Linking Mailing Addresses to a Household Sampling Frame Based on Census Geography

Katherine Morton, Vincent Iannacchione, Joseph McMichael,
James Cajka, Ross Curry, and David Cunningham

RTI International

Abstract

Using Census geography as the primary sampling unit in area-based household surveys ensures complete geographic coverage of the target population. In addition, the use of Census geography facilitates matching to external demographic data. For these reasons, sampling frames for household surveys are often based on Census geography at the first stage of sample selection. Subsequently, counting and listing methods can be used to create household frames within sampled areas for the final stage of sample selection. However, these methods are often cost and time prohibitive. RTI International has conducted research to investigate potential advantages of using residential mailing lists instead of the traditional onsite enumeration methods in an area-based survey. An inherent challenge in this process is that of linking mailing addresses to a corresponding frame. For this research, Geographic Information System (GIS) and Global Positioning System (GPS) technologies have been used to identify addresses located within the sampled block. Other challenges addressed in this paper are those related to inclusion of areas known to have poor mailing list coverage such as rural areas, areas without home delivery of mail, and areas with simplified mailing addresses. Secondary sources are often needed to supplement the residential mailing lists in these areas.

KEY WORDS: area sampling frame, residential mailing lists, counting and listing, GIS, GPS, geocoding

1. Introduction

Traditional onsite enumeration (or counting and listing) can be used to create household frames within sampled areas for the final stage of sample selection in area-based household surveys. However, these methods are often time and cost prohibitive. Because mailing address lists are inexpensive and can be purchased closer to the start of data collection, recent research has focused on residential mailing lists as an alternative to counting and listing.

Mailing address lists can be purchased by postal geography (postal carrier routes). However, postal carrier routes do not cover complete geographic areas and are subject to change periodically. As shown in Table 1, a sampling frame based on Census geography is preferred to a frame based on postal geography in area-based household surveys. Using Census geography ensures complete geographic coverage of the target population and allows matching to external data (for example, rural or urban identifiers, housing unit counts, and population counts are available at the Census block level). However, linking mailing addresses to a frame based on Census geography is an inherent challenge.

Table 1. Census Geography versus Postal Geography

	Pros	Cons
Census Geography	• Complete geographic coverage of target population • Facilitates matching to external data	• Geocoding error
Postal Geography	• Complete list of mailing addresses within sampled area	• Incomplete geographic coverage • Boundaries are dynamic

Areas known to have poor mailing list coverage present another challenge. As reported by Staab and Iannacchione (2003), mailing lists have poor coverage in rural areas. Some rural areas may not have home delivery of mail or may use simplified addresses (name, city, and state only). Also, Dohrmann et al. (2006) reported undercoverage of group quarters in residential mailing lists. Despite these drawbacks, RTI was able to devise a method for linking mailing addresses to Census geography. This paper evaluates this method and suggests sources for improving the coverage of mailing lists.

2. Research Study

RTI International conducted a research study which compared residential mailing lists to traditional onsite enumeration methods. For this study, 50 area segments consisting of one or more Census blocks were selected across the state of North Carolina with probability proportional to size. A Geographic Information System (GIS) was used to overlay the postal carrier routes with the Census geography and mailing lists were purchased for all routes within the sampled areas. To ensure that all

mailing addresses were obtained, mailing lists were also purchased for the neighboring Census blocks.

Field staff members were sent to each sample segment to enumerate households using traditional count and list methods. Then, field staff revisited each segment and located addresses from the residential mailing address list of the sampled area. To assist in the matching post-process, a Global Positioning System (GPS) device was used to record the coordinates of each household during each visit.

3. Evaluation of Geocoding

A secondary objective of the research study was to evaluate geocoding as a method for linking mailing addresses to Census geography. Geocoding is the process by which geographic coordinates are assigned to mailing addresses by interpolating the address location along a section of street based on address ranges assigned to that section. Using geocoding, a mailing list vendor can determine the Census geography (e.g., blocks, block groups, and tracts) for any given address. The vendor can then provide all mailing addresses within the sampled Census geography.

In order to evaluate the accuracy of geocoding, geocoded coordinates were obtained for all mailing addresses identified within the segment boundaries. The sample represented 165 blocks, 62 block groups, 50 Census tracts, and 36 counties in the state of North Carolina. Table 2 shows the cumulative percent of mailing addresses that geocoded to the correct block, block group, Census tract, and county. One rural mailing address did not geocode to the correct county; therefore, the cumulative total in rural areas is 99.9 percent. Households were more likely to geocode to the correct block in urban areas than in rural areas with 73.4 percent of the urban and 37.6 percent of the rural households geocoding to the correct block. In both areas, the majority of the households geocoded to the correct block group and Census tract.

Table 2. Cumulative Level of Geocoding Accuracy by Rural/Urban (Percentages)

Level of Accuracy	Rural (n=1,200)	Urban (n=4,318)	Total (n=5,518)
Block (165)	37.6	73.4	65.6
Block Group (62)	89.9	91.7	91.3
Census Tract (50)	91.7	92.6	92.4
County (36)	99.9	100.0	100.0

Note: *n* is the number of mailing addresses found in the sampled areas.

For the next portion of the analysis, GPS and geocoded coordinates for mailing addresses located within the segment boundaries were compared. A comparison of all addresses that were found inside the sample segments but geocoded outside of the segments showed that most addresses geocoded within 100 meters of the actual location. The distances ranged from about 10 meters to 10,000 meters (approximately 6 miles).

Figure 1 shows GPS and geocoded coordinates for one sampled area. The GPS coordinates are denoted by yellow dots. Geocoded coordinates for in-sample mailing addresses which geocoded outside of the segment boundaries are denoted by purple dots. Blue lines are drawn to show distances between GPS and geocoded coordinates for these addresses. For example, 2617 Billhooks Road geocoded 857 meters northwest of the actual location which would have placed it outside of the segment boundary. The majority of the geocoding errors occured on or near the boundary of the sample segment. Finally, Census block groups are denoted by different shades of green. With the exception of the addresses in the northwest corner of the map, addresses geocoded into the correct block group.

4. Improving the Coverage of Mailing Lists Purchased by Census Geography

In order to maximize coverage of the target population, it is important to supplement mailing lists that are subject to geocoding error. The half-open interval (HOI) frame linking procedure is commonly used in area-based sampling to pick up new or missed households (Kish, 1965). The procedure requires that a geographic order be established for the households. Then any new or missed household in the interval between the sampled household and the next household on the list is considered part of the sample. By ordering mailing addresses by postal delivery sequence, the HOI procedure can also be used to supplement mailing list frames.

In areas with no locatable addresses, a secondary source may be needed to supplement the frame. For example, households without city-style addresses may be listed in telephone books, driver's license files, or credit card files. If time and budget allows, areas with poor coverage of mailing addresses could also be listed using traditional in-field enumeration. Finally, if group quarters are part of the target population, a secondary source could be used to improve the coverage of these units. For example, an education survey frame could be used to enumerate college dormitories.

5. Conclusions

The use of Census geography in area-based household surveys ensures complete geographic coverage of the target population. Our research indicated that mailing lists

based on Census geography are subject to some amount of geocoding error. Geocoding is more accurate for larger geographic areas. Therefore, we recommend that block groups or Census tracts serve as the primary sampling units when sampling from a residential mailing list. Sampling households across larger geographic areas will also decrease the variance of estimates since clusters are less homogeneous (i.e. there is less intra-cluster correlation).

We expect that the accuracy of geocoding will improve over time. We also expect that the coverage of mailing lists will improve as more rural areas are being converted to city-style addresses for household emergencies (E911 systems). In the mean time, supplemental sources can be used to improve the coverage of mailing lists.

Acknowledgements

This research was funded by RTI International as part of an internal research and development project. The authors thank RTI staff members James Chromy, Jamie Ridenhour, and Amanda Lewis-Evans for their contributions to this research.

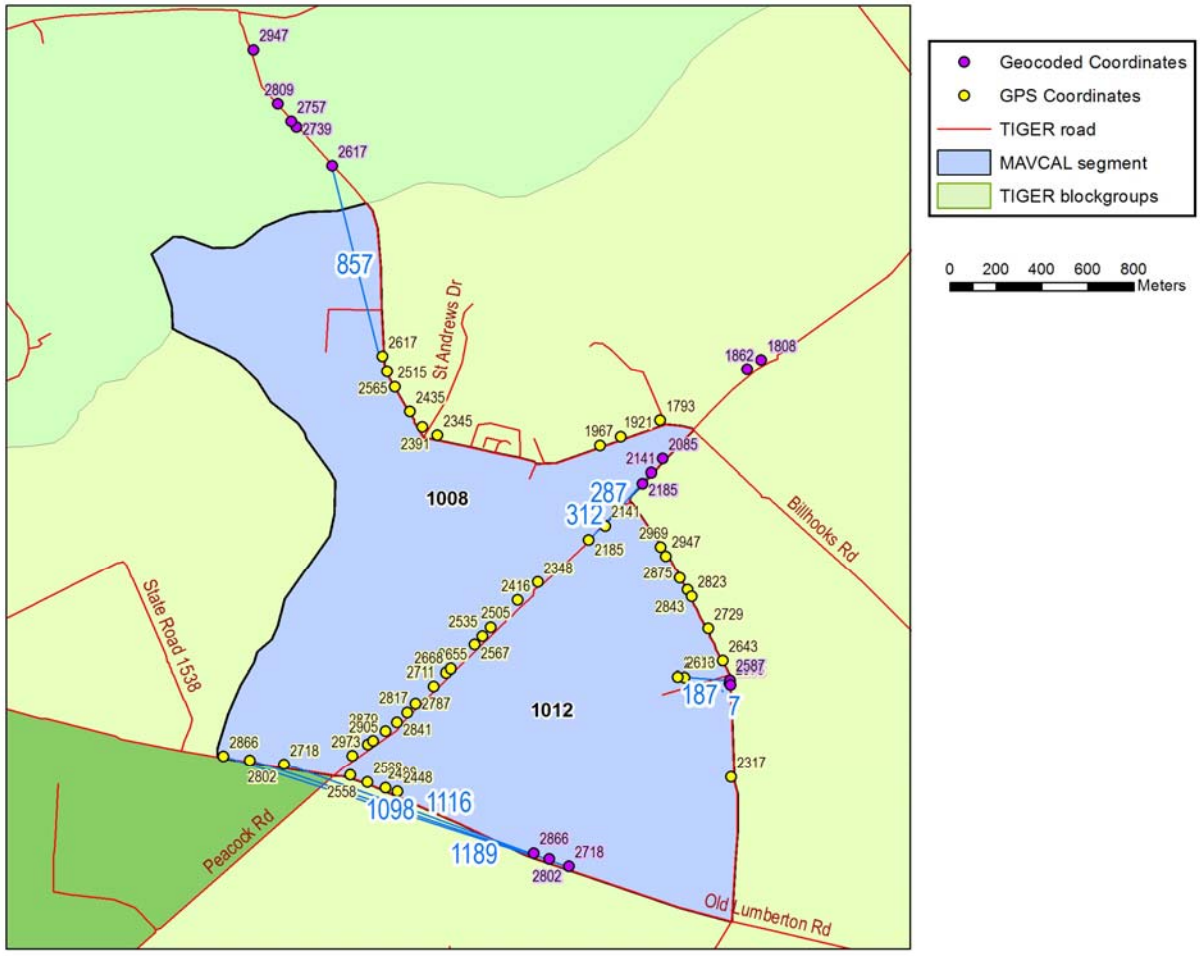
References

Dohrmann, Sylvia, Daifeng Han, and Leyla Mohadjer (2006). Residential address lists vs. traditional listing: Enumerating households and group quarters. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 2959-2964.

Kish, Leslie (1965). *Survey Sampling.* New York: John Wiley & Sons.

Staab, Jennifer and Vincent Iannacchione (2003). Evaluating the Use of Residential Mailing Addresses in a National Household Survey. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pp. 4028-4033.

Figure 1. GPS and Geocoded Coordinates in One Area Segment



Predicted segment coverage = # Locatable Addresses #Households + #Group Quarters Estimating ABS coverage by this formula brings in several potential sources of error. The numerator, number of locatable mailing addresses, is a count of addresses within a segment from the ABS mailing address list. The addresses on the list for that segment are subject to geocoding error.Â 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. Morton, K.B., J.P. McMichael, J.L. Ridenhour, and J. Bose. Census can increase nonsampling errors. In large populations to maintain accuracy a sample is in favor of census. Sample design process.Â Example- a household. sampling frame. Is a representation of the elements of the target population. Examples - telephone book, association directory listing the firms in an industry, mailing list purchased from a commercial organization, city directory, or a map. Sample size. Number of elements to be included in a study.Â Then random sample of clusters is selected based on probability sampling techniques such as simple random sampling. For each selected cluster either all the elements are included in the sample or sample of elements is drawn probabilistically. Sampled households can follow the simple instructions given in the notification letters to activate and complete the online questionnaire on their own. Questionnaire content. The GHS questionnaire is designed to collect information on the labour force, employment, unemployment and underemployment, as well as the demographic and socio-economic characteristics of the population.Â Households who wish to check the identity of the census officer visiting / calling them can contact C&SD by the following means: Address. 21/F, Wanchai Tower, 12 Harbour Road, Wan Chai, Hong Kong.Â E-mail address. household-survey@censtatd.gov.hk. Statistical data.