Improving the Efficiency of Address-Based Frames with the USPS No-Stat File

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

Abstract

Address-Based Sampling (ABS) frames are typically based on the Computerized Delivery Sequence (CDS) file, which the United States Postal Service (USPS) makes available through licensing agreements with qualified vendors. Research based on the CDS file has found the coverage of ABS frames for in-person surveys to be sufficient in urban areas but problematic in rural areas. Because of low rural coverage, researchers often resort to hybrid sampling frames based on both ABS and traditional field enumeration (FE). With a hybrid frame, areas where ABS coverage is expected to be sufficient are allocated to ABS while areas where poor ABS coverage is anticipated are allocated to FE. The more areas that are allocated to the ABS portion of the hybrid frame, the greater the cost savings.

Since 2009, the USPS has made available the No-Stat file, a supplement to the CDS file that contains approximately seven million predominately rural locatable addresses not found on the CDS file. Previous research has indicated that supplementing the CDS file with the No-Stat file could be a cost-effective strategy for improving rural ABS coverage for in-person surveys. Although the overall coverage gains provided by the No-Stat file are modest, No-Stat addresses are clustered in relatively small geographic areas. This clustered aspect of No-Stat addresses means that they could significantly improve ABS coverage in some localized areas. In a hybrid frame design, these coverage improvements move some areas that otherwise would rely on FE to the ABS portion of the frame, which lowers field costs. This paper measures the efficiencies that are gained by including the No-Stat file for a specified hybrid frame design and coverage estimator. Efficiency gains vary widely across states, and are most significant for higher coverage thresholds.

Key Words: ABS, in-person surveys, frame supplementation

1. Introduction

Address-Based Sampling (ABS) is commonly used either as an alternative to or in conjunction with traditional field enumeration (FE) methods for in-person surveys (Iannacchione 2011). FE frames are constructed by specially trained field staff who canvas selected areas and enumerate potential housing units (HUs). ABS frames are derived from the United States Postal Service's (USPS) Computerized Delivery Sequence (CDS) file, which is made available to select vendors through nonexclusive licensing agreements with USPS. ABS frames based on the CDS file provide both time and cost savings over traditional FE because ABS frames do not require field staff to visit area segments in advance of sample selection.

Despite the time and cost savings of ABS designs, concerns remain about the coverage of ABS frames for in-person surveys, particularly in rural areas (Dohrmann et al. 2007; Iannacchione et al. 2007; O'Muicheartaigh et al. 2007). An evaluation of the ABS frame for the National Survey on Drug Use and Health (NSDUH) estimated ABS household coverage to be 96 percent in urban areas, but only 72 percent in rural areas (Shook-Sa and Currivan 2011).

ABS undercoverage comes from several sources. For an in-person survey, field staff must be able to physically locate sampled HUs on the ground. For this reason, mailing addresses that cannot be linked to physical HUs on the ground are typically excluded from the sampling frame (e.g. Post Office Boxes, Rural/Highway Contract Routes, and Simplified Addresses). Undercoverage also occurs due to geocoding error. The CDS file is organized based on postal geography (e.g. postal routes and ZIP codes), but samples are typically selected based on Census geography (e.g. census tracts or census block groups) to allow appended external data to be used in the sample design and weighting. The process of allocating CDS addresses into Census geographies is called geocoding. Error in this process can lead to undercoverage, particularly in areas where geocoding is less accurate (Eckman and English 2012; Morton et al. 2007).

In addition to mailing addresses that are not locatable for an in-person survey and geocoding error, undercoverage occurs because some types of addresses are excluded from the CDS file. The USPS No-Stat file is a supplemental file maintained by USPS that contains approximately seven million locatable residential addresses. Four types of locatable addresses are included on the No-Stat file[1]:

- *Rural Throwbacks*: These are locatable addresses for residents on rural postal routes who specify that their mail be delivered to a PO Box rather than to their residence.
- *Internal Drop*: The No-Stat file provides unit designators (e.g. Apt B) for a small proportion of the drop point addresses contained on the CDS file.
- *New Growth*: The No-Stat file provides addresses associated with residences under construction that are not yet receiving mail.
- *Rural Vacant*: Addresses on rural postal routes that have been classified as vacant for 90 days or longer are moved to the No-Stat file.

Both throwback and vacant addresses that are located on city-style routes are included on the CDS file, and are therefore included on the sampling frames of most ABS studies. However, for administrative purposes USPS relegates these types of addresses to the No-Stat file when they are located on rural routes[2].

The No-Stat file was made available to companies holding licenses to the CDS file starting in 2009. There is very little overlap between the CDS and No-Stat files, so the two files can be easily combined into a single, No-Stat supplemented ABS frame. Prior findings indicate that approximately 21 percent of No-Stat locatable addresses are associated with occupied HUs (households). While this is much lower than the

[1] No-Stat addresses can be classified into these four categories using the PO Box Throwback Indicator, the Delivery Point Type Code, and the CDS No-Stat New Growth Indicator as outlined in the CDS User's Guide (USPS 2013).
[2] Email communication with Pat Wiley of Compact Information Systems (CIS).

occupancy rate for HUs associated with CDS addresses (90 percent), the No-Stat file contains the addresses for a non-trivial number of households that are excluded from the CDS file (approximately 1.3 million). It provides an estimated 4 percent increase in coverage for households in rural areas[3] (Shook-Sa et al., in press).

While the No-Stat file does provide modest household coverage gains for in-person surveys, it does not eliminate rural undercoverage. Even with the No-Stat file, rural household coverage is incomplete. Based on prior No-Stat findings, the combined CDS and No-Stat file provide 77 percent coverage of households in rural areas (Shook-Sa et al., in press). Alternative sampling approaches are needed in areas where ABS coverage is not adequate.

One alternative approach is to develop a hybrid sampling frame, where areas that are expected to have adequate ABS coverage are allocated to the ABS portion of the frame, and areas that are expected to have poor ABS coverage are allocated to the FE portion of the frame. ***Figure 1*** depicts the implementation of a hybrid sampling frame. Following the selection of segments, the ABS coverage in each selected segment is estimated using a pre-specified coverage estimator, which is a pre-established method for predicting the ABS coverage of each segment. A pre-specified ABS coverage threshold is also established, and is typically study-specific. Segments that exceed the ABS coverage threshold are allocated to the ABS portion of the frame, while segments where ABS coverage is not estimated to achieve the threshold are allocated to FE.

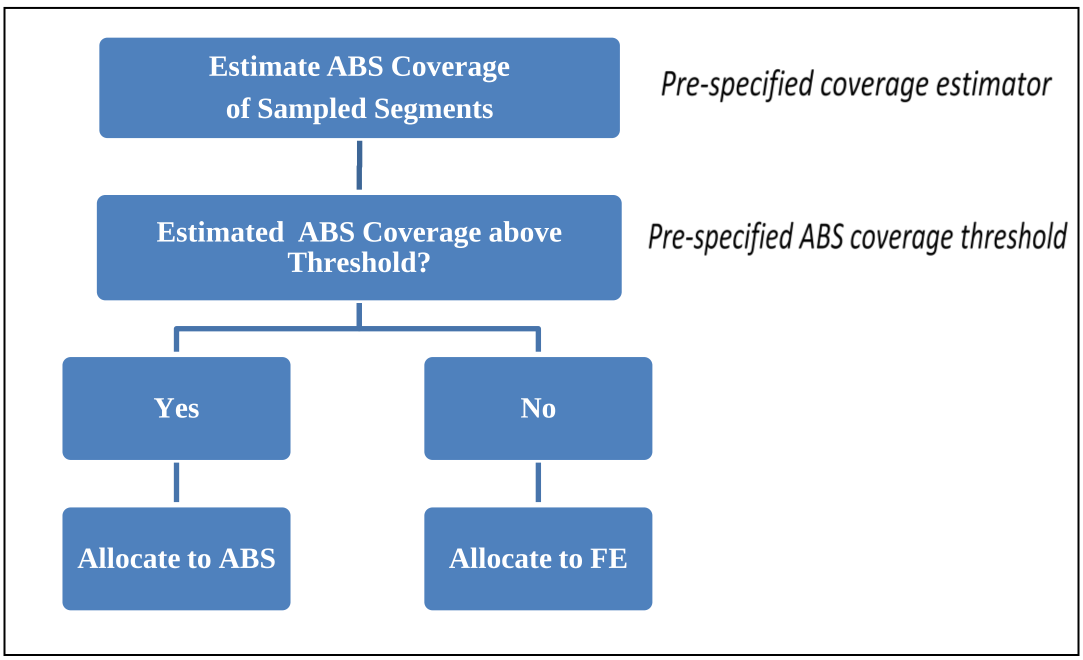

Figure 1: Hybrid Sampling Frame Design. ABS=Address-Based Sampling, FE=Field Enumeration

[3] Area segments were comprised of collections of census blocks. Segments containing only rural census blocks were included in rural coverage estimates.

The estimated ABS coverage for the bulk of segments in the ABS portion of the frame will exceed the pre-specified coverage threshold, so this threshold does not estimate the level of coverage expected from the ABS portion of the frame. The pre-specified ABS coverage threshold is simply a minimum bound for allocating segments to the ABS portion of the frame.

The hybrid sampling frame design provides the cost savings of ABS where possible, but retains FE where necessary to maintain adequate household coverage. In general, the more segments that are allocated to the ABS portion of the frame, the higher the cost savings. The highest cost savings would be realized for lower ABS coverage thresholds, where more segments meet the pre-specified ABS coverage threshold criteria and are allocated to ABS. However, there is a trade-off between cost and coverage, as the expected coverage of the frame decreases as the ABS coverage threshold decreases.

The map in ***Figure 2*** displays the percentage of locatable addresses in each county that come from the No-Stat file (i.e. the number of locatable addresses from the No-Stat file divided by the number of locatable addresses on the combined CDS/No-Stat file). Because No-Stat addresses are clustered in primarily rural areas, the No-Stat file has the potential to move area segments that would otherwise rely on FE to the ABS portion of the frame. If this occurs in enough segments, the No-Stat file could provide significant cost savings in a hybrid design, without having to compromise coverage by reducing the ABS coverage threshold. This paper measures the efficiencies that are gained by including the No-Stat file in a hybrid frame design.

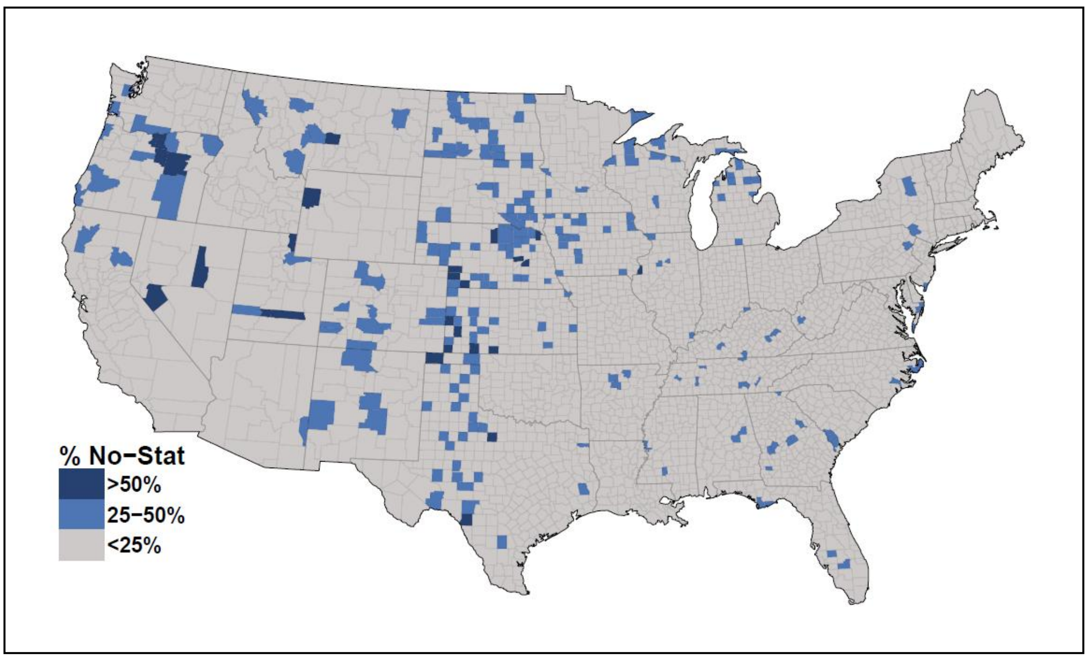

Figure 2: Percentage of Locatable Addresses from No-Stat File by County

2. Methods

The No-Stat file has the potential to move segments that would otherwise rely on FE to the ABS portion of a hybrid frame. Clusters of No-Stat addresses can increase the expected coverage of a segment enough that the expected coverage moves from below to above the pre-specified ABS coverage threshold, effectively moving the segment from FE to ABS.

To evaluate the level of efficiencies this could create, I compared two potential sampling frames. The first is the CDS-only frame, which is the traditional ABS frame. The second is the combined CDS and No-Stat frame. I constructed a national list of area segments. The 215,547 area segments were defined as census block groups or collapsed census block groups, and accounted for the entire United States[4]. After forming area segments, I obtained the March 2013 CDS and No-Stat files that had been geocoded into census block groups and could therefore be linked to the area segments. Within each area segment s, I calculated the number of locatable CDS addresses ($L_{s,CDS}$) and the number of locatable addresses on the combined CDS/No-Stat frame ($L_{s,CDS+}$)[5].

To compare the two potential sampling frames (CDS-only vs. CDS/No-Stat), I estimated the ABS coverage in each area segment for each frame using a pre-specified coverage model, as defined below:

$$C_{s,f} = \frac{L_{s,f}}{HU_s}$$

Where $C_{s,f}$ is the expected coverage of segment *s* based on sampling frame *f*; $L_{s,f}$ is the number of locatable addresses on sampling frame *f* in segment *s* (as defined above); and HU_s is the estimated number of HUs in segment *s* (from the 2010 Census).

After calculating the expected coverage of each segment for both sampling frames, I compared the allocation of segments to ABS and FE under a hybrid sampling design for various potential coverage thresholds. I assumed a national probability proportional to size (PPS) design where the size measure was the adult population from the 2010 Census. PPS sampling allows the probability of selecting each area segment to be proportional to its size measure (in this case, persons 18 or older). This ensures that persons have similar probabilities of selection, regardless of the size of the segments in which they reside.

I examined the expected distribution of segments across ABS and FE under this design to make national and state-level estimates for the differences between the two sampling frames. The difference between the CDS-only percentage of segments allocated to FE and the combined CDS/No-State percentage of segments allocated to FE provides an estimate of the efficiencies gained by including the No-Stat file.

---

[4] Less than one percent of census block groups had to be collapsed when forming area segments because they did not meet the minimum size criteria of 100 HUs (at least one of which was occupied). HU estimates were derived from the 2010 Census.

[5] Locatable addresses from the No-Stat file were merged onto the CDS file. To avoid multiplicities on the combined CDS/No-Stat frame, addresses appearing on both files were removed from the No-Stat file. Internal drops were also excluded from the No-Stat file because the associated drop points are already included on the CDS file.

For a given ABS coverage threshold t, the percentage of segments moved from FE to the ABS portion of the hybrid frame by the No-Stat file (P_{NS_t}) is equal to the percentage of segments allocated to FE for the CDS-only frame (P_{CDS_t}) minus the percentage of segments allocated to FE for the combined CDS/No-Stat frame (P_{CDS+_t}). That is,

$$P_{NS_t} = P_{CDS_t} - P_{CDS+_t}$$

I estimate P_{NS_t} for the specified sample design overall and by state for various potential ABS coverage thresholds.

2. Results

No-Stat addresses were quite concentrated within the 215,547 area segments. While 63 percent of area segments contained one or more No-Stat addresses, segments ranged from having only 0.01 percent of locatable addresses from the No-Stat to being comprised solely of No-Stat addresses. The median percentage of locatable addresses coming from the No-State file was 4 percent.

Figure 3 compares the percentage of segments allocated to FE for the CDS-only frame (P_{CDS_t}) to the CDS/No-Stat frame (P_{CDS+_t}) for four different ABS coverage thresholds t. The differences between the bars represent the total percentage of segments that the No-Stat file moved from FE to ABS (P_{NS_t}). This is a measure of the efficiency gains resulting from inclusion of the No-Stat file. For all ABS coverage thresholds, the No-Stat file provided efficiency gains by boosting the expected coverage of segments enough that a significant number of segments moved from FE to ABS. Efficiency gains were larger for the higher coverage thresholds. For example, at the 90 percent coverage threshold an additional 7.6 percent of segments could rely on ABS rather than FE when the No-Stat file was included, compared to a 1.7 percent gain at the 60 percent coverage threshold.

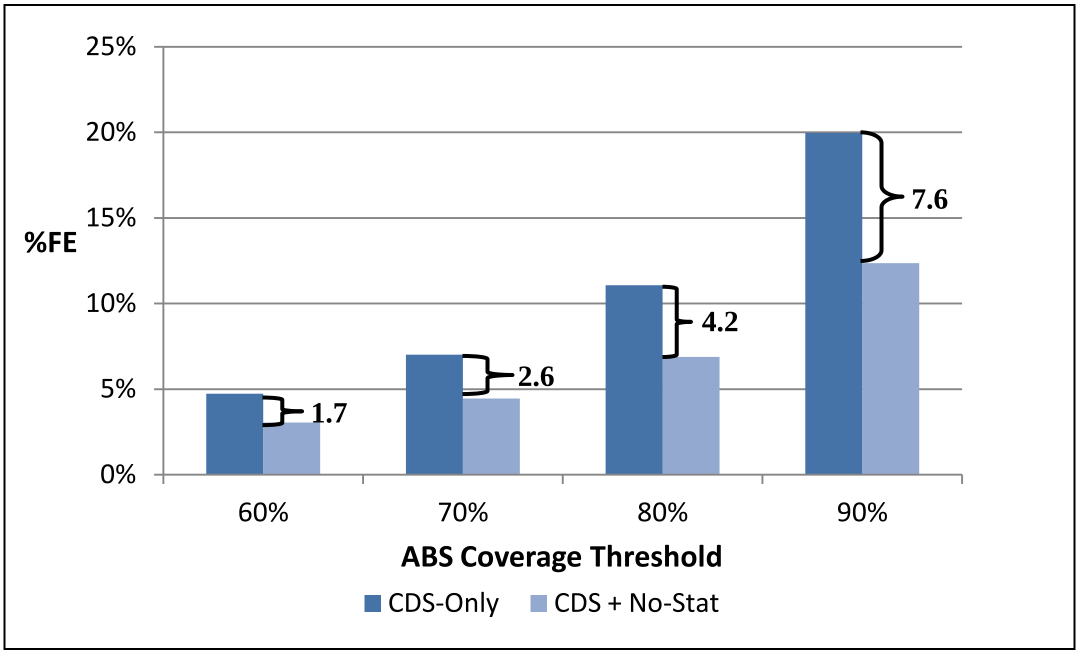

Figure 3: Percentage of Segments Allocated to FE by ABS Coverage Threshold

Because No-Stat addresses are clustered and would not be expected to achieve uniform coverage gains across the country, I also compared the efficiency gains resulting from No-Stat inclusion at the state level. The map in ***Figure 4*** shows the total percentage of segments that rely on ABS rather than FE due to the inclusion of the No-Stat file at the 90 percent coverage threshold ($P_{NS_{90}}$). There are large differences across the states in efficiency gains resulting from inclusion of the No-Stat file. At the 90 percent coverage threshold, the highest efficiency gains are in Iowa, where 20.1 percent of all segments are moved to the ABS portion of the hybrid frame due to the No-Stat file. This is compared to the lowest efficiency gains in Washington, DC, where only 0.7 percent of total segments move from FE to ABS when the No-Stat file is included on the frame. In general, the largest efficiency gains occur in the South and Midwestern states, while smaller gains are realized in more urban states (e.g. California and New Jersey).

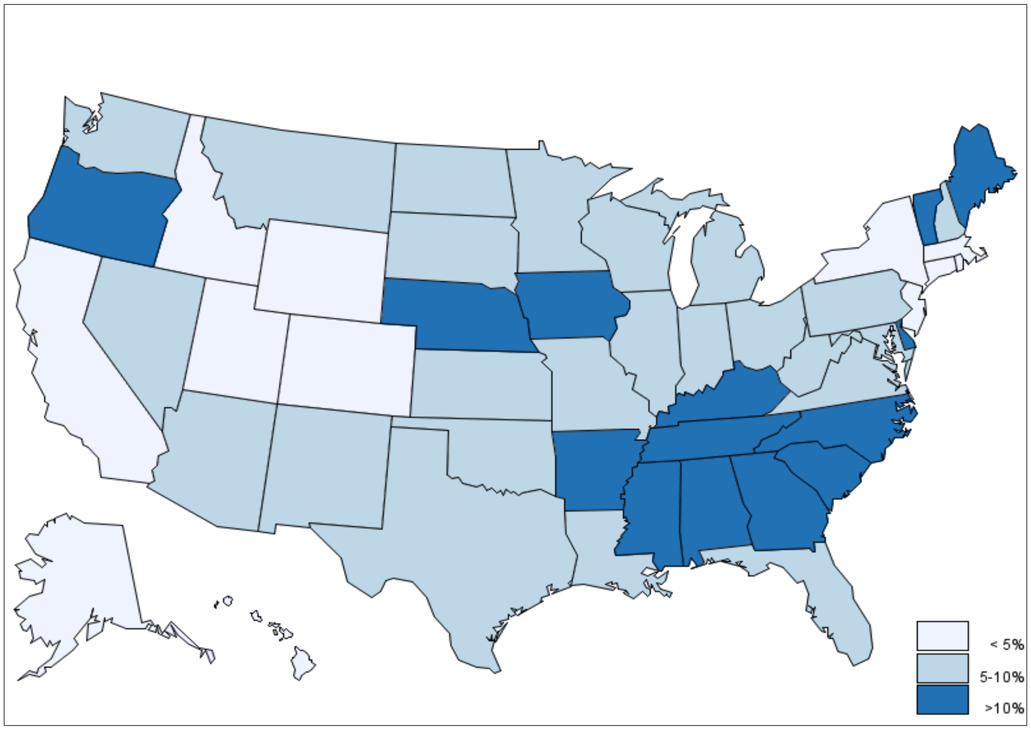

Figure 4: Percentage of Segments No-Stat File Moved to ABS Portion of Hybrid Frame by State (90% ABS Coverage Threshold)

Table 1 provides state-level estimates of the total percentage of segments moved to ABS by the No-Stat file, (P_{NS_t}), for four potential ABS coverage thresholds: 60, 70, 80, and 90 percent. The maximum value of P_{NS_t} is bolded for each coverage threshold *t*. While states with the largest No-Stat gains for a particular coverage threshold tend to see substantial gains for all four thresholds, the state distributions do vary across the four thresholds. For example, North Dakota has the largest gains from the No-Stat file at the 60 percent coverage threshold ($P_{NS_{60}}$= 10.8 percent), but ranks twenty-fifth for No-Stat gains at the 90 percent coverage threshold ($P_{NS_{90}}$= 8.3 percent). These differences are driven by the distribution of expected ABS coverage from the CDS file alone relative to the distribution of No-Stat addresses across the segments.

Table 1: Percentage of Segments No-Stat File Moved to ABS Portion of Hybrid Frame by State and ABS Coverage Threshold[1]

	ABS Coverage Threshold					ABS Coverage Threshold			
STATE	**60%**	**70%**	**80%**	**90%**	**STATE**	**60%**	**70%**	**80%**	**90%**
IA	8.3	12.3	**17.0**	**20.1**	NH	3.2	5.0	5.3	8.1
SC	2.3	4.0	7.8	18.3	OK	3.4	4.1	6.1	7.7
AL	1.3	3.3	8.2	17.9	AZ	1.3	1.6	3.3	7.4
AR	4.7	7.7	12.4	17.8	NM	2.2	3.5	4.9	7.4
MS	2.2	3.9	6.9	17.7	IN	1.5	2.8	4.4	7.3
NC	2.5	4.6	8.5	16.0	PA	1.5	2.2	3.5	6.2
OR	4.7	6.7	9.8	15.8	NV	0.4	0.8	2.9	6.1
VT	8.6	**13.4**	14.5	15.8	LA	1.8	2.2	3.4	6.0
GA	1.3	2.7	5.5	13.9	OH	0.7	1.4	2.7	5.9
DE	1.4	3.4	5.8	12.9	IL	1.4	2.4	3.6	5.8
KY	1.8	3.5	7.0	12.4	TX	1.6	2.2	3.6	5.8
ME	5.7	9.6	10.4	12.0	MD	0.8	1.4	2.5	5.0
TN	0.6	1.4	3.9	11.8	CO	1.6	2.1	2.8	5.0
NE	8.1	10.3	11.9	11.5	NY	1.3	1.9	2.9	4.7
MT	4.1	4.6	7.9	10.0	UT	1.1	1.1	2.3	4.4
SD	7.6	7.2	9.6	9.6	RI	0.1	0.6	1.8	4.1
WA	2.7	3.1	5.1	9.5	WY	4.9	3.6	6.3	4.1
MI	2.6	3.8	5.5	9.3	CT	0.7	0.9	1.3	4.1
MN	3.7	5.1	7.1	9.2	MA	0.9	0.9	2.1	3.9
WI	2.7	4.4	6.4	9.1	HI	2.1	2.3	2.3	3.6
MO	1.2	2.4	4.6	8.8	ID	1.8	2.7	2.3	3.0
FL	0.8	1.7	3.4	8.7	NJ	0.6	0.8	1.3	2.8
VA	1.3	2.4	4.5	8.5	AK	2.9	4.5	4.1	2.3
WV	3.0	4.2	5.5	8.3	CA	0.4	0.6	0.9	1.9
ND	**10.8**	11.7	9.5	8.3	DC	0.1	0.1	0.0	0.7
KS	3.4	5.2	7.0	8.3					

[1] States are sorted by the 90 percent coverage threshold. The maximum value for each coverage threshold is bolded.

3. Discussion

For the specified ABS coverage estimation model, the No-Stat file improves the efficiency of ABS frames for in-person surveys by moving segments that would otherwise rely on FE to the ABS portion of the frame. Under this model, efficiency gains are greater for higher coverage thresholds and vary widely across states.

These results are dependent on the coverage predication model and ABS coverage threshold for a given study, so care should be taken when applying these results if the coverage predication model deviates from the model assumed above. These results are specific to in-person surveys utilizing hybrid sampling frames, and estimates are dependent on the sample design.

While this analysis focused on including the No-Stat file on ABS sampling frames for in-person surveys, further research is needed to assess its utility for other types of surveys (e.g. mail or multi-mode designs). Research is needed about success mailing to No-Stat addresses, appending phone numbers to No-Stat addresses, and the accuracy of some of the indicator variables present on the No-Stat file (e.g. new growth and drop indicators). Despite these areas for future research, the No-Stat file appears to be a useful source for supplementing the CDS file for an in-person ABS design. It provides modest coverage gains overall, but localized coverage gains could lead to significant efficiencies in a hybrid design.

Acknowledgements

The author would like to acknowledge RTI staff members Vincent Iannacchione, G. Lance Couzens, and Rachel Harter for their contributions to this research. I would also like to acknowledge the contributions of Pat Wiley of Compact Information Systems (CIS).

References

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.

Eckman, Stephanie, and Ned English. 2012. "Creating Housing Unit Frames from Address Databases: Geocoding Precision and Net Coverage Rates." Field Methods 24(4): 399-408.

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

Morton, Katie, Joe McMichael, Jamie Cajka, Ross Curry, Vincent Iaanacchione, and David Cunningham. 2007. "Linking mailing addresses to a household sampling frame based on census geography." *Proceedings of the American Statistical Association, Section on Survey Research Methods* 3971-3974.

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., McMichael, J. P., & Iannacchione, V. G. (in press). Beyond the USPS CDS file: Extending the coverage of ABS frames. *Public Opinion Quarterly*.

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. (2013). "CDS User Guide." Retrieved from http://ribbs.usps.gov/cds/documents/tech_guides/CDS_USER_GUIDE.PDF