TRADITIONAL AND ENHANCED FIELD LISTING FOR PROBABILITY SAMPLING

Colm O'Muircheartaigh, Stephanie Eckman, Charlene Weiss
NORC, University of Chicago, 55 E Monroe, Chicago, IL 60603

Keywords: Listing, Geocoding, United States Postal Service, Sampling Frame

Introduction

In area probability samples, secondary sampling units (SSUs), or segments, are often made up of census blocks, aggregated to reach a minimum housing unit (HU) count. The next step is to create or obtain a list of all housing units (HUs) in each segment. The third stage of sampling is then the selection of HUs from this list.

NORC, in common with most major field survey organizations, carries out its own listing for area samples. Trained listers are sent out to each segment to do the listing, following a strict protocol. The listers walk around each selected block, starting at the northwest corner and continuing around the block clockwise, writing down the address or description of every HU they find. These lists are of good quality, but are expensive to produce. Costs include training, travel, room and board, and listing time. In this paper we will refer to this listing process as *traditional listing*.

An Alternative to Traditional Listing

The United States Postal Service (USPS) maintains a list of residential addresses that the mail carriers update continually. The list is ordered by zip code, carrier route and walk sequence number. This sequence of codes uniquely identifies every delivery point in the country. Direct mail marketers license the list from the USPS and resell it. The list they sell is not as complete as the one used by the USPS, because households can request that their addresses not be sold.

The availability of this list to survey organizations raises the question: could it serve as a replacement for or as a supplement to our traditional listing efforts? Iannacchione, Staab and Redden (2001) discuss the use of this address list as a sampling frame in Dallas County, Texas. They selected a sample of 2,724 addresses and were able to locate 97% of them in the field.

Theoretically, the USPS list contains every delivery point in the country. In many parts of the country, delivery points are household addressees. In these areas, the USPS list most closely approximates the HU list we need for our third stage of sampling. In the more rural parts of the country, the list contains many rural route and post office box numbers, which are not useful for our sampling frame. In these areas, the USPS list cannot assist in developing an HU list.

Households that receive mail only at rural route or post office boxes are not represented on the list by housing unit address. Their absence is one source of incompleteness of the USPS lists. Opting out by households who wish not to receive mass marketing direct mail is another source of undercoverage. We felt, however, that the list was promising enough to serve as the basis for local area sampling; this paper reports on an investigation of the quality of the USPS list for this purpose.

Evaluating the USPS List

In late 2001, NORC conducted tests of the usability of the USPS address list as a sampling frame. At that time, as a supplement to our existing national master sample, we listed 79 segments for the General Social Survey, a national survey conducted in alternate years by NORC on behalf of the National Science Foundation and others. These segments were listed using the traditional protocol. We also purchased the USPS list for the zip codes that covered each segment. The list for 27 of the segments did not have any street-style addresses. In 14 of the remaining 52 segments we carried out an additional field test — we sent *different* listers into the field to list the segments again. In this second listing effort, which we call *enhanced listing*, we provided each lister with the geocoded USPS addresses for their blocks and asked them to make corrections to this list. Thus, instead of starting from blank listing sheets, these listers already had an HU list that they updated in the field. We tried to make the process of updating the list as similar to traditional listing as possible, in order to isolate the contribution of the USPS list to list quality.

In the 52 segments we were able to compare the traditional list with the USPS list. In 14 segments a three-way comparison was possible – the *traditional* list, the *USPS* list and the *enhanced* list. This paper focuses on the three-way comparison in these fourteen segments.

Implementation of Enhanced Listing

After purchasing the USPS list for the zip code(s) that contained each of the 14 test segments, the next step was to identify, by geocoding, which of these addresses were within the segments. Geocoding software takes in a street address and assigns a latitude and longitude to the address, referring to its city, zip and street level databases. The software can match the address at many levels: using just the zip code, it can match the address to that zip's centroid, or to its zip+4 centroid. If the address is complete and in standard format (e.g. 254 S State St, Chicago IL 60602) it can match the address at the street level. The street database contains the name, location and address ranges of all streets—the program then interpolates the position of the given address. For example: 254 S State St would be assigned a latitude and longitude that would place it half way down the 200 block of south State Street. More than 99 percent of the addresses (excluding post-office and rural route boxes) geocoded at the street level. Once we knew which census blocks the addresses were in, we could classify each as inside or outside our segments. See Table 1.

Table 1: USPS list numbers

Test Segments	14
Zip codes that cover segments	18
Addresses in these zips	232,810
Post-office boxes	3723
Rural route boxes	12,648
City-style addresses	216,439
Addresses geocoded inside segments	2336

To follow traditional listing as closely as possible, we then sequenced these addresses within each block, starting at the NW corner and continuing around the block in a clockwise direction. NORC uses this method in all of its traditional listing and we did not break with this convention in our experiment.[1]

Next we prepared maps and listing sheets for listers to use in the field. We printed a segment overview map for each segment and a block map for each block within each segment. Stars on these maps showed the listers where we expected the addresses to be. The listing sheets were preprinted with the ordered addresses from the list. We left four blank lines between every two addresses on these listing sheets, as a reminder to the listers that they should look carefully for missing HUs.

Comparison of Three Listing Approaches

Issues in Enhanced Listing

Our test uncovered several issues with the USPS list and the geocoding process. As we discussed above, the USPS list does not provide HU addresses for households that receive all their mail at a post office box or a rural route box. These post-office and rural route box addresses are not useful for NORC's purposes. In some zip codes, the USPS list is made up entirely of these types of addresses; in these areas enhanced listing as we have tested it is not possible.[2] Households may also request that vendors not include their addresses when they sell the list.

The USPS list is very good at finding hidden HUs: basement apartments with an entrance in the rear of what looks like a single-family home, small apartments behind store fronts. Listers often miss these types of HUs. In enhanced listing, the listers need only confirm the existence of the hidden HUs that are already on the listing sheets. Enhanced listing (E) thus captures more of these hidden HUs than traditional listing (T).

Some problems experienced during enhanced listing arose from errors made by the geocoding software. The software uses rules to assign a latitude and longitude to each address. For instance, in Chicago it might use the rule that all odd addresses are on the east side of the street and the evens on the west side and that house numbers run the full range of the street's allotted numbers (e.g. from 200 to 299 on south State Street). These rules would work for many blocks in Chicago but not for all. For instance, some streets have house numbers only up to the middle of their assigned range (from 200 to 254). The geocoding software would place 254 S State Street in the middle of the block, though it may actually be at the south end. Another common error in some parts of the county was to assign the even numbered houses to the side that actually had the odds, and vice versa.

[1] This sequencing of the addresses within each block was the most time-consuming part of the experiment for us. In the future we expect to change the enhanced listing process so that this task is no longer necessary.

[2] We are currently investigating using different vendors to purchase addresses for these rural areas; other vendors appear to have more city-style addresses. We are not ready to make any statements about the use of these other lists in enhanced listing.

Both these geocoding errors had repercussions for enhanced listing. We used the geocoding software to determine which HUs were inside the segments. An error in placing the address on a map could mean that the HU was assigned to the wrong block and thus mistakenly included in or excluded from) our U. More accurate geocoding would lead to a more complete and accurate U which needed fewer corrections in the field.

Our enhanced listing project was particularly sensitive to these geocoding errors. We tried to use geocoding to replicate the methods of traditional listing. In future enhanced listing projects, using the postal service's own ordering (zip code-carrier route-walk sequence) may help.

Cost Comparison

Listing costs are largely fixed: anytime we send a lister into the field we must pay for training, travel, and room and board. We cannot reduce these expenses; however, we can reduce the time it takes to list a segment. Enhanced listing cuts costs by decreasing the time spent in the field.

We have found that this marginal cost of traditional listing is twice as expensive as enhanced listing. Enhanced listing takes much less time than traditional listing does. The lists are correct enough that in most cases the lister is just checking off the addresses that are already on the listing sheets.

An additional cost associated with enhanced listing is the price of the USPS address list. These lists cost $100 per 10,000 addresses. In this test we spent just over $2000 on the USPS addresses list for our fourteen segments. This expense is much less than the amount saved in enhanced listing.

Results of Enhanced Listing in One Segment

Once the updated listing sheets were returned, we could begin to compare the three HU lists we had for these fourteen segments. The three lists were:

T – Traditional listing HU list
U – USPS address list, geocoded inside the segments
E – U enhanced in the field

We geocoded all three lists, which allowed us to construct maps to compare them. Figures 1 and 2 show the results of the different lists for one of the segments. Figure 1 shows the overlap between E and U and Figure 2 the overlap between E and T.

In Figure 1, black stars represent addresses that are on both the enhanced list and the USPS list; red plus signs indicate addresses added to the USPS list by the enhanced lister; grey circles mark addresses that were on the listing sheets but that the enhanced lister could not find in the field and dropped from the list. The most striking feature of the map is the degree of agreement between E and U; most addresses in the segment were captured by both methods. The enhanced lister did delete three HUs from the list, and also added four—including three that appear to be just off the block in the northeast corner.

These three addresses illustrate the geocoding problems discussed above. They do lie inside the block; the software mistakenly placed them outside. Though these addresses were on the USPS list we purchased, they geocoded outside the segment and thus were not included on U and were not on the listing sheets for our test. The USPS list was correct, but due to geocoding errors we did not think these addresses were inside our segment.

Figure 2 is a similar map showing the agreement between E and T. Again black stars indicate matches; red pluses represent addresses that are unique to the E list, and grey circles, addresses that are unique to T. There were no addresses listed by the T lister that E did not also list, though there were several listed by E that T missed.

These maps suggest that enhanced listing shows promise. In order to determine whether E can give us a better quality list than T or U, we needed to compare the three lists across all of our test segments.

Comparison of T, U and E in All Test Segments

To compare the coverage of E, T and U, we first needed to match addresses across the lists. Although "254 S State St", "254 South State Str." And "254 State So" clearly all represent the same address, it was not easy to match addresses from different sources. We used address parsing and geocoding to decide on the equivalence of addresses in different formats.

Table 2: Intersection percentages across all segments

U in E	95%
E in U	93%
T in U	87%
U in T	77%
T in E	92%
E in T	81%
E in USPS	96%

Comparing E and U gives us a sense of how accurate the USPS list would be without field enhancement. Ninety-five percent of the addresses in U are also in E; 93 percent of those in E are in U. See Table 2. Together these numbers mean that while E contains more addresses overall, the overlap between the two lists is high.

Comparing U and T shows whether simply purchasing the USPS list can give us the same list we can get from traditional listing. Eighty-seven percent of the HUs listed in T are also in U; 77 percent of U is in T. These figures point to difficulty in matching T's descriptions with U's addresses. T listers are told to use descriptions whenever an address is not visible: "2 story brick between 11 and 15 Main St". In the cases where the home does indeed have an address, that address is on the U list ("13 Main St"); but back in the central office, it is not easy to match the two as referring to the same HU. Therefore, the actual match rate between T and U is higher than it appears in Table 2.

Another explanation for the low T in U percent is that we found errors in the way T was done, where listers walked past their assigned blocks and listed HUs outside the segment. Three difficult segments have a low U/T match rate and thus pull down this overall rate. Enhanced listing is less prone to these sorts of errors because we have upgraded our maps and materials to do the geocoding and address sequencing. The U in T percent is also contaminated by our inability to perfectly match T description to U addresses. However, U clearly captures more HUs that T does.

Comparing E and T shows the improvement of enhanced listing over traditional. Ninety-two percent of addresses in T are in E; 81 percent of those in E are in T. These numbers tell us that E contains nearly all the addresses that T does, but also picks up many more and at less cost. (As above, the actual T in E percent is probably higher; if we could match more of the addresses and descriptions, we would see this.)

We also looked at how many of the addresses missed by U were actually in the USPS list, but did not appear on our listing sheets due to geocoding error. We matched all addresses added to U by E (the red pluses in Figure 1) to the larger USPS list. Recall that one of the first steps in implementing enhanced listing was to use geocoding to pare the large USPS list down to just those addresses that were inside the test segments. Any errors in geocoding could lead to mistakes in deciding which addresses were inside and which were outside. We found that the percent of E addresses that are in the larger USPS list is 96 percent. Looking at both the E in U and E in USPS percents, three percent of the addresses in E were added because of geocoding errors on the original USPS list—we mistakenly thought these addresses were not in the segment. This calculation tells us that if we can reduce the geocoding errors or lessen their impact, the USPS list is an even better approximation of the enhanced list.

Conclusion

In rural segments, the USPS list does not provide us with usable addresses, because the list consists of post-office or rural route boxes, rather than household addresses. In these segments traditional listing is the only method available, though we have other research projects to explore ways to improve it.[3]

In the non-rural segments (which are the majority) enhanced listing is clearly superior to traditional listing. This research has shown that T is inefficient: it produces an inferior HU list at greater cost. Enhanced listing produces the most complete list of HUs in a segment, and corrects for errors introduced during geocoding. However, the purchased addresses alone are also a good quality HU list. Sampling directly from U without enhancing and using a half-open interval might pick up the HUs that U misses, particularly those dropped from the list at the request of the household.

This small test does not allow us to draw any strong conclusions about the quality of enhanced listing, though the results for our 14 segment are encouraging. Since completing this test of enhanced listing, we have undertaken a new study to compare the performance of a sample from an enhanced list with a sample from an unenhanced list. This next study focuses on deprived urban neighborhoods, which were not represented in our initial listing experiment. We hope to report these results in 2003.

Reference

Iannacchione, Stabb and Redden. 2001. "Evaluating the Use of Residential Mailing Addresses in the Metropolitan Household Survey." Paper presented at AAPOR, May 2001.

[3] As another part of this research project, we sent our listers out with handheld GPS devices in order to collect latitudes and longitudes of all listed HUs in several test segments; we do not have the space to discuss this research here. Future traditional listing efforts will also benefit from technological improvements made to do enhanced listing.

Figure 1: Comparison of E, U lists in One Segment

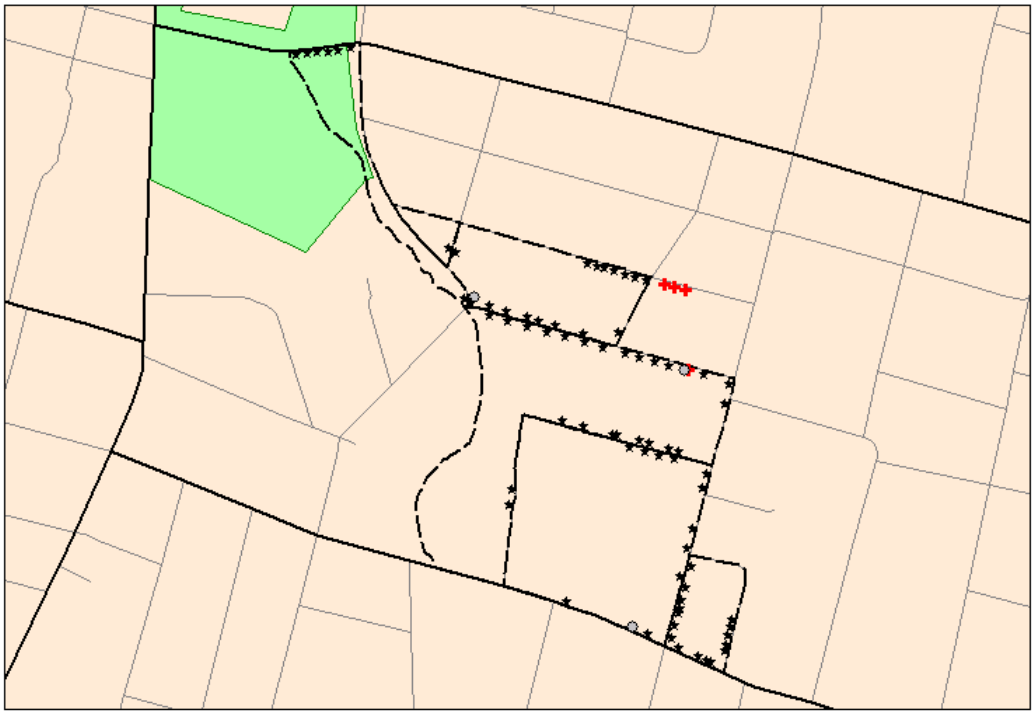

Figure 2: Comparison of E, T lists in One Segment