Supplementing Address-Based Sampling Frames with Physical Addresses of Housing Units with Unlocatable Mailing Addresses

Bonnie E. Shook-Sa[1] and Doug Currivan[1]
[1] RTI International, PO Box 12194, Research Triangle Park, NC 27709

Abstract

When utilizing an address-based sampling (ABS) frame for an in-person household survey, households with unlocatable mailing addresses are a primary source of undercoverage. Unlocatable mailing addresses such as post office boxes, rural route boxes, highway contract boxes, and simplified addresses cannot be linked to housing units on the ground. For this reason, approximately two million residential unlocatable mailing addresses are typically excluded from the national ABS frame for in-person surveys. Most of the undercoverage occurs in rural areas where unlocatable addresses are more prevalent. Low ABS coverage in rural areas often leads researchers to resort to a hybrid frame that supplements ABS coverage with costly field listing methods in areas without adequate ABS coverage.

Commercial databases of listed addresses - derived primarily from telephone white pages - contain physical addresses for some households with unlocatable mailing addresses. Our previous research indicated that commercial databases could add a significant amount of coverage to the ABS frame, particularly in rural areas. However, these initial findings assumed that the entire frame would be supplemented, which would not be cost effective for large studies. Additionally, this research was conducted prior to the CDS No-Stat file being made commercially available so it did not take into account the effect of the No-Stat file on supplemental coverage.

We estimate the gains in coverage provided by a targeted commercial database address supplementation approach by combining commercial database address lists with locatable mailing addresses from a commercially available version of the United States Postal Service Computerized Delivery Sequence (CDS) file and the No-Stat file. We find that targeted commercial database address supplementation would not be cost effective for most studies in the presence of the No-Stat file.

Key Words: mailing lists, white pages, frame supplementation procedures, coverage, in-person surveys, CDS file, No-Stat file

1. Introduction

For in-person household surveys, address-based sampling (ABS) frames comprise all active, locatable mailing addresses derived from commercially available versions of the United States Postal Service's (USPS) Computerized Delivery Sequence (CDS) file. The

CDS file is made available to select vendors through non-exclusive licensing agreements with the USPS. Valassis is one company that holds a CDS license, and the Valassis Lists file accounted for all but 35,000 of the addresses on the CDS file in July 2010 (Iannacchione 2011).

ABS undercoverage in rural areas for in-person surveys has been reported by numerous survey researchers (Dohrmann et al. 2007; Iannacchione et al. 2007; O'Muicheartaigh et al. 2007). One of the primary sources of rural undercoverage is the exclusion of addresses from the ABS frame prior to sample selection. While nearly all households have active mailing addresses, not all mailing addresses are amenable to in-person data collection. City-style addresses, which consist of a street number and street name along with the city, state, and zip code are locatable (i.e. they can be linked to housing units on the ground). Unlocatable mailing addresses (UMAs) include "Only Way to Get Mail" (OWGM) Post Office boxes[1] (PO boxes), rural route and highway contract boxes, and simplified addresses[2]. These addresses must be excluded from the sampling frame prior to sample selection because they cannot be linked to housing units on the ground.

There are approximately 1.9 million active UMAs in the United States which account for 1.5 percent of unique addresses on the Valassis Lists file. The national distribution of residential mailing addresses by locatability and address type is displayed in Table 1 below. The vast majority of UMAs are OWGM PO boxes. While UMAs account for a small percentage of addresses overall, they are more highly concentrated in rural areas where 6 percent of addresses are unlocatable[3].

Table 1. National Distribution of Residential Mailing Addresses by Address Type

Address Type	Count [1]	Percentage
Locatable	**122,291,026**	**98.5**
Unlocatable	**1,867,373**	**1.5**
"Only Way To Get Mail" Post Office Boxes	1,499,175	1.2
Rural Route Boxes	222,673	0.2
Highway Contract Boxes	112,826	0.1
Simplified Addresses	32,699	0.0
Total	**124,158,399**	**100.0**

[1] Count of residential addresses from the January 2011 Valassis Lists file.

Field supplementation methods have been developed to improve ABS coverage in rural areas. Two such procedures are the Check for Housing Units Missed procedure and the Waksberg approach (McMichael et al. 2008; Dohrmann et al. 2006). Another option for improving coverage is to implement a hybrid sampling frame, where ABS is used in segments where coverage is expected to be adequate and traditional field listing is used in low ABS coverage areas (Iannacchione et al. 2010). Although a hybrid frame can

[1] Only Way to Get Mail (OWGM) post office boxes are those identified by the CDS file vendor as being associated with households that are located in areas without home delivery of mail.

[2] Simplified addresses do not include a street number or street name. Mail is addressed to the resident's name along with the city, state, and zip code.

[3] Counts of addresses within census block groups were based on geographic classifications from the 2011 Valassis Lists file.

improve coverage, it is both costly and time consuming because field enumeration has to be done prior sample selection.

A less costly alternative is to supplement the ABS frame with supplemental address lists prior to sample selection. Databases of addresses derived from telephone white pages and other commercial database sources contain physical addresses for some housing units with UMAs. For example, someone who receives her mail at "PO BOX 200" could be listed in a commercial database with her physical address, "100 MAIN ST". While we could not send mail to "100 MAIN ST," this address can be sampled and located in the field for in-person data collection. Supplementing the ABS frame with commercial databases prior to sample selection would increase the coverage of the sampling frame in areas containing UMAs.

2. Using Commercial Databases for Frame Supplementation

In addition to containing physical addresses for some households with UMAs, commercial databases also contain a significant number of locatable addresses that are already on the ABS frame. When performing commercial database address supplementation, only addresses from the commercial database that are not on the ABS frame would be included as supplemental addresses. To avoid multiplicities on the sampling frame, commercial database addresses must be purged of valid locatable mailing addresses by matching the commercial database addresses to the ABS frame prior to sample selection.

Our previous research indicated that supplementing the ABS frame with commercial database addresses prior to sample selection could add 7.6 percent coverage to the sampling frame in rural areas (Shook-Sa and Currivan 2011). This estimate assumes that commercial database lists would be purchased for all sampled segments, which would be costly for large studies.

UMAs are clustered together primarily in areas without home delivery of mail (i.e., areas with OWGM PO Boxes). Rather than supplementing the entire ABS list, including areas that do not contain any UMAs, a more efficient supplementation strategy would be to target only areas near the post offices that are associated with UMAs. This targeted supplementation approach would provide cost savings by supplementing only the areas where commercial database address supplementation is expected to add a significant amount of coverage.

The first step to using a targeted supplementation approach is to determine the level of geography that will be used to identify segments for commercial database address supplementation. We limited the targeted areas considered for supplementation to those defined by census geography because physical addresses associated with UMAs cannot be mailed to and are therefore not amenable to postal geography. Targeting based on census geography enables us to identify physical addresses in close proximity to the post offices associated with UMAs.

We used the Valassis Lists file to identify census block groups, census tracts, and counties containing at least one residential UMA. These geographies are the potential geographic levels that could be used to identify and target areas for commercial database address supplementation. For example, if supplementation was performed at the county level, then all sampled area segments within counties containing UMAs would be

supplemented with commercial database addresses, while segments in counties without concentrations of UMAs would not be supplemented.

When selecting a targeted supplementation strategy, there is a balance between the added coverage associated with larger geographic areas and the costs associated with purchasing more supplemental addresses and combining them with the ABS frame. For targeted supplementation to be effective, the physical addresses associated with UMAs need to be allocated (i.e., geocoded) into the same geographic areas as their UMAs. The odds of UMAs and the corresponding physical addresses of the housing units geocoding into the same geographic area increase for larger levels of geography. However, the larger the geographic area used for targeting, the more costly supplementation would be. We considered three potential levels for targeted supplementation:

- **Census block groups:** Using census block groups to identify areas for targeted supplementation would be the least costly approach of those considered, but assuming that OWGM PO Boxes addresses and the physical addresses associated with them geocode into the same census block groups is problematic. The post office associated with the PO Boxes might geocode to one census block group, while the physical addresses of those housing units could geocode to an adjacent census block group.

- **Counties:** Expanding the targeted supplementation approach to counties would lead to supplementing the vast majority of counties in the United States (72 percent). This approach would include all areas that we expect to benefit from commercial database supplementation but would be very costly due to the large amount of addresses that would need to be purchased and combined with the ABS frame. Less than two percent of addresses in counties with one or more UMA are associated with UMAs, making supplementation highly inefficient at the county level.

- **Census tracts**: We used targeted commercial database address supplementation based on census tracts. Supplementing based on census tracts allows us to be confident that we are appropriately identifying areas that would benefit from commercial database supplementation while controlling costs by not purchasing addresses in areas where we do not expect to find UMAs. Areas without concentrations of UMAs are made up of only city-style addresses and would therefore not benefit from commercial database address supplementation.

We identified census tracts likely to benefit from commercial database address supplementation by obtaining the physical addresses of post offices associated with twenty or more UMAs[4]. For each post office, we then identified the census tract that it is located in and the next closest census tract. Nationally, there were 4,308 "tract clusters," or groups of adjacent tracts thought to be associated with concentrations of UMAs. With this targeted supplementation approach, all segments contained in these tract clusters would be supplemented prior to sample selection.

[4] We limited the frame to post offices associated with twenty or more UMAs for efficiency purposes. The requirement of being associated with twenty or more UMAs eliminated 208 post offices from the frame but covered 99.2% of UMAs.

3. Methods

Our prior research indicated that commercial database address supplementation could add a significant amount of coverage to the ABS frame, particularly in rural areas. However, this research assumed that commercial database address supplementation would be implemented for all area segments in the sample. In addition, because of the timing of when this research was conducted, it did not take into account the USPS CDS No-Stat file. The No-Stat file is a supplemental file to the CDS that was made commercially available to vendors holding licenses the CDS file starting in 2009 (USPS 2009). The No-Stat file supplements the CDS file with: 1) addresses on rural and highway contract carrier routes that have been vacant for 90 days or longer. 2) Locatable city-style addresses for PO Box throwbacks[5] on rural and highway contract carrier routes. 3) Locatable city-style addresses for drop units including unit type and number, e.g., APT 14. 4) Addresses of residences under construction.

Based on a comparison of counts of commercially available versions of the No-Stat and CDS files and census data, the combined CDS and No-Stat files likely provide a near-complete listing of housing units in the United States (Iannacchione 2011). While the No-Stat file does not contain physical addresses associated with UMAs, it does contain addresses for a significant number of housing units in rural areas that are not included on the CDS file. Because they are not contained on the CDS file, these addresses are presumed to be primarily inactive addresses associated with vacant housing units.

Using the targeted supplementation strategy outlined in Section 2, we estimated the number of supplemental addresses that are not contained on commercially available versions of the CDS and No-Stat files and would thus be added to the frame and provide additional coverage. Including the No-Stat file in this analysis ensured that we would not include commercial database addresses associated with vacant housing units as supplemental addresses. Because the commercial databases are frequently updated and consist of active listings, we expected the number of commercial database addresses matching to the No-Stat file to be relatively low.

For each of the 4,308 tract clusters thought to contain concentrations of UMAs, we defined a size measure (M_i) equal to the expected number of supplemental addresses contained on the commercial database list in the tract cluster.

$$M_i=(UMA_i + NSU_i /(LMAs_i+UMAs_i+ NSU_i+ NSL_i))* CD_i$$

where
UMA_i = the number of UMAs on the Valassis Lists file associated with tract cluster i
NSU_i = the number of UMAs on Valassis' No-Stat file associated with tract cluster i
$LMAs_i$ = the number of locatable mailing addresses on the Valassis Lists file associated with tract cluster i
NSL_i=the number of locatable addresses on Valassis' No-Stat file associated with tract cluster i
CD_i = the number of commercial database addresses in tract cluster i

[5] A throwback address is a locatable, city-style address where residents elect to receive mail at PO Boxes rather than at their residences.

We then sorted the sampling frame by state to ensure a reasonable geographic spread of tract clusters, selected a probability-proportional-to-size sample of 100 tract clusters, and purchased all of the active, locatable commercial database addresses[6] from these sampled tract clusters. To determine which addresses were supplemental addresses, we matched the sampled commercial database addresses to the Valassis Lists file and Valassis' No-Stat file and purged the sample of all matching addresses.

3. Results

All but 3,277 of the 342,654 commercial database addresses matched to either the Valassis Lists file or Valassis' No-Stat file. These are the supplemental addresses associated with the sampled tract clusters that would be added to the sampling frame if commercial database address supplementation were implemented. Table 2 contains the weighted and unweighted sample counts by match status.

Table 2: Distribution of Commercial Database Address Sample by Match Status

Match Status	n	Wtd n	Lower 95% CI
Valassis Lists Match	326,792	21,454,791	8,716,120
Valassis No-Stat Match	12,585	646,395	435,272
Supplemental Address	3,277	194,331	70,195
Total	342,654	22,295,517	9,334,229

Without the No-Stat file, we estimate that this targeted supplementation strategy would add 840,726 addresses to the Valssis Lists file alone. This estimate is in line with our expectations based on our previous coverage results that were calculated before the No-Stat file was commercially available. However, most of the addresses in the commercial database address sample that were not on the Valassis Lists file were contained on Valassis' No-Stat file. We estimate that our targeted supplementation strategy would only add 194,331 addresses to the combined Valassis Lists and No-Stat files nationally.

4. Discussion

These results indicate that the benefits of commercial database supplementation are more limited than we expected based on our previous findings. For a large national study, it would not be cost effective to purchase all of the commercial database addresses from targeted segments and purge the file of addresses already contained on the ABS frame and the No-Stat file in order to obtain so few supplemental addresses. One probable explanation that explains both our previous findings and these findings is that the No-Stat file provides a similar level of coverage to the commercial databases and that, in the presence of the No-Stat file, commercial databases are not a very effective source of supplementation.

While the No-Stat file does not contain physical addresses associated with UMAs, it does contain locatable addresses that are not contained on the CDS file, primarily from rural routes. We previously thought these addresses were primarily associated with vacant housing units, but because so many of the active commercial database addresses matched

[6] The commercial database sample was purchased from Marketing Systems Group (MSG) and consisted of addresses from two commercial database sources.

to Valassis' No-Stat file we have reason to believe that the No-Stat file contains locatable addresses for a significant number of occupied housing units not included on the CDS file.

Our next steps will be to examine the potential coverage benefits of the No-Stat file to estimate how many locatable addresses associated with occupied housing units are contained on the No-Stat file. If the No-Stat file does contain a significant number of addresses associated with occupied housing units, it should be considered as the primary supplementation source for ABS studies. It has the potential to significantly improve rural ABS coverage.

Acknowledgements

The authors would like to acknowledge RTI staff members Vincent Iannacchione, Jill Dever, Joseph McMichael, Katie Morton, Joey Morris, and Jamie Cajka for their contributions to this research. We would also like to acknowledge the contributions of David Malarek and Steven Dintino of Marketing Systems Group and Art Hughes and Joel Kennet of the Substance Abuse and Mental Health Services Administration (SAMHSA).

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* 2959-64.

Dohrmann, Sylvia, Daifeng Han, and Leyla Mohadjer. 2007. "Improving Coverage of Residential Address Lists in Multistage Area Samples." *Proceedings of the American Statistical Association, Section on Survey Research Methods* 3219-26.

Iannacchione, Vincent, Katherine Morton, Joseph McMichael, David Cunningham, James Cajka, and James Chromy. 2007. "Comparing the Coverage of a Household Sampling Frame Based on Mailing Addresses to a Frame Based on Field Enumeration." *Proceedings of the American Statistical Association, Survey Research Methods Section* 3323-32.

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.

Iannacchione, Vincent. 2011. "The Changing Role of Address-Based Sampling in Survey Research." *Public Opinion Quarterly* nfr017v1-nfr017:1-20.

McMichael, Joseph, Jamie Ridenhour, and Bonnie Shook-Sa. 2008. "A Robust Procedure to Supplement the Coverage of Address-Based Sampling Frames for Household Surveys." *Proceedings of the American Statistical Association, Section on Survey Research Methods* 4329-35.

O'Muircheartaigh, Colm, Edward English, and Stephanie Eckman. 2007. "Predicting the Relative Quality of Alternative Sampling Frames." *Proceedings of the American Statistical Association, Section on Survey Research Methods* 3239-48.

Shook-Sa, B. E., & Currivan, D. B. (2011, May). "Supplementing address-based sampling frames with physical addresses of housing units with unlocatable mailing addresses." Presented at the annual conference of the American Association of Public Opinion Research, Phoenix, AZ.

United States Postal Service. (2009). "CDS User Guide." Retrieved from http://ribbs.usps.gov/cds/documents/tech_guides/CDS_USER_GUIDE.PDF