Enhanced Listing for Improving Frame Coverage: A Review

Ned English[1], Colm O'Muircheartaigh[1,2], and Katie Archambeau[1]
[1]NORC at the University of Chicago, 55 E. Monroe Street, Ste 3100, Chicago, IL, 60603
[2]University of Chicago Harris School of Public Policy, 1155 E. 60th Street, Chicago, IL, 60637

Abstract
While the USPS Computerized Delivery Sequence File (CDS or CDSF) has been shown to be an effective sampling frame for in-person, mail, and multi-mode surveys in most environments, areas with non-city style addresses delivery or those experiencing demolition and new-construction will be at risk for differential coverage in some modes. While databases such as the NoStatistics or "NoStat" may contain some of the housing units missing from the CDS, physical listing may still be necessary. Our paper will discuss enhanced (or dependent) listing, a method employed by multiple organizations to improve undercoverage and reduce overcoverage associated with vendor-provided address lists. We describe the advantages and drawbacks associated with enhanced listing, and discuss the potential impacts on costs, training, and resulting survey data.

Key Words: Address-based samples, address lists, listing, frame coverage

1. Introduction and Background

As well-described in survey research literature, the United States Postal Service Computerized Delivery Sequence File (CDS or CDSF) has become the basis of most address-based (ABS) sampling frames over the past decade, regardless of mode (Harter et al. 2016). Relatedly, there is a considerable body of work validating the coverage and utility of CDS as sampling frame as a replacement for in-field listing, the issue being that the CDS was designed for postal delivery and not household surveys (Iannacchione 2011, Link 2010, Iannacchione, Staab, and Redden 2003, O'Muircheartaigh, Eckman, and Weiss 2003). The consensus in our industry is that the CDS is generally equivalent to traditional listing in most settings with respect to coverage, as defined as the proportion of addresses in reality that are captured on a given list. Secondly, areas of limited coverage may be predicted through comparison to available benchmarks. It is also true that all frames will have deficiencies in some areas, and so there is a question of how best to proceed in areas of known or predicted coverage limitations.

An important factor when considering coverage and related coverage deficiencies in ABS frames is the influence of mode, whether face-to-face, mail, telephone, or in combination (Harter et al. 2016). The choice of mode can impact the types of delivery points that may be included in a frame, and thus the effective coverage and yield. For example, surveys based on mail contact may include non-city-style addresses such as post-office and rural-route boxes, acknowledging the potential for lower eligibility or response-rates. Their inclusion would not be feasible for face-to-face data collection, however, as they do not link directly to dwelling units. As such there may be the potential for coverage issues when

excluding non-city-style addresses in such situations. CDS-derived sampling frames also carry the challenges of special delivery types, such as drop points, seasonal, educational, and simplified addresses which may not be feasible sampling units in all modes (Amaya et al. 2014). Drop points present a particular challenge in that they may not always be accurately mailed to, but could be contacted face-to-face reliably through the derivation of individual ordinal delivery points (Dekker et al. 2012).

National coverage evaluations have produced a range of estimates of how many households are contained on the CDS, with Link et al. (2008) estimating 98% and others in the low to mid 90% (Harter et al. 2016). Such a range is likely due to differential inclusion criteria and research designs between evaluations. While we have seen improvement in CDS coverage over time in general, the question remains of what to do in situations with insufficient coverage for a specific area or study remains. For example, national studies would expect rural areas or those experiencing change or demolition to have coverage deficiencies. There are in fact multiple options of what actions to take in such situations, depending on mode choice. A key question is if one is conducting face-to-face data collection as part of a study, which would render non city-style addresses unusable and thus necessitate different considerations for frame augmentation. Regardless, it would be possible to enhance a CDS-based address frame with external databases. Examples of external sources include address files compiled from commercial vendors, or the USPS no statistics or "NoStat" file (Harter et al. 2016, Shook Sa 2013). In instances where commercial vendors or the NoStat are insufficient for coverage needs, it may be necessary to conduct a listing (Harter, Eckman, English, and O'Muircheartaigh 2010).

Traditional listing is the most basic form of listing, and occurs "from scratch" as listers do not start with any list (Eckman and Kreuter 2011). Alternatively, it is possible to edit an existing list in-person using the process known as "dependent" or "enhanced" listing (Harter and English 2018). Relatedly, one could instead train field interviewers to execute a linking procedure during field interviewing to attempt to find missed housing units as per the half open interval process described by Kish (1965).

Results of prior evaluations have shown CDS-based frames to be at least comparable to traditional listings in urban and suburban areas, especially in environments with regular blocks, single-family homes, and relative stability (Iannacchione 2011). The consistent message of coverage evaluations is that lists have undercoverage and overcoverage error in rural areas (Harter et al. 2016). Such issues are the intersection of discrete processes, including non-city style delivery points, limited geocoding street databases which create geocoding errors, and infrequent database updates. Apparent geocoding error will also be influenced by segment size and morphology, in that errors will be reduced in larger segments.

At question is how best to improve lists should it be necessary, given multiple options. Essentially, the goal would be to minimize undercoverage, resulting from incomplete lists, non-city-style addresses (depending on mode), geocoding error, and vintage issues. At the same time, researchers would prefer to limit overcoverage that resulting from geocoding error, duplication between lists, and vintage (demolition, change). The purpose of this paper is to briefly contextualize and review available options for frame coverage enhancement typically undertaken in survey research today.

2. Options for Frame Enhancement

We can consider options for frame enhancement in two categories: those that require in-field augmentation or listing, and those that do not. One option that does not require listing in advance of field work is the half-open interval or "HOI". The HOI is a Frame correction approach post-sample selection, where field staff are asked to find missing housing units between a selected unit and the next on frame (Kish 1965). For example, if "55 E. Monroe St" were the selected unit and "59 E. Monroe St" was identified as the 'check' or 'next' address, their job would be to see if there were an address in-between, e.g., "57 E. Monroe St". Found units are entered into the study and given a chance of selection based on that of the originally-selected unit (O'Muircheartaigh and Eckman 2011). The HOI has been historically used by many high-profile studies such as NSFG, GSS, HRS, and has the advantage of being inexpensive with staff already in field. In addition, it theoretically could ameliorate all coverage deficiencies if executed properly, but carries challenges if the input frame isn't sorted geographically as USPS files are sorted by carrier route and walk-sequence (McMichael et al. 2008).

One fundamental question is how accurately field interviewers are able to implement the HOI during field data collection, as it has been generally assumed the HOI is implemented with low error rates. O'Muircheartaigh and Eckman (2011) conducted an experiment where they deliberately removed addresses from frame in order to create false "missed housing units" to see if the interviewers would discover and add them back in. In reality, interviewers only found 11% (15/140), which would not have positively impacted coverage substantially. Moreover, of those addresses that interviewers added, 82% were already on the frame. O'Muircheartaigh and Eckman (2011) concluded that the HOI did not ameliorate undercoverage and at the same time contributed to overcoverage and felt it was not an appropriate method to correct frame deficiencies.

A second option that would not require listing would be to conduct database enhancement, such as by using the No Statistics or "NoStat" file. The NoStat File is an administrative supplement to the CDS (Shook-Sa 2013) that contains both active and inactive addresses, including long-term vacant records that no longer receive mail. Shook-Sa (2013) conducted a test that found fielding addresses on the NoStat file did ameliorate undercoverage somewhat while contributing to overcoverage, with records in fact having a 21% occupancy rate. Moreover, commercial vendors provide specialized, targeted lists that could be used to "fill-in" gaps as part of a hybrid frame. Doing so could be conducted in sync with a dependent listing described later.

Thirdly, one could conduct an enhanced or dependent listing. Traditional listing (Kish 1965) had long been considered the "gold standard", but has been recently demonstrated to introduce both undercoverage and overcoverage (Kwait 2009, Eckman and Kreuter 2013) with listers missing addresses that exist and erroneously adding ones that do not. In E-listing, listers have a map and initial frame, which could be from a previous listing or a postal database. The function of the listers is to update the frame in the field by adding missing addresses, deleting inappropriate addresses, and confirming or editing existing addresses by traversing selected blocks in a systematic manner. Enhanced listing is conducted by the US Census Bureau for their Master Address File (MAF) update as well as by the Centers for Disease Control and Prevention (CDC) for the National Survey of Family Growth. Moreover, there are variations described in Harter and English (2018)

such as CHUM (RTI) and ACE (Westat), each having their own characteristics related to implementation and statistical efficiency. Theoretically, enhanced listing should correct for undercoverage by adding missed units at the same time as correcting overcoverage resulting from geocoding error or non-existing units. Starting from an existing list also reduces cost in comparison with traditional listing. Disadvantages of enhanced listing include requiring separate trip and training for staff, and the fact that it would need to be conducted prior to sample selection as part of frame construction.

There are potential limitations for enhanced listing that parallel the same concern with the HOI in the form of "confirmation bias" (Eckman and Kreuter 2011). "Confirmation bias" is the tendency for listers to maintain errors on the initial input list, and is shown at other levels of quality-control in surveys. Eckman and Kreuter (2011) found that dependent listers exhibited both "failure to delete" and "failure to add" in an experiment on the NSFG, with de-duplication between lists representing another source of error found in their experiment. Nonetheless, it is clear that frame enhancement methods have specific advantages and disadvantages in given situations.

3. Evaluation Results

A number of evaluations have been undertaken to ascertain the coverage of the CDS, with or without enhancement (Iannacchione 2011). For example, in 2006 O'Muircheartaigh et al. conducted a national evaluation that compared the intersection of a validated "best" frame of housing unit addresses with the entire CDS, a set of addresses that geocoded in specific segments, and traditional listing. The authors found that the CDS was superior to traditional listings (84% vs. 80%), while geocoding error did introduce some undercoverage as the match rate of those that geocoded inside target segments was less than those overall (74% vs. 84%). The results were encouraging, however, as they indicated the CDS was superior to traditional listing and some frame enhancement could potentially ameliorate over and undercoverage.

In 2010, NORC conducted a second evaluation related to the National Children's Study (English et al. 2012) where segments were paired based on similarity in two counties with rural, suburban, urban, and "small town" environments. One member of each pair was listed traditionally, the other with enhanced listing, with the results independently checked. Finally, the address list was matched to the entire CDS. We then conducted logistic regression to understand method preference by environment (English et al. 2012). Our results found that both the enhanced and traditional listings captured nearly all of reality. The enhanced listing did carry cost-savings in comparison with traditional listing, amounting to 25% in the selected rural county and 5% in urban/suburban areas. Two important caveats in this evaluation include that we did not control for lister experience and our study was not nationally representative.

4. Discussion and Conclusions

As researchers we should be fundamentally concerned about bias, with the impact varying depending on the measure under consideration (Amaya et al. 2018). That is, the fact that a list contains more addresses does not mean it is more representative of a given measure. In addition, one would need to would need to consider differences in key variables between groups. CDS quality has shown to be predictable, with areas requiring augmentation well documented. As such researchers could implement a "surgical" approach to frame construction. Enhanced listing is still necessary in some situations if implementing face-to-face data collection. Rural areas often do have some CDS addresses as starting point, and so we argue for executing enhanced listing in instances where the CDS is not suitable alone. Enhanced listing has the advantage of improving coverage in all environments with lower or equivalent costs relative to traditional. In addition, technology adds the possibility of further improving enhanced listing through the collection of photographic imagery of housing units as well as GPS coordinates to facilitate downstream data collection.

Looking ahead, there are a number of areas where future research would be valuable. First, creating more sophisticated modeling that considers changes in the CDS over time could help focus enhancement efforts. Secondly, integrating multiple lists for commercial sources could improve coverage, with the expectation of substantial over coverage that may need to be remediated. Thirdly, advances in machine learning could allow researchers to process aerial imagery effectively to discover housing units remotely. Integrating aerial imagery with multiple geocoded lists and sophisticated raster population data such as "LandScan" suggest an approach to improving coverage while avoiding in-field listing.

References

Amaya, A. E., Zimmer, S., Morton, K., & Harter, R. 2018. "Does undercoverage on the United States address-based sampling frame translate to coverage bias?" Sociological Methods & Research. https://doi.org/10.1177/0049124118782539

Amaya, Ashley, Felicia LeClere, Lee Fiorio, and Ned English. 2014. "Improving the utility of the DSF address-based frame through ancillary information." Field Methods 26:70-86.

Dekker, Katie, Ashley Amaya, Felicia LeClere, and Ned English. 2012. "Unpacking the DSF in an attempt to better reach the drop point population." Pp. 4596-4604 in Proceedings of the Section on Survey Research Methods: American Statistical Association.

Eckman, S., and F. Kreuter. 2011, "Confirmation Bias in Housing Unit Listing," Public Opinion Quarterly, 75, 139–150.

Eckman, S., and C. O'Muircheartaigh. 2011. "Performance of the Half-Open Interval Missed Housing Unit Procedure," Journal of the European Survey Research Association, 5, 125–131.

English, Ned, Colm O'Muircheartaigh, Katie Dekker, Ipek Bilgen, Lee Fiorio, Mark Clausen, and Tamara Brooks. Predicting When to Adopt Given Frame Construction Methods: Modeling Coverage and Cost Benefits. Presented at the American Association for Public Opinion Research Conference, May 2012, Orlando, FL.

Rachel Harter and Ned English. 2018. Overview of Three Field Methods for Improving Coverage of Address-Based Samples for In-Person Interviews. Journal of Survey Statistics and Methodology, https://doi.org/10.1093/jssam/smx037.

Harter, R., M. P. Battaglia, T. D. Buskirk, D. A. Dillman, N. English, M. Fahimi, M. R 2016. Frankel, T. Kennel, J. P. McMichael, C. B. McPhee, J. M. DeMatteis, T. Yancey, and A. L. Zukerberg (2016), "Address-based Sampling." Prepared for AAPOR Council by the Task Force on Address-based sampling, Operating Under the Auspices of the AAPOR Standards Committee. Oakbrook Terrace, Il. http://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-(2).pdf.aspx Accessed March 1, 2016. 140 pages.

Harter, Rachel, Stephanie Eckman, Ned English, and Colm O'Muircheartaigh.2010. Applied Sampling for Large-Scale Multi-Stage Area Probability Designs. In Handbook of Survey Research, Second Edition, P. Marsden and J. Wright, eds. Elsevier.

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

Iannacchione, Vincent G., Jennifer M. Staab, and David T. Redden. 2003. "Evaluating the use of residential mailing lists in a metropolitan household survey." Public Opinion Quarterly 67:202-210.

Kish, L. 1965. Survey Sampling, New York: John Wiley & Sons.

Kwiat, Aliza. 2009. "Examining blocks with lister error in area listing." in Proceedings of Survey Research Methods: American Statistical Association.

Link, Michael W. 2010, "Address-based sampling: What do we know so far?", Retrieved October 14, 2015, (http://www.amstat.org/sections/SRMS/AddressBasedSampling11-29-2010.pdf).

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad. 2008. "A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys." Public Opinion Quarterly 72:6-27.

McMichael, Joseph P., Jamie L. Ridenhour, Susan Mitchell, Kristine Fahrney, and Wanda Stephenson. 2008. "Evaluating the use and effectiveness of the half-open interval procedure for sampling frames based on mailing address lists in urban areas." In Proceedings of the American Statistical Association, Section on Survey Research Methods: American Statistical Association.

O'Muircheartaigh, Colm, Ned English, Stephanie Eckman, Heidi Upchurch, Erika Garcia Lopez, and James Lepkowski. Validating a Sampling Revolution: Benchmarking Address Lists Against Traditional Field Listing. 2006 Proceedings of the American Statistical Association, AAPOR Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

O'Muircheartaigh, Colm, Stephanie Eckman, and Charlene Weiss. 2002. "Traditional and enhanced field listing for probability sampling." Pp. 2563-2567 in Proceedings of the Social Statistics of the American Statistical Association.

Shook-Sa, Bonnie E., Douglas B. Currivan, Joseph P. McMichael, and Vincent G. Iannacchione. 2013. "Extending the coverage of address-based sampling frames beyond the USPS computerized delivery sequence file." Public Opinion Quarterly 77:994-1005.