Experiences with the Use of Addressed Based Sampling in In-Person National Household Surveys

Jennifer Kali, Richard Sigman, Weijia Ren, Michael Jones
Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract

When selecting multistage samples for in-person household surveys, the final stage of sampling typically involves sampling dwelling units or addresses from lists of these units within sampled geographic areas generally known as segments. The use of address-based sampling (ABS) frames based on USPS-lists as the source of address lists is a cost-effective alternative to the traditional listing of dwelling units by field staff. This paper discusses the results of an application of the use of an ABS frame for a recently completed national in-person household survey. An Address-Coverage Enhancement (ACE) procedure, which involves the sampling of geography-based units in which field staff record potential off-of-frame addresses and the sampling of confirmed off-of-frame addresses for assignment to data collection, was used to address coverage issues with the ABS frame. The usefulness of the vacancy indicator and the educational institution indicator which are available on the ABS frame will be discussed. Methods used to sample clustered units (called drop points) which are included on the frame will also be evaluated.

Key Words: Address-Based Sampling, Area Probability, Multi-stage, in-person survey, coverage, frame

1. Introduction

A typical design for a national in-person household survey is a multi-stage design in which at the last stage housing units are sampled within sampled geographic areas called segments. Traditionally, listings of housing units within the geographic boundaries of the segment have been compiled for use as the sampling frame for the housing unit sample. Listing or field enumeration is expensive. Recently there has been a move to replace field enumeration with a much less expensive address-based sampling (ABS) frame which replaces the housing unit lists with address lists provided by the United States Post Office (USPS). Iannacchione (2011), Kalton, Kali, and Sigman (2014), and Dohrmann, Montiquila, Buskirk, and Hyon (2014) provide thorough background on the ABS frame. This paper describes the use of an ABS frame for sampling addresses for the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC-III).

2. Summary of the Sample Design for NESARC-III

The NESARC-III is a large national in-person survey conducted in 2012-13 of adults living in the United States in over 65,000 sampled households. The NESARC-III household sample was generated from a four-stage sample design. At the first stage of sampling, a stratified probability-proportional-to-size (PPS) sample of 150 PSUs, consisting of individual counties or combinations of contiguous counties, was selected from the 50 states and the District of Columbia. PSUs were sampled with probability proportional to the number of housing units according to the 2010 Population Census.

High and medium minority PSUs were oversampled. At the second stage, a stratified PPS sample of around 48 area-segments was selected in each of the selected PSUs, where the segments were defined in terms of blocks or combinations of adjacent blocks with a minimum of 60 households per segment based on the 2010 Population Census. Segments were sampled with probability proportional to the number of housing units according to the 2010 Population Census. High and medium minority segments were oversampled. At the third stage, a systematic sample of addresses was selected from each sample segment. In all but three highly rural counties, the main address sample was selected from the USPS addresses that geocoded to the segment based on a list of addresses provided by a vendor Marketing Systems Group (MSG) from the USPS's Computerized Delivery Sequence (CDS) file. Kalton et al. (2014) provide a description of the CDS file. In the three rural counties, the main address sample was selected from a list resulting from field enumeration. In a sample of the other sampled counties, the Address Coverage Enhancement (ACE) procedure was used to supplement the main address sample with a sample of addresses that were either not on the USPS lists or not locatable from those lists. At the fourth-stage of selection, one or two adults were sampled from each sample household.

3. Main and Supplemental Samples

3.1 Main Sample Selection

A typical design for an in-person household survey is a multi-stage design in which housing units are sampled within sampled geographic areas, called *area segments*. Kalton et al. (2014) defines another type of segment – the *list segment* **-** that can be useful when utilizing an ABS sampling frame to select the household sample. The list segment is the set of addresses in the vendor's ABS database that geocode into the area segment. See Dohrmann, Kalton, Montaquila, Good, and Berlin (2012) and Eckman and English (2012) for more details on geocoding.

Because of geocoding errors, some of the addresses in a list segment may be for housing units that are not physically located in the area segment. Conversely, because of geocoding errors, one or more of the addresses in the vendor's data base for housing units that are physically located in an area segment may not be included in the list segment. According to the *list segment eligibility rule* defined in Kalton et al. (2014) and used in NESARC-III, any address that geocodes into a sampled area segment is eligible to be sampled regardless of the physical location of the housing unit. The list segment eligibility rule increases the coverage of the ABS frame over an *area segment eligibility rule*, which excludes housing units that geocode to the sampled area segment but are physically located outside the area segment.

For NESARC-III, area segments for NESARC-III were formed by grouping Census blocks such that there were a minimum number of occupied housing units based on counts from the 2010 Decennial Census. The minimum number of housing units per area segment was 60, although on average the number of housing units per area segment was 100. At the time that segments were being formed for NESARC-III, MSG had not yet geocoded addresses for the entire nation to the 2010 Census blocks so that the area segments could not be formed to take account of the ABS counts of addresses that geocoded to the segments.

Table 3.1.2 provides details on the differences between the ABS counts of addresses in a list segment and the Census counts of occupied housing units in the corresponding area segment. For almost half of the segments, the two counts are within 10 percent of each other. There are some large differences between the two counts, although in only 2.1 percent of segment is the ABS count of addresses more than twice the Census count of housing units. In 2.1 percent of segments, although there were at least 60 housing units in the area segment according to the Census count, there were no addresses on the CDS file in the associated list segment and thus no household sample was selected for these list segments.

Table 3.1.2. Segment Differences Between ABS count of Addresses and the Census Count of Occupied Housing Units

	% of segments
ABS count within 5% of Census count	28.0
ABS count within 10% of Census count	47.6
ABS count within 20% of Census count	68.4
ABS count within 50% of Census count	88.2
ABS count within 100% of Census count	97.9
ABS count more than twice Census count	2.1
0 addresses on ABS frame	2.1

The differences between the number of households according to the Census and the number of addresses on the ABS frame led to a highly variable within-segment sample size for the NESARC-III main sample. Subsequent studies have formed segments based on the count of addresses on the ABS frame and used the count of addresses on the ABS frame as the MOS for PPS selection. Since the sample of household addresses is also selected from the list of addresses on the ABS frame, the using the same measure in both stages of selection produces less variation in within-segment sample sizes. In rural areas that are not well-represented on the CDS file, however, formation of segments based on the count of addresses on the ABS frame leads to geographically large segments that are difficult to field. A hybrid approach in which the Census count is used to form segments in some rural areas may be beneficial.

In some rural areas, the proportion of household addresses that are on the ABS frame is very low. In three counties in PSUs sampled for the NESARC-III, segments were listed via field enumeration because the number of addresses on the ABS frame was so small. The decision to enumerate segments was made at the county level to avoid issues with geocoding error. For example, suppose that segment A is listed via field enumeration and the neighboring segment, segment B, the list of addresses is taken from the ABS frame. Due to geocoding errors, addresses physically located in segment A may geocode into segment B. Under the list segment eligibility rule, the housing units that are physically in segment A but geocode into segment B would be eligible for the survey if segment B is selected. However, because segment A is a listed segment, households that are physically located in segment A would be eligible for the study if segment A is sampled. Therefore, households that are physically located in segment A but geocode to segment B would have two chances of selection. To avoid this scenario, all sampled segments in entire counties were field listed.

2.1 Supplemental Sample Selection

Using the list segment eligibility rule means that every address located on the USPS list is eligible for sampling. However, there are households that, for various reasons, are not on the ABS frame. The Address Coverage Enhancement (ACE) procedure described in Kalton, et al. (2014) was used in the NESARC-III to address under-coverage due to addresses missing from the ABS frame. For the ACE procedure, addresses within area segments that are not on the ABS frame are eligible for supplemental sampling. Because the supplemental sample is selected from only those addresses not contained on the ABS frame, the ACE sample is mutually exclusive of the main sample. Thus, the area segment is associated with two types of segments: the list segment and the ACE segment.

2.1.1 Synopsis of the ACE Procedure

The ACE procedure consists of the following six tasks:

1. Select ACE segments. The ACE procedure is performed in a random subsample of the sampled segments (apart from those that were field listed) Segments are selected for the ACE procedure with probability $P(i) = k_i r_i$ where k_i is the under/over sampling factor for ACE segment i, and r_i is the within-segment sampling rate for sampling addresses in segment i. For the NESARC-III, about ten percent of sampled segments were sampled for the ACE procedure. Segments for NESARC-III were sampled by PPS sampling with the measure of size based on the difference between the area segment count and the list segment count and the county-level urbanicity of the segment according to the Beale Code (collapsed to two levels – urban and rural). Subsequent studies using the same design based the definition of urbanicity on the Census 2010 Type of Enumeration Area (TEA) Delineation which is available at the block level and also includes an indicator of the quality of mail coverage.

2. Obtain the vendor's ABS database addresses for the selected ACE segments. Specifically, obtain all the addresses from the vendor's ABS database that geocode into each ACE-selected area segment. These addresses are the list segment sampling frame for the ACE segment.

3. Perform the ACE field procedure. The addresses from the vendor's ABS database for the ACE segments are loaded into a laptop computer. Field staff, called *listers*, canvas each ACE-selected area segment in a systematic manner. They determine for each housing unit they encounter within the boundaries of the area segment whether the address is on the list segment sampling frame that is preloaded into their laptop computer. If so, they assign the address a status of "located." If not, they record the address (and the laptop application flags the address as "added in the field"). Note that this procedure is performed by a highly trained team performing this task alone, separately from the task of completing interviews.

4. Match added addresses to the ABS database. The reconciled addresses added in the field are checked against the vendor's ABS database to determine if they are truly missing from the database, or if they are present in the ABS database but geocoded into another list segment.

5. Sample the non-matching added addresses and assign them for data collection. Non-matching added address j in ACE-selected segment i is selected for data collection with probability $P(j|i)=1/(w_i k_i)$ where w_i is the ratio of the sampling weight of a sampled added address in segment i to the sampling weight for an addresses sampled from the vendor's ABS database in segment i.

6. Confirm Address for Added Addresses Addresses are confirmed or corrected for all sampled non-matching added addresses during the screener. Corrected addresses are sent to the vendor to match to ABS frame. Sample base weights for sampled non-matching added addresses found on the ABS frame are adjusted for duplicate chances of selection.

Figure 3.2.1 illustrates the steps of the ACE procedure.

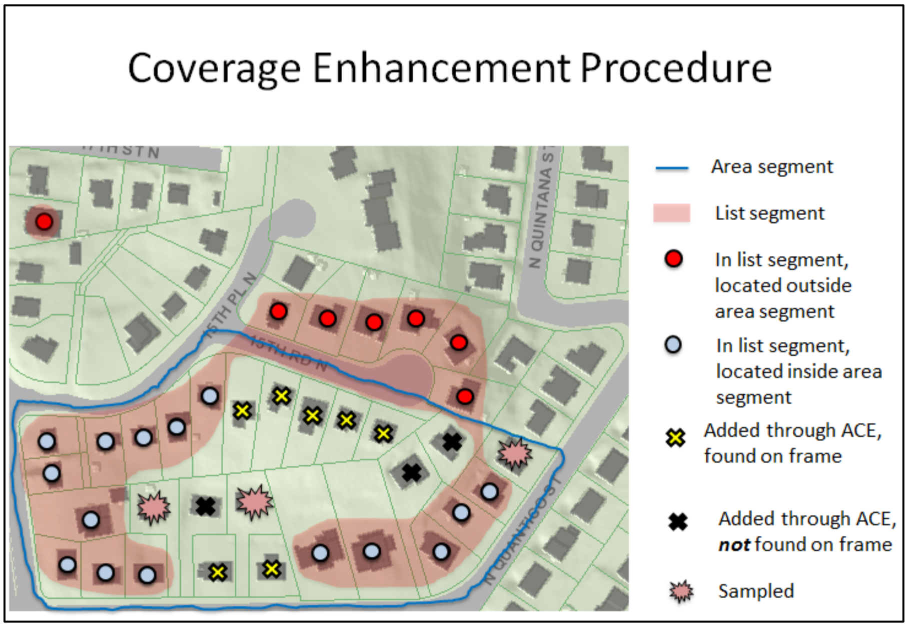
Coverage Enhancement Procedure
11TH ST N
N QUINTANA ST
15TH PL N
15TH RD N
N QUANTICO ST
Area segment
List segment
In list segment, located outside area segment
In list segment, located inside area segment
Added through ACE, found on frame
Added through ACE, *not* found on frame
Sampled

Figure 3.2.1 Map Illustrating the Address Coverage Enhancement (ACE) Procedure

3. ACE Results for NESARC-III

Table 4.1 shows unweighted segment averages for the number of addresses added through the ACE procedure for NESARC-III. Recall that the average NESARC-III segment had 100 occupied housing units based on the 2010 Decennial Census.

On average, the number of addresses added within a segment was 29.4. This means that in an average segment with 100 occupied housing units, about 29 addresses will be located within the area segment but not contained in the list segment.

Of those 29.4 addresses, 2.2 added addresses on average had unknown components--that is, some part of the address could not be determined by the verifier, such as the house number. These addresses could not be matched to the frame because of the unknown elements and thus must be given a chance of selection during the field period. To ensure the proper weights for sampled added addresses with unknown components, screener respondents were asked for corrected addresses, which were sent to MSG to be matched to the ABS frame. The match rate was 61 percent. The weights of these households which were fielded but then later found to be on the ABS frame were adjusted for their duplicate chances of selection.

The remaining 27.2 addresses on average were sent to MSG to match to the frame. On average, 12.8 of those addresses matched to the frame--that is, 12.8 addresses per segment on average are on the frame, but because of geocoding errors they were not in the list segment. These addresses were not eligible for sampling because they had a chance of selection through another list segment. The weighted match rate of added addresses overall was 50 percent--that is, 50 percent of the ACE added addresses were geocoded to other segments and were not eligible for sampling. The match rate varies for urban and rural counties, with a 55 percent match rate for urban counties and a 43 percent match rate for rural counties. The list segment eligibility rule reduces the extra workload required for fielding added addresses because half of addresses added in the field were not eligible for sampling.

The remaining 14.4 addresses on average that did not match to the frame, along with the 2.2 addresses on average that had unknown components, were eligible for sampling. For the NESARC-III, an average of 4.0 addresses per ACE segment were sampled from the frame of added addresses. Of those, 2.7 addresses on average were found to be occupied housing units. A weighted rate of 65 percent of the sampled added addresses were occupied housing units for the survey. This is lower than the 87 percent eligibility rate for the main sample.

Table 4.1. Unweighted Segment Average Counts of Added Addresses Per ACE Segment

Added addresses	29.4
Added addresses with unknown components	2.2
Added Addresses sent MSG for matching	27.2
Added addresses not matched to CDS file	14.4
Added addresses eligible for sampling (including unknown components)	16.6
Added addresses sampled	4.0
Occupied sampled added addresses	2.7

Table 4.1 summarizes these averages per ACE segment. Individual segment results varied considerably, however, as illustrated in Table 4.2. Twenty-four percent of the ACE segments had no added addresses, and an additional 6 percent had no added non-matching addresses --that is, in 32 percent of segments sampled for ACE, no additional addresses were sampled. Some segments had very large numbers of added addresses. In some of these segments, the number of sampled added addresses had to be reduced to control interviewer workloads. More work is needed to develop a better measure of size for selecting segments for the ACE procedure such that there is better targeting of the areas with the most undercoverage on the CDS file.

Table 4.2 Unweighted Segment Quartiles of Added Addresses per ACE Segment

	Minimum	*25th Percentile*	*Median*	*75th Percentile*	*Maximum*
Added Addresses	0	1	12	48	419
Added Addresses Not Matched to CDS File	0	0	3	22	291
Added Addresses Sampled	0	0	2	7	25

Geocoding accuracy was verified in all segments sampled for ACE. The lister canvassed the segment and confirmed the location of all addresses on the ABS frame. Segment assignment was accurate for 92 percent of addresses, though it varied by urbanicity. The geocoding accuracy rate was 93 percent in urban counties and 84 percent in rural counties. Segments for NESARC-III were quite small, with a minimum of 60 housing units per segment. Forming larger segments would likely result in better geocoding accuracy. Because of the list segment eligibility rule, geocoding errors do not have an effect on the main sample. However, more accurate geocoding would reduce the ACE workload and improve the MOS for subsampling segments for use in the ACE procedure.

Results of the ACE procedure provided estimates of the coverage of the CDS file. The locatable addresses on the CDS file cover 92 percent of all household existing in the United States. Coverage is higher is urban counties, with 96 percent coverage, and lower in rural counties with 78 percent coverage. Ideally, the ACE procedure brings the coverage from 92 percent to 100 percent.

4. Drop Points Addresses (clusters) on the ABS Frame

Drop points are addresses where mail is delivered for several units and it is then distributed internally among the individual housing units, called drop units. Drop points represent clusters of households that require special sampling procedures. The proportion of addresses on the CDS which are drop points is small. A weighted analysis of the ABS frame data for the 150 PSUs sampled for NESARC-III informed the sample design for sampling drop points and drop units. The analysis found that less than one percent of addresses on the residential CDS file are drop points. When expanded to represent the number of housing units, drop units represent less than two percent of housing units on the CDS file. Approximately 93 percent of drop points are urban addresses. The percentage of addresses that are drop points on the CDS file varies greatly by PSU, ranging from no drop points in a PSU to 15 percent of addresses on the frame for the PSU designated as drop points. Large city PSUs have the largest percentages of addresses designated as drop points. Approximately 97 percent of drop points are very small, having two or three units. Of those with more than three units, half have four units and three-quarters have less than ten units. A few drop points have very many units, with some drop points having more than 500 units.

Because the presence of drop points on the frame is rare and the size of the clusters is mostly small (typically two or three units), the “take-all up to three” sample design described in Kalton et al. (2014) was utilized. The CDS file provides two variables on the frame which, if they are accurate, allow for proper sampling of the individual units. The file identifies drop points by an indicator variable And also provides a count of a drop pont’s drop units . PPS sampling was utilized to sample addresses from the CDS file for the NESARC-III. The sampling frame of addresses contains non-drop-point addresses, which each represent only one housing unit, and drop point addresses, which each represent two or more housing units. The MOS for the non-drop-point addresses was 1, while the MOS for the drop points was dependent upon the number of units. For drop points with two or three drop units, the MOS was 1; otherwise, the MOS was the number of units divided by three. Once a drop point has been sampled, the protocol was to interview all the household drop units if the observed number of drop units was three or less and to interview a random sample of household drop units if the number of observed drop units was four or more. If the observed number of units was equal to the number of units listed on the CDS file, the expected number of units sampled was three. The observed number of units often varied from the CDS count, however. If the observed count was less than the CDS count, the sampling rate was not changed if the expected number of sampled units was greater than or equal to one. If the expected number of sampled units was less than one, then the sample rate was increased so that that expected number of sampled units would equal one. If the observed count was greater than the CDS count, the sampling rate was usually decreased so that the expected number of sampled unit was equal to three. Decreasing the sampling rate, however, increased the sampling weights for the sampled drop units. In a few instances, when the observed count was much larger than the CDS count, more than three drop units were selected so that the increase in the sampling weights of the sampled drop units was no more than a factor of three. More than three drop units were also selected if the PPS sampling of the drop point had selected it more than once, which could happen if the list segment contained few or no non-drop-point addresses.

This “take-all up to three” approach is efficient from a field perspective as it allows the field staff to simply interview all households for the majority of drop points without first contacting the home office. When a sample of the households in a cluster is required, the interviewer has to send the list of drop units to the home office, which then selects the sample to be interviewed.

As was expected based on the prevalence of large drop points in the frame, less than one percent of sampled addresses were drop points. However, because of the PPS design which gave a larger probability of selection to larger drop points, 15 percent of the sampled drop points had more than three units.

Sampling drop points based on the information provided on the frame requires the frame data to be reasonably accurate. Of the drop points sampled with a MOS of one (those with two or three units according to the frame information), 43 percent were observed to contain only one unit. Two percent contained more than three units, nearly all of those containing four or five units. However, one drop point which was indicated on the CDS file to contain only three units was observed to contain 60 units. Of the larger drop points (according to the frame information), 54 percent contained fewer units than the CDS file count. Eight percent contained more units than the CDS file count. All but three of these contained fewer than twice the number recorded on the CDS file, with the largest being six times as large as indicated on the CDS file.

5. Auxiliary Variables Present on the ABS Frame

The CDS file contains two variables which, if accurate, could be useful to sampling. The vacancy indicator and seasonal delivery information could be useful in creating a sampling frame that excludes ineligible housing units. These variables are provided on the CDS file by USPS and are not enhanced by MSG.

The vacancy indicator on the CDS file flags any address that has been vacant for at least 90 days. Addresses flagged as vacant on the CDS file were not removed from the frame of addresses for the NESARC-III. Less than three percent of the records on the CDS file were coded as vacant. Reviewing the final disposition of addresses coded as vacant on the CDS file found that 40 percent were eligible for NESARC-III. Removing the addresses coded as vacant according to the CDS file would result in an undercoverage bias, although given the small prevalence of addresses coded as vacant, the bias would be small.

People living in college dormitories were ineligible for NESARC-III . Students in dorms were sampled through their parents' residence. The seasonal delivery variable on the CDS file has an indicator for addresses associated with educational institutions. Addresses flagged as educational institutions on the CDS file were not excluded from the sampling frame for the NESARC-III. Less than one percent of addresses on the CDS file were coded as educational institutions. However, less than 5 percent of addresses flagged as educational institutions were found to be college dormitories; 85 percent were found to be occupied housing units. The designation of educational institutions on the CDS file does not appear to be useful for sampling.

6. Concluding Remarks

Our experience with fielding a large-scale nationally representative in-person household survey utilizing an ABS sample design has been informative. Similar to other studies, the coverage of the ABS frame was found to be quite good, though it is much better in urban areas than in rural areas. Using the list segment eligibility rule increases the coverage of the frame and reduces the amount of coverage enhancement required. The ACE procedure proved to be a useful method to enhance the coverage of the frame. Subsequent studies have improved the urbanicity variable used to create the MOS for sampling segments for the ACE procedure. More work is still needed to improve the MOS so that it accurately targets the segments with the most undercoverage.

At the segment level, geocoding was found to be fairly accurate, especially in urban areas. Larger segments would have even larger rates of geocoding accuracy at the segment level. Also, geocoding accuracy has been improving in recent years and more improvement may be possible. However, since the list segment eligibility rule avoids the reliance on geocoding accuracy for the main sample, greater geocoding accuracy would not affect coverage of the main sampling frame but would improve the ACE procedure, both in terms of creating the MOS and reducing the ACE workload.

Drop points are relatively rare and most are very small, with only two or three units. A PPS sample design paired with the "take-all" procedure was found to be an efficient method for handling drop points. The auxiliary frame variables on the CDS file do not seem to be useful for sampling.

References

Dohrmann, S., G. Kalton, J. Montaquila, C. Good, and M. Berlin (2012), "Using Address Based Sampling Frames in Lieu of Traditional Listing: A New Approach," *Joint Statistical Meetings*, *Survey Research Methods Section*, 3729-3741.

Dohrmann, S., J. Montaquila, T. Buskirk, A. Hyon (2014), "Address-Based Sampling Frames for Beginners," *Joint Statistical Meetings*, *Survey Research Methods Section*, to appear

Eckman, S., and N. English (2012), "Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates," *Field Methods*, 24, 399-408.

Iannacchione, V.G. (2011), "Research Synthesis: The Changing Role of Address-Based Sampling in Survey Research," *Public Opinion Quarterly*, 75, 556-575.

Kalton, G., J. Kali, R. Sigman (2014), "Handling Frame Problems When Address-Based Sampling is Used for In-Person Household Surveys," *Journal of Survey Statistics and Methodology*, 2, 283-304.