Coverage Rates and Coverage Bias in Housing Unit Frames

Ned English[1], Colm O'Muircheartaigh[2],
Katie Dekker[1], Michael Latterner[1], and Stephanie Eckman[1]
[1]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603
[2]Harris School of Public Policy Studies at the University of Chicago, 1155 E. 60th Street, Chicago, IL 60637

Abstract

Address databases, derived from the USPS Delivery Sequence File, can serve as frames for face-to-face area-probability samples as well as multi-mode address based samples. Working with the National Children's Study (NCS), we continue our work to understand the coverage properties of these databases in order to determine what households tend to be covered vs. missed. Recent in-field listings by the NCS give us an opportunity to compare the coverage of the address databases with a frame created in the field. After matching the listed addresses with two versions of the DSF, we returned to those housing units that were missed by one or more sources to collect additional data.

Key Words: Address-based samples, delivery sequence file, National Children's Study

1. Introduction

The National Children's Study (NCS) is a major new initiative whose goal is to use a panel of 100,000 children to understand environmental impacts on child development (Montaquila et al. 2009). It is intended that children will be enrolled in the study before birth and will be followed until age 21 for periodic testing. The NCS will therefore be one of the largest and most complex surveys ever undertaken, and so has been subject to considerable planning with respect to the sample design (Michael and O'Muircheartaigh 2007).

The NCS sample design is based on a housing unit frame generated through traditional listing. "Traditional listing" is a method of address frame generation created by field staff, known as "listers", who record all residential addresses in defined geographies in a systematic manner (Kish 1965). This method of frame creation has been considered the "gold standard" in the survey research industry since the early days of in-person studies (O'Muircheartaigh et al. 2006, O'Muircheartaigh et al. 2007).

Motivated by the high costs associated with traditional listing, survey research organizations have been undertaking recent research into using the United States Postal Service Delivery Sequence File (DSF) as a replacement, at least in urban areas (Montaquila et al. 2009, O'Muircheartaigh et al. 2007). An early use of the

DSF in Dallas County, TX suggested that coverage was adequate for an urban sample (Iannacchione et al. 2003). NORC then began an assessment of the coverage properties of DSF-derived frames with an evaluation in a subset of segments for the General Social Survey (GSS) in 2001 and 2002 (O'Muircheartaigh, Eckman, and Weiss 2003). NORC continued DSF evaluation using a set of inner-city surveys from 2002-2004 (O'Muircheartaigh et al. 2007). The sum total of this research has been a robust finding that the USPS DSF performs at least comparably to traditional listings in urban and suburban areas, and so may be used as a replacement. Rural areas, however, contain a larger share of non city-style addresses, such as PO and rural route box addresses. Consequently, the coverage of the DSF in rural areas is not yet adequate for in-person surveys, which require a housing unit address for sampling purposes. Non city-style addresses may be sufficient for mixed mode surveys, however.

The National Children's Study (NCS) made the global decision to use traditional listing to create a frame in all segments, including areas where previous research has shown the USPS delivery-sequence file (or DSF) to be comparable or superior. The NCS thus presents an opportunity to evaluate traditional listing against DSF addresses through direct comparison.

Our current research investigates key aspects of frame construction and coverage, with implications for in-person surveys beyond the National Children's Study. Our primary goal has been to determine if traditional listing is the ideal method for National Children's Study frame construction by comparing the collected listings to those from the USPS DSF. In so doing we explore the overlap between traditional and DSF-derived lists in common areas, and thus quantify the relative coverage of each. We use two sources of the DSF in the current research, which are from the Valassis (formerly ADVO) and CIS vendors. In the near future we will examine the categories of housing units that are missing from given lists, and thus the types of households that would be expected to be under or over-covered.

2. Methods

Our current evaluation focuses on the Waukesha, WI National Children's Study site, where the field-work is contracted to NORC. Waukesha County is known as a "vanguard site" by the National Children's Study, as it is one of the first counties to undergo fieldwork as a pilot for the remaining sites. Waukesha County is located in west-suburban Milwaukee, and had a population of 360,767 at Census 2000. At that time Waukesha County was approximately 97% White non-Hispanic, and so is not meant to be nationally representative. While Waukesha may generally be described as a suburban county, it does have rural and urban components as discussed below.

For NCS data collection, NORC field staff traditionally listed a representative sample of 17 segments across Waukesha County during the fall of 2008. The 17 segments contained approximately 13,000 housing units, and were spread across the county to capture rural, urban, and suburban environments. We categorized these segments for analytical purposes as being primarily "urban", "suburban", and "rural" based on their population density, street composition, and location within the county. Our method classified five segments as urban, eight as suburban, and four as rural.

NORC then geocoded the USPS delivery sequence file (DSF) provided by the Valassis Corporation for Waukesha County in November, 2008, identifying those addresses that were inside the 17 selected segments. The DSF file provided by Valassis is known as the 'ADVO' file and so we describe it as such in this paper. We matched the two lists using LinkPlus probabilistic matching software package which permits "fuzzy" matching, and so tolerates differences in format, variations in spelling, etc. Our basis for doing so was that if the ADVO database covered the same population of housing units as the traditional listings, we should see a very high rate of overlap between the geocoded database and the traditionally-listed frame. As a last step, we manually reviewed non-matched lines from each source to resolve outstanding issues using internet resources and other reconnaissance. Note that we omitted non-city style addresses from the ADVO DSF file, such as PO BOXes and rural-route boxes. Such delivery points do not provide a direct link with housing units, and so they are not useful for the current analytical processes.

We then repeated the matching process with another version of the delivery sequence file acquired from the vendor CIS, with a data vintage of August, 2008. Our goal in this second match was to discern the variation in coverage between vendors that theoretically offer the same product, with the acknowledgement that the data vintages were somewhat different. Following the second set of matching we had a three-way match, with addresses being present on any of the traditional listings, the ADVO DSF, or the CIS DSF. For purposes of notation we can describe addresses from the traditional listings as being in the "T" frame, those

from the ADVO file as the "A" frame, and those from the CIS file as the "C" frame.

While at this point in the process we had a composite list of addresses within the selected segments in Waukesha County, we had no way to resolve any differences between lists (i.e., we did not know which list was correct in instances of disagreement). To determine which addresses were actually present, we sent interviewers back into the field to validate the veracity of all members of our composite address list. This additional field verification was conducted from late December 2008 through January 2009. Field staff were sent to Waukesha County with the composite list to confirm the existence of each address or to note that the address "did not exist." Staff were also instructed to add any addresses that were present in the segments but not in the union of the T, A, and C lists, which were subsequently de-duplicated against the existing frames by central office staff. The resulting edited and augmented list can be described as the "best" or "B" frame, as it would be the most complete representation of reality.

3. Results

Overall, as shown in table 1, 95% of the B frame was represented by the traditional listings, 92% by ADVO, and 89% by the CIS address vendor. None of the three frames captured all of the "reality" verified in the Waukesha segments, but all were near or above 90%. Note that while these results are unweighted, the National Children's Study sample design introduces very little variation at the segment level, and so the weighted results are essentially identical.

Table 1: Intersection of B with Individual Frames

Intersection	*Percent of B*
Traditional Listings (T)	95%
ADVO (A)	92%
CIS Addresses (C)	89%

Table 2 shows the B ("reality") broken into its component intersections. We can see that 84% of the B frame lies in the intersection of all three frames (A, T, and C). A relatively substantial (6%) portion of the best frame was captured only by the traditionally listed frame. In addition, two percent of the best frame was added by the listers who went back into the field to verify addresses. These are housing units missed by all three of the frames, which would include new construction built during the period when the two DSF lists were compiled and the lists were validated.

Table 2: Components of B

Intersection	*Percent of B*
All Three (A, T, and C)	84%
ADVO and Traditional (A,T)	4%
ADVO and CIS (A,C)	3%
CIS and TRAD (C,T)	1%
Traditional Only	6%
ADVO Only	1%
CIS Only	0%
New Adds	1%
	100%

We would expect the quality of each frame to vary by urbanicity, as shown in table 3. Here we break urbanicity into three categories at the segment level: urban, suburban and rural, as described in section 2. The traditional listings performed better than ADVO or CIS in the urban and rural segments, but were equivalent to ADVO in the suburban segments. Each frame also contained addresses that were "only" on that frame, because they could not be matched to any other source. Relative quantities of such "only" addresses are shown in table 4.

The lower coverage rates for the two DSF frames in the urban areas is somewhat surprising, given robust findings that these lists provide the best coverage in urban (vs. rural) areas. We speculate that urban areas are characterized by a greater proportion of derelict (long-term vacant) buildings than others; such buildings are not present on the DSF file because are vacant and do not receive mail, but would be included by traditional listers. Consequently, the degree of overlap between the best frame and the traditionally listed frame is higher in urban areas, and there are more "traditional only" records in rural areas as shown in table 4. Results of the National Children's Study enumeration will be necessary to determine the qualities of the included and excluded housing units, such as the issue of derelict housing units being added by the traditional listers. In rural areas, the DSF consists largely of PO and RR BOX addresses, which explains the expected superior performance of the traditional listings as non city-style addresses cannot be directly matched to housing units. There were also some addresses that were on the ADVO list but not on any other frame in the rural areas. We can say from the results in tables 2 and four that the ADVO or CIS listings do not include or omit the same addresses as the traditional listings.

ADVO does appear to perform better than the CIS list in the Waukesha segments we examined. Table 1 shows a 3% overall advantage for ADVO, which is present across all three segment categories in table 3. We believe the apparent discrepancies were due to differences in processing and data vintage, rather than fundamental deficiencies with the CIS file. For example, the CIS list was geocoded and subset to the segment geographies by the vendor, while NORC performed these operations on the ADVO file.

Table 3: Intersections of Each Frame with the Best Frame by Segment Urbanicity

Source	*Urban (n=5)*	*Suburban (n=8)*	*Rural (n=4)*	*Overall (n=17)*
ADVO (A)	92%	96%	87%	*92%*
Traditional (T)	97%	96%	94%	*95%*
CIS (C)	93%	93%	82%	*89%*

Table 4: Addresses Only in One List by Urbanicity

Source	*Urban (n=5)*	*Suburban (n=8)*	*Rural (n=4)*	*Overall (n=17)*
ADVO (A)	0%	0%	1%	*1%*
Traditional (T)	7%	3%	9%	*6%*
CIS (C)	0%	0%	0%	*0%*

Because it appears that traditional and DSF-based frames are optimized in disparate environments, generally rural vs. urban, it may be worth considering using the union of multiple lists as a sampling frame for in-person studies such as the National Children's Study. Table 5 shows the coverage of pairs of frames, as well as the union of all three frames (without additional field updating). It is clear that the benefit of combining both DSF-based frames is quite small, as B in ADVO was 92% while B in both ADVO and CIS is 94%. Combining a DSF-based frame with the traditional listings, however, approaches full coverage. There is no benefit of using all three frames. It appears that the DSF-based frames "fill in the gaps" of the traditional listings, and vice-versa.

Table 5: B in Multiple Lists

Source	*Urban (n=5)*	*Suburban (n=8)*	*Rural (n=4)*	*Overall (n=17)*
ADVO (A) and CIS (C)	93%	97%	90%	*94%*
ADVO (A) and Traditional (T)	100%	100%	99%	*99%*
CIS (C) and Traditional (T)	100%	99%	98%	*99%*
All Three Methods	100%	100%	99%	*99%*

4. Discussion and Conclusions

Our multiple list evaluation found that traditional listing produced results that were closest to reality in Waukesha County, WI, with 95% of the traditional listings in the 'best' frame. Of the three frames, ADVO had the second highest coverage at 92%, with CIS at 89%. ADVO coverage was comparable to the traditional listings in suburban areas of the county, but under-performed somewhat in both urban and rural sections. We feel, however, that the difference in urban areas is due to chronically-vacant units that have been removed from the

ADVO file. Such addresses would be considered "out of scope" for field interviewing, and thus may not be of concern. In line with previous findings, rural areas are best covered by traditional listing, due to the presence of non city-style addresses.

If one chose two methods to use concurrently, it would be most effective to pair the DSF database offered by ADVO with traditional listings. Such an approach would be a relatively inexpensive way to improve the coverage within Waukesha County, WI, and we speculate it could be the same in others. However, it would require list matching coupled with manual de-duplication as in the present analysis.

It is clear from this and other evaluations that database-derived and traditionally listed frames are indeed different, and therefore may be expected to include dissimilar types of households. For example, DSF-based lists may favour households that tend to exhibit particular consumer behaviors, while the coverage of traditional listings may be influenced by housing type and tenure. Going forward, we will be researching the categories of households that are missing from particular lists in a number of ways. First, we will be conducting interviews with all housing units in the selected segments for the National Children's Study, and will thus have screener data to compare. We have also collected qualitative information about the segments at the block face data, in addition to photographs of individual units, and so can describe the kinds of housing units and their neighbourhoods that tend to be missed. Lastly, we will be matching addresses to household-level demographic data provided by the MSG vendor.

References

Iannacchione, V.G., Staab, J.M., and Redden, D.T. 2003. Evaluating the use of Residential Mailing Addresses in a Metropolitan Household Survey. *Public Opinion Quarterly*, 67, 202-210.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley and Sons, Inc.

Michael, R.T. and O'Muircheartaigh, C. 2008. Design Priorities and Disciplinary Perspectives: The Case of the U.S. National Children's Study. *Journal of the Royal Statistical Society: Series A*, 171, Part 2, 465-480.

Montaquila, J., V. Hsu, J. Michael Brick, N. English, and C. O'Muircheartaigh. 2009. A Comparative Evaluation of Traditional Listing vs. Address-Based Sampling Frames: Matching with Field Investigation of Discrepancies. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

O'Muircheartaigh, C., Eckman, S., and Weiss, C. 2003. Traditional and Enhanced Field Listing for Probability Sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

O'Muircheartaigh, C., English, N., Eckman, S., Upchurch, H., Garcia, E., and Lepkowski, J. 2006. Validating a Sampling Revolution: Benchmarking Address Lists against Traditional Listing. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

O'Muircheartaigh, C., English, N., Eckman, S. 2007. Predicting the Relative Quality of Alternative Sampling Frames. *Proceedings of the Survey Research Methods Section, American Statistical Association*.