A Review of the Address Coverage Enhancement Scheme for In-person Household Surveys

Sylvia Dohrmann[1], Michael Jones[1], Graham Kalton[1], Jean Opsomer[1]
[1]Westat, 1600 Research Blvd, Rockville, MD 20850

Abstract
The U.S. Postal Service's Computerized Delivery Sequence (CDS) file is widely used as the sampling frame for household surveys conducted by face-to-face interviewing. However, the file suffers from some undercoverage for such surveys, partly because unlocatable addresses are dropped from the frame. This paper assesses the strengths of and challenges with the Address Coverage Enhancement (ACE) procedure that aims to supplement the CDS file to deal with the undercoverage. One key strength of ACE is that it is applied in only a subsample of the sampled segments, so that it can be implemented at reasonable cost. Another strength is that ACE retains addresses sampled from the CDS file that are erroneously geocoded to a sampled segment. However, as a consequence, addresses added through the ACE procedure have to be compared to the CDS file to determine if they are elsewhere on the file. This task has sometimes proved time consuming. An important consideration in implementing the ACE procedure is the determination of an efficient sampling design for selecting segments in which the procedure is to be applied.

Key Words: ABS frame supplementation, address coverage enhancement, address-based sampling, geocoding, sampling frame coverage

1. Introduction

In recent years, address-based sampling (ABS) frames have largely replaced traditionally listed frames for household surveys conducted by face-to-face interviewing. However, ABS frames are not without well-documented limitations, which have prompted development of frame enhancement procedures in order to improve coverage for these types of surveys (Kalton, Kali, and Sigman, 2014). Harter and English (2018) review three such enhancement procedures: Enhanced Listing, Check for Housing Units Missed, and Address Coverage Enhancement (ACE). Developed at Westat, ACE is described in this paper.

ACE has been applied in several face-to-face household surveys, including the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC-III), the Population Assessment of Tobacco and Health Study, and the 2017 U.S. Programme for the International Assessment of Adult Competencies. The ACE procedure has unique characteristics that provide advantages over alternative enhancement procedures. However, ACE is not without its own challenges.

The ACE procedure has been described in the literature with varying amounts of detail (Dohrmann, Kalton, Montaquila, Good, and Berlin, 2012; Kali, Sigman, Ren, and Jones, 2014; and Kalton, Kali, and Sigman, 2014). We provide a brief description of the procedure

in Section 2. Section 3 describes the strengths of ACE as an ABS frame enhancement procedure, and Section 4 discusses some challenges of implementing the procedure. Section 5 summarizes some of the improvements planned for future implementations of ACE.

2. Description of the ACE Procedure

The core of an ABS frame consists of the addresses found on the U.S. Postal Service Computerized Delivery Sequence (CDS) file. Dohrmann, Buskirk, Hyon, and Montaquila (2014) provide a comprehensive review of the CDS, including a discussion of how vendors qualify to hold a license in order to obtain the CDS, the processing vendors may perform on the file prior to making it available to survey researchers, and the information available on the file.

Overall, the CDS provides excellent coverage of residential addresses across the United States, which makes it extremely effective for household mail surveys. However, using the CDS addresses in lieu of traditional field listing can be problematic for studies that collect the survey data by face-to-face interviewing. Coverage rates suffer when non-locatable addresses, such as those on rural routes, are dropped from the frame. Also, at the time of sampling, some new construction in high-growth areas might not be included on the CDS. Kali et al. (2014) point out some of the issues faced when using the CDS as the sampling frame for the NESARC-III.

The frame enhancement methodology behind ACE was first proposed in Dohrmann, Han, and Mohadjer (2006) and then revised and further developed by establishing rules for checking missed units against the frame to address geocoding error, as described in Dohrmann et al. (2012). An important distinction associated with ACE compared to other enhancement procedures is that sampled addresses subject to geocoding error are retained in the sample with the ACE procedure, whereas they are dropped from the sample with the other procedures; the other procedures aim to pick up the addresses that are mis-geocoded as part of the enhancement process. As a result, the amount of noncoverage is less with the ACE procedure than it is with the other procedures.

The ACE procedure distinguishes between *area* segments and *list* segments. An area segment is the geographic area (e.g., a group of contiguous census blocks or block groups) that is sampled. In contrast, the corresponding list segment is the set of CDS addresses that geocode to that area. Ideally, all addresses forming the list segment, and only those addresses, would correspond to housing units physically located within the geographic boundaries of the area segment. However, this is often not the case. Figure 1 gives an example of an area segment (outlined in blue) and the addresses on the CDS frame that are geocoded to, or linked to, the area segment (shaded pink). The addresses shaded in the figure form the list segment and the basis of the sampling frame for the segment. In this example, some addresses in the list segment correspond to housing units that fall outside the upper-right boundary of the area segment but that were erroneously geocoded to the area segment. Note that this is a rudimentary example for illustrative purposes. In reality, segments are often larger with more complex shapes, and addresses located miles from an area segment can be geocoded to it (Van de Kerckhove, Krenzke, Mohadjer, and Ren, 2019).

Figure 1: Illustration of an area segment versus a list segment

The addresses for housing units in Figure 1 that are physically inside the area segment but were not assigned to the associated list segment are identified by the ACE procedure in segments in which the procedure is implemented. These addresses may or may not be on the CDS. Any such addresses that are on the CDS would be assigned to a list segment associated with a different area segment. A sample of those addresses not on the CDS is added to the addresses sampled from the list segment.

The steps of the ACE procedure are described by Dohrmann et al. (2012), Kalton et al. (2014), and Kali et al. (2014). They are broadly as follows:

1. Assign probabilities of selection for the ACE procedure to the area segments selected for the main sample and select a random subsample of those segments to undergo the ACE procedure.
2. Canvass each selected area segment to identify addresses physically inside the geographic boundaries but not in the associated list segment.
3. Compare addresses identified in (2) with the CDS, and remove those found elsewhere on that frame. The rest are called "ACE added addresses."
4. Select a sample of the ACE added addresses and add these addresses to the list segment sample.

We discuss these steps briefly in Sections 2.1 through 2.3 below, with the strengths and challenges associated with each step discussed in Sections 3 and 4, respectively.

2.1 ACE Segment Probabilities of Selection

The ACE procedure is performed in only a subsample of the segments in the main sample. The subsample probabilities of selection for ACE vary across segments depending on the potential number of added addresses predicted for each segment. The probabilities of selection are determined as discussed below (see Kalton et al., 2014, for additional details).

Let P(i) be the probability of selecting area segment i for the ACE procedure, let $P(j|i) = 1/k_i$ be the probability of selecting address j added through ACE for selected area segment i, and let r_i be the within-segment sampling rate for list segment i. By setting $P(i) = k_i r_i$, the selection probabilities of the list sample addresses and the ACE added addresses are the same in segment i. If all ACE added addresses are to be included in the sample, $P(i) = r_i$. If a large number of added addresses is anticipated for a given segment, then a value of

$k_i > 1$ may be used so that only a subsample of added addresses is drawn. To achieve this outcome while retaining the equal selection probabilities, the segment is subsampled at a rate $P(i) > r_i$. For example, set $k_i = 3$, so that $P_i = 3$ and the probability of selecting ACE added addresses in segment i becomes 1/3 in order to achieve the same overall selection probabilities for the ACE added addresses as for the listed addresses in the segment. Thus, if, say, 60 ACE added addresses were found, then 20 of them would be sampled for the survey.

In a few study segments, the CDS coverage might be expected to be so poor that such segments should be added to the ACE sample of segments with certainty, that is, P_i=1. For those segments, k_i would be set to $1/r_i$. On the other hand, if the CDS coverage is expected to be very high, k_i might be assigned a value less than 1 in order to reduce the number of segments subsampled for the ACE procedure. For example, with $k_i = 0.8$, $P_i = 0.8r_i$. In this case, if all the added addresses are included in the subsample, the weights of these addresses would need to be adjusted by a factor of 1.25.

2.2 Canvassing Area Segments

Before canvassing work is performed in area segments selected for the ACE procedure, the addresses for the corresponding list segments are loaded onto tablet computers. Field staff trained to perform ACE then systematically canvass each selected area segment. They compare the addresses for each housing unit they encounter within the boundaries of the area segment to the list of addresses on the preloaded list on their computer. If the address for a housing unit is found on the preloaded list, field staff assign the address a status of "located." If the address is not found, they record the address in the system and the computer application flags the address as "added in the field."

Figure 2 shows the four types of addresses the field staff encounter. The addresses denoted by the red dots are the list segment addresses for housing units that are located outside the area segment but that, because of geocoding error, have been assigned to the area segment. List segment addresses correctly geocoded to the area segment are denoted by the blue dots. The addresses denoted by either a yellow or a black dot will be added to the computer by the field staff during canvassing. The yellow addresses are elsewhere on the CDS and are not included as ACE added addresses. The black addresses are the ACE added addresses, from which a sample will be drawn for the survey. However, the distinction between these two types of addresses identified during the canvass cannot be made until they are compared to the CDS as discussed in the next section.

Figure 2: Hypothetical segment with address status

2.3 Comparing Canvass-Identified Addresses to the CDS

Once a segment has been canvassed by field staff, it needs to be determined which, if any, addresses found by the ACE procedure are elsewhere on the CDS. After the addresses identified by ACE are transmitted to the home office, the following steps are performed:

1. The addresses are reviewed in house for accuracy and format.
2. The addresses are sent to the vendor and compared to the CDS.
3. The vendor returns the addresses, indicating which addresses are on the CDS.
4. The addresses indicated as not being on the CDS are reviewed a second time to determine if an edit is possible to increase their match likelihood.
5. Any edited addresses from (4) are sent back to the vendor and again compared to the CDS.
6. The vendor returns these addresses, indicating which are on the CDS.

The first review of the addresses (Step 1) serves two purposes: (1) to apply the usual quality control assessment for such a canvass and (2) to ensure that the addresses identified during the canvass are formatted in the same way as the addresses in the list segment, in order to increase the probability of the vendor finding a match on the CDS. If the format is different, edits may be made to the addresses (e.g., changing a street name from “St. Augustine” to “Saint Augustine”). Upon receipt of the addresses from the vendor indicating which were not on the CDS, Westat performs a second review of addresses that did not match to the CDS (Step 4) to determine if additional edits are necessary. After the addresses found to be elsewhere have been removed, a sample of the ACE added addresses is selected for the survey.

3. Strengths of the ACE Procedure

We will describe three significant strengths of the ACE procedure compared to other enhancement procedures.

The first strength is that ACE is applied in only a subsample of sampled segments. Our experience indicates that the procedure works well with a subsample of about 10 percent of the sampled segments. The probabilistic approach discussed in Section 2.1 allows for the enhancement of the frame to be focused in the areas with the least coverage, while also controlling interview workload.

Another strength of the ACE procedure is the incorporation of geocoding error into its design. Because other procedures do not differentiate between area and list segments, addresses corresponding to housing units not physically located inside the area segment are not eligible to participate in the survey. These procedures therefore rely on their enhancement procedure to identify such addresses and remove them from the sampling frame. With the ACE procedure, all addresses in the list segment frame are retained, and the procedure is relied upon to only enhance the frame and not to reduce it. This is an attractive feature because all field canvassing procedures are imperfect, and requiring both the addition and removal of addresses during the canvass increases the chance for error.

A third strength of ACE is that the timing of the procedure is flexible. The segment canvassing can be done either before or during the data collection period, depending on the needs of the study. If it is done before the data collection period, specially trained field staff, rather than interviewers, can be employed to perform the canvassing step, to better ensure high-quality work. If interviewers are tasked with the canvassing step before the start of data collection, they can perform the work single-mindedly without any conflict with their interview workload in the segment. Another advantage of constructing the frame of added addresses in a segment before data collection is that the sampling of addresses from the list frame and from the addresses added through ACE can be coordinated, so that interviewing at all selected addresses can begin at the same time. However, for timing reasons, it may be necessary to carry out the canvassing during the data collection period. In this case, the most up-to-date CDS can be used as a basis for the fieldwork, and the canvass can be performed by the interviewers already assigned to the segments. The optimal timing of the procedure may vary depending on the study.

4. Challenges of the ACE Procedure and Some Potential Improvements

In Section 3, we discussed three strengths of the ACE procedure compared to other enhancement procedures. However, with these strengths come some challenges.

One challenge is determining the probabilities of selecting segments for ACE, that is, the $P(i)$. With $P(i) = k_i r_i$ and r_i being predetermined as the sampling rate to be applied to the CDS addresses geocoded to the segment, the determination of $P(i)$ reduces to the choice of the value to be assigned to k_i. This choice is based on the expected coverage of the CDS in the segment. If the expected coverage is high, a value of $k_i = 1$ can serve well; with $k_i = 1$, including all the ACE added addresses in the sample gives the added addresses the same overall selection probabilities as the addresses sampled from the CDS frame. Various indicators of coverage have been used when determining whether to increase, decrease, or not adjust the $k_i = 1$ ACE segment selection probabilities. One simple and generally available coverage indicator is the ratio of the number of CDS addresses in the list segment to the number of housing units from the most recent decennial census for the area segment (Kalton et al., 2014). However, this indicator confounds coverage and geocoding error; it is difficult to ascertain if a small ratio is the result of poor CDS coverage or of geocoding error.

The urbanicity of the segment provides another useful CDS coverage indicator. Experience has shown that the likelihood of CDS coverage being poor is greater for rural segments than for urban segments. Urbanicity can be defined in a number of different ways. We have

investigated the relationships of some of the urbanicity definitions with coverage in order to try to determine which is best suited for our purpose.

As discussed in Section 3, geocoding errors do not reduce CDS coverage with the ACE procedure. However, geocoding errors do result in some inefficiencies during canvassing and interviewing. More geocoding error means that more addresses are identified during the canvass, which will then require review and comparison with the CDS. More geocoding error also means that more of the list segment sample addresses will be located outside the area segment; as a result, the set of sampled addresses will be less compact and interviewers' travel will be increased.

Most implementations of ACE have used the geocoding method of Westat's address vendor to form the list segments and as the basis for the field canvass. The most precise level of geocoding provided by the vendor is street address interpolation. This relies on having accurate street address ranges for street sections and interpolates the position of an address along the street section based on the address's street number. When the ranges are inaccurate, addresses can be erroneously geocoded outside (or inside) the area segment. See Dohrmann and Sigman (2013) for more information on geocoding.

In a recent study of ACE, we attempted to improve the segment-geocoded CDS address lists by increasing the number of CDS addresses corresponding to housing units physically within each area segment. We purchased the addresses that the vendor geocoded to each segment and, in addition, we purchased addresses that the vendor geocoded to areas larger than the ACE area segments, by buffering each segment by an amount equal to 15 percent of the square root of the segment's area. We then applied an in-house geocoding method, called land parcel geocoding, to the addresses in the larger areas to create alternative lists of addresses that geocoded to the sampled segments. The most precise level of land parcel geocoding places geocoordinates on the actual address property, with the next level being the street address interpolation. This resulted in a list of addresses that was at least as accurate as the one created using the vendor's method.

As part of the ACE procedure, the field staff record whether the addresses listed on their computers are located within or outside the geographic boundaries of the area segment. For this study, we used this information to compare our vendor's geocoding accuracy to the geocoding based on our in-house land parcel methodology, as applied to the larger area. The results presented in Table 1 show that greater proportions of the addresses assigned to the sampled segments by the land parcel geocoding were located in the segments than was the case with the vendor's geocoding: the land parcel geocoding was much more effective in rural segments, and slightly more effective in urban segments.

Table 1: Percentage of CDS addresses correctly geocoded into the area segments

Type of segment	*Vendor street address interpolation*		*Westat land parcel geocoding*	
	Number of segments	*% of Addresses geocoded correctly*	*Number of segments*	*% of Addresses geocoded correctly*
Urban	388	93	340	97
Rural	201	78	261	92
All	589	91	601	96

To assess the possible gains in efficiency from a more accurate geocoding method, we calculated the percentage of addresses identified during the field canvass that were actually on the CDS but geocoded to another area segment. These results are presented in Table 2.

Table 2: Percentage of addresses added during the ACE canvass that geocoded to a different area segment

Type of segment	*Vendor street address interpolation (%)*	*Westat land parcel geocoding (%)*
Urban	68	47
Rural	37	20
All	52	32

The percentage of addresses that geocoded to a different area segment was significantly reduced using land parcel geocoding. This result implies that fewer addresses needed to be identified during the field canvass, resulting in a more efficient field procedure and fewer addresses needing review and comparison with the CDS. Because the land parcel geocoding method was not used to form the CDS sampling frame for the list segments, we cannot directly assess gains in efficiency resulting from fewer addresses being sampled outside the geographic boundaries of the area segment. Nevertheless, these results indicate that improved geocoding at the segment level would be helpful in improving the efficiency of ABS in general.

A third challenge in applying the ACE procedure is the amount of time and resources necessary to determine which of the addresses identified during the canvass are elsewhere on the CDS. The current method is outlined in Section 2.3 and involves two steps of address review in house, with the first having two components, and two comparisons to the CDS frame. We have found that the first frame comparison identifies 95 percent of the addresses added by the canvass that are on the CDS. In the future, we plan to eliminate the second comparison in order to reduce the amount of time and resources dedicated to the comparison task.

One way to further reduce the resources spent on this task is to restrict the remaining address review to those sampled for interviewing. After selecting a sample of addresses from those identified during the canvass, only those addresses would be reviewed (according to the first step of review described in Section 2.3), to increase the likelihood of a match to the CDS prior to sending them to the vendor. Another possible approach is to only perform a quality control assessment of the field canvass before selecting the sample of addresses for interviewing and then sending the sample to the vendor for comparison to the frame. With this approach, it is likely that some addresses in the selected sample are on the CDS frame, but these may be resolved after obtaining the mailing addresses from the household screening interview and comparing those addresses to the CDS frame. Under both approaches, weight adjustment would be required if some of the sampled housing units are determined to belong to unsampled list segments.

5. Conclusions

The ACE procedure provides an efficient method for enhancing the CDS frame. It is applied in only a subsample of sampled segments and retains all the addresses listed on the frame. Building on the list of addresses geocoded to an area segment, the canvassing

operation can be performed in a short time, often in only a fraction of a day. As with all enhancement procedures, the canvassing needs to be performed with great diligence and be subjected to thorough quality control processes. Otherwise, a sizable proportion of addresses not listed on the sampling frame may still be missed.

When ACE was first applied in our field interview work, we had not anticipated the amount of effort needed to remove addresses identified through the canvassing that were on the CDS outside the given segment. The procedures described in Section 2.3 were established under the assumption that they could be implemented with little effort. With that assumption proving to be false, we have been developing revised procedures for future implementation. Two have been mentioned here: the first is to use improved geocoding methods; the second is to relax the requirement that addresses that are added during the canvass and that are on the CDS frame be eliminated before subsampling. With this relaxation, only the subset of ACE-identified addresses selected using the specified subsampling fraction would require comparison to the CDS frame. A weight adjustment can then account for the fact that some of the selected addresses are removed from the sample prior to interviewing.

We are also considering the timing and level of review that occurs prior to comparison to the CDS frame, with one possibility being to restrict the initial review of the addresses to just the quality control assessment prior to an initial comparison. This would remove a high proportion of the potential addresses before the sample of ACE added addresses is drawn, but might allow some addresses to still be unidentified as being on the CDS frame, and therefore be eligible for sampling and interviewing. Respondents obtained from such addresses can be identified by comparing their reported mailing addresses to the CDS frame. Because these respondents had more than one chance to be included in the sample, their weights would need to be adjusted to reflect this.

In summary, the ACE procedure is proving to be a very efficient method for enhancing the CDS frame. However, the method can be made more efficient in various ways. We are in the process of assessing the efficiency gains from the use of these alternatives, taken singly or in combination.

References

Dohrmann, S., T. D. Buskirk, A. Hyon, and J. Montaquila. 2014. Address based sampling frames for beginners. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 1009-1018.

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

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. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 3729-3741.

Dohrmann, S., and R. Sigman. 2013. Using an area linkage method to improve the coverage of ABS frames for in-person household surveys. *Proceedings of the 2013 Federal Committee on Statistical Methodology (FCSM) Research Conference.* Retrieved September 8, 2019. (https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/05/C1_Dohrmann_2013FCSM.pdf)

Harter, R., and N. English. 2018. Overview of three field methods for improving coverage of address-based samples for in-person interviews. *Journal of Survey Statistics and Methodology*, 6, 360-375.

Kali, J., R. Sigman, W. Ren, and M. Jones. 2014. Experiences with the use of address based sampling in in-person national household surveys. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 3050–3059.

Kalton, G., J. Kali, and 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.

Van de Kerckhove, W., T. Krenzke, L. Mohadjer, and W. Ren. 2019. Evaluation of dwelling unit frame coverage enhancement: Case study of the 2017 PIAAC Survey. *Proceedings of the 2019 Survey Research Methods Section of the American Statistical Association.*