

Table 5: Over-Coverage Geocoding Error Odds Ratios

Variable	Wtd. Pct.	OR	95% CI
Address-Level Variables			
Delivery Type			
• Residential Curb, Cluster Box Unit, or Central	70.6%	1.00	(1.00, 1.00)
• Residential Other	29.4%	0.27	(0.13, 0.56)
Segment-Level Variables			
Rural/Urban			
• Rural	11.1%	2.03	(1.16, 3.55)
• Urban	88.9%	1.00	(1.00, 1.00)
Census Division			
• New England and Mountain	24.4%	3.48	(1.62, 7.50)
• All Other	75.6%	1.00	(1.00, 1.00)

4. Conclusions and Future Work

Geocoding accuracy at the segment level is quite poor and varied significantly by urbanicity. Rural segments have a much higher rate of under-coverage geocoding error (23.4 percent) compared to urban segments (7.5 percent). Geocoding accuracy improves significantly at the census block group level for both rural and urban segments, with 99.3 percent of addresses geocoding into the correct census block group (99.8 percent urban, 96.5 percent rural). These findings should be considered when designing ABS studies that are based on census geography. Geocoding error can be a significant source of under-coverage and sampling inefficiencies if segments are smaller than census block groups, especially in rural areas.

There are several characteristics of addresses and segments that are related to segment-level geocoding error. Segment-level geocoding is more accurate in urban areas than rural areas. Geocoding error also varies by census division, postal route and delivery type, the area of the segment, the proportion of new homes, and median home values within the segment.

Our future work will examine various geocoding methods and determine which geocoding methods are the most accurate. We will also incorporate segmentation analysis in our regression models to control for significant interactions between predictor variables.

5. Acknowledgements

This project is funded by the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract no. 283-2004-00022 and Project no. 0209009. The views expressed in this paper do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

The authors would like to acknowledge RTI staff members Katie Morton, Amanda Lewis-Evans, and Jamie Cajka for their contributions to this research. We would also like to acknowledge the contributions of David Malarek of Marketing Systems Group and Art Hughes and Michael Jones of SAMHSA.

6. References

Iannacchione, V., K. Morton, J. McMichael, B. Shook-Sa, J. Ridenhour, S. Stolzenberg, D. Bergeron, J. Chromy, and A. Hughes. 2010. The best of both worlds: a sampling frame based on address-based sampling and field enumeration. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association.

McMichael, J.P., J.L. Ridenhour, B.E. Shook-Sa, K.B. Morton, and V.G. Iannacchione. 2010. Predicting the coverage of address-based sampling frames prior to sample selection. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Morton, K. B., J.P. McMichael, J.C. Cajka, R.J. Curry, V.G. Iannacchione, and D.B. Cunningham. 2007. Linking mailing addresses to a household sampling frame based on census geography. In *JSM Proceedings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association. 3971-3974.

O'Muircheartaigh, C., N. English, M. Lattermer, S. Eckman, and K. Dekker. 2009. Modeling the need for traditional vs. commercially-available address listings for in-person surveys: results from a national validation of addresses. In *JSM Proceedings*, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Research Triangle Institute (2008). "SUDAAN Language Manual, Release 10.0." Research Triangle Park, NC: Research Triangle Institute.