Comparing the Coverage of Alternative Address Frames for the 2009 Residential Energy Consumption Survey

Rachel Harter[1], Katie Dekker[2], Anna Wiencrot[2], Ned English[2], Jizhou Fu[2], Mary Hess[2], Hina Shah[2], Brad Parsell[2]
[1]RTI International, P.O. Box 12194, Research Triangle Park, NC 27709
[2]NORC at the University of Chicago, 55 E. Monroe St., Suite 3000, Chicago, IL 60603

Abstract
Address-based studies require high quality address sources for frame construction. The 2009 Residential Energy Consumption Survey (RECS) for the Energy Information Administration (EIA) employed an address frame based on the USPS Delivery Sequence file (DSF) supplemented by traditional and enhanced listing. The purpose of this research was to understand differences in coverage among the DSF, traditional listing, and enhanced listing in areas of questionable coverage. This research set out to: 1) match the addresses in segments that were used in both the 2005 and 2009 RECS in order to understand changes over time and changes in method, and 2) examine the coverage of the DSF in 2009 RECS segments that were proximate to the threshold where we could have used any of the methods.

Key Words: Sampling frames, address-based sampling, area probability studies

1. Introduction

First conducted in 1978, the Residential Energy Consumption Survey (RECS), sponsored by the Energy Information Administration (EIA), collects energy-related data for a nationally representative sample of occupied primary housing units. The thirteenth RECS, conducted in 2009, collected data for 12,083 households. These data are tabulated for the four Census regions, the nine Census divisions, and 16 states. These 16 states vary by geography, climate, and population size (U.S. Energy Information Administration, 2011).

The 2009 RECS differed from previous study rounds in that, in concurrence with EIA and NORC at the University of Chicago, the sample would employ an address frame largely based on the USPS delivery sequence file (or DSF) in addition to traditional listing. Segments could have used the DSF, traditional listing, or an enhanced version of the DSF as the sampling frame depending on their nature. The purpose of this research was to understand differences in coverage among the DSF, traditional listing, and enhanced listing in areas that could have employed any of the methods. Three study questions were examined for this research: 1) What is the coverage of the DSF in segments close to the threshold where we could have field listed? 2) Would it have been possible to use the DSF in segments that were traditionally listed for the 2009 RECS? 3) Would it have been possible to use the DSF in segments that were enhanced listed for the 2009 RECS? We explore each of these research questions in Sections 2-4, following an introduction to each of the three methods of listing.

1.1 Traditional Listing

Historically, the in-person address frames have been created by traditional listing in the field. Traditional listing is the decades-old process of sending a field representative (lister) to a selected segment to collect addresses (Kish, 1965). The lister's advance information on the segment is generally limited to a segment ID, overview maps, an expected measure of size, and perhaps some demographics for the general area. By NORC convention, the lister begins in the northwest corner of the first numbered block and proceeds around the block, recording addresses as she goes. When she returns to her starting point on the first block, she then proceeds to the starting point on the next numbered block. The hand-written list of addresses is examined for consistency and errors, sent for data entry, and then quality-checked again. The resulting list is sequenced by block and in sequential order within block. Traditional listing is considered to be the most costly listing method, and has been shown to be subject to under-coverage (O'Muircheartaigh et al., 2006; O'Muircheartaigh et al., 2007). Additionally, traditional listing may result in the creation of an address frame of non-standard format. For example, in rural areas with non-city-style addresses, the listed "addresses" are often descriptions, as housing units may not have formal addresses because they receive their mail at Post Office boxes or Rural Route Boxes. Still, traditional listing is often the only viable option for sampling frame construction in areas without city-style addresses.

1.2 DSF-based Listing

In urban and suburban areas dominated by city-style addresses, the U.S. Postal Service Delivery Sequence File (DSF) has emerged as a considerably more efficient alternative to traditional listing. DSF-based listing is a procedure used since the early 2000's whereby office staff geocode addresses from a copy of the U.S. Postal Service Delivery Sequence File (DSF) and extract the addresses corresponding to the selected segment (Iannacchione, Staab, & Redden, 2003; O'Muircheartaigh, Eckman, & Weiss 2003; Staab & Iannacchione, 2003). The list is often further refined to exclude businesses and specialized housing such as educational housing (dormitories) and seasonal housing. The list is in the order in which mail carriers deliver mail, by block. Because no travel or field work is involved to create the list, DSF-based listing is the least expensive listing process that NORC uses for producing address lists in advance of sampling. The coverage of DSF-based listing has been evaluated to be comparable or superior to traditional listing in urban and suburban areas, but less effective in rural areas with non-city-style addresses.

1.3 Enhanced Listing

Enhanced listing (sometimes referred to as e-listing) is a hybrid approach in which a lister travels to the selected segment equipped with a preliminary list. The lister's job is to "check" the list by verifying the existence of addresses, adding housing units that do not appear on the list, and deleting housing units that are not present on the ground. The lister also makes corrections to the information on the preliminary list. NORC continues to evolve its processes for enhanced listing, which are currently being deployed nation-wide in the National Children's Study. In some versions of enhanced listing employed by NORC, the lister puts the preliminary list in a traditional listing order. Enhanced listing contains advantages of both traditional and DSF-based listing. While it is more costly than DSF-based listing because of the field travel costs, it is less time-consuming than

traditional listing. Enhanced listing also has theoretical coverage advantages over either approach since the starting list is verified independently in the field.

Enhanced listing can be performed with any reasonable preliminary list. For the 2009 RECS, the e-listers used the 2005 traditional listings as the starting point.

1.4 2009 Decision Rules for Determining Appropriate Use of the DSF to Update the 2005 RECS Listings

The rules for determining whether to use the DSF to update 2005 listings for the 2009 RECS were as follows:

1. 2005 segments containing blocks with the 2000 Census-defined TEA[1] code having value 1 were preliminarily considered "urban"; otherwise they were considered "rural".

2. We further assessed the ratio of DSF counts to the Claritas[2] scaled[3] count derived from Census projections. If the ratio (DSF/Claritas scaled) was greater than or equal to 0.80, the segment was classified for DSF updates.

3. Next, for the remaining segments, we evaluated the ratio of the DSF count to 2000 census counts. If this ratio was greater than or equal to 0.9, the segment was classified for DSF updates.

4. Finally, the remaining segments were manually reviewed through satellite imagery and a review of the vacant housing unit counts to determine whether DSF or field visits were appropriate.

The rules for determining whether to use the DSF or field listing for new 2009 RECS segments were identical to the rules listed above. These rules were based on past experience with determining where it was possible to employ the DSF or field listing (English et al. 2009; O'Muircheartaigh et al. 2007). Such research demonstrated the influence of the share of addresses that are city style (analogous to TEA code) and the ratio of DSF counts to Census-based housing unit controls (such as Claritas).

As a practical matter, the initial TEA designation for "urban" or "rural" was ultimately unnecessary for purposes of determining whether DSF or field listing was appropriate. (The TEA designation was essential for determining the segment measure of size prior to segment selection, but that was a different function.) For listing, the TEA designation was

[1] TEA is an acronym for "type of enumeration area", and indicates how the U.S. Census Bureau conducted their enumeration for the 2000 decennial census. Of importance is whether enumeration could be conducted by mail (generally urban, TEA 1) or by in-person canvassing (non-TEA 1).

[2] Claritas is a product of the A.C. Nielsen market research firm and provides annual estimates of households down to the block group level based on decennial census figures and proprietary modeling.

[3] NORC scaled the 2000 Census counts using 2009 Claritas counts. We assumed that the average growth rate from 2000 to 2009 was constant within a block group and that the growth rate for an individual census block was the same as that of its census block group. (Claritas counts were purchased at the block group level, and RECS needed estimates at the block level). We then computed the block group rate of growth as M = (Claritas 2009 HUs) / (Census 2000 HUs) and scaled the Census 2000 HUs for individual blocks by this factor M.

overruled by the ratio rules. For this research, we focused on cases whose DSF/Claritas ratios were in a specified range close to .85.

2. Creation of New Listings for a Sample of Segments that used the DSF for the 2009 RECS

As discussed in the previous section, a decision process was used to determine which segments for the 2009 RECS would be traditionally listed, which would have existing listings updated, and which would use DSF listings. To determine how well this decision process worked, we examined a sample of segments close to the boundary of the decision rule fundamentally based on the ratio of DSF counts to Claritas counts as described above.[4] This section examines the first study question: What is the coverage of the DSF in segments close to the threshold where we could have field listed?

2.1 Methods

We selected a sample of segments[5] from the original 2005 segments that used the DSF as the 2009 frame. Approximately 2/3 of the sample was from the 1,148 segments defined as being "urban" according to the Census TEA code. We defined segments as urban if at least 95% of the housing units in the segment were in blocks classified as TEA code 1. We refer to such segments as "TEA code = 1" and those below the TEA threshold as "TEA code = 0". Of these "urban" segments, half were from those segments where the DSF/Scaled Claritas ratio was less than .85, and the other half had ratio values at least .85. We selected the other third of the sample from the 84 segments with TEA code = 0 and where the DSF/Claritas ratio was between .80 and .90. We also selected two additional segments purposively where the DSF/Scaled Claritas ratio was substantially smaller. Ultimately, 48 segments were selected. These segments were then enhanced listed, using the September 2010 Valassis DSF as both the preliminary list for enhanced listing and the DSF frame for comparison, in order to reduce timing issues.

2.2 Findings

The basic summary of the comparisons of this confirmation process for all 48 segments is shown in Table 1 below. The enhanced list contains slightly more lines (6987) than the DSF (6489). One explanation for the higher counts in enhanced listing is that field listers are trained to include all HU structures they identify, whether or not they are vacant, while the DSF omits addresses that have been vacant (not received mail) for an extended period of time. Our experience is that HUs added during the enhancement process tend to have considerably lower eligibility rates than those confirmed during the enhancement process, as they are more likely to be vacant (O'Muircheartaigh, Eckman, & Weiss, 2003). The tradeoff of including vacant HUs is one of coverage versus efficiency. Including them gives the HUs a chance of selection in case the HU is occupied at the time of data collection. Conversely, truly vacant HUs are not eligible, and sampling them inflates the initial sample and work assignments.

[4] If the DSF/Scaled Claritas ratio was greater than or equal to 0.80, the segment was classified for DSF updates.

[5] For cost and scheduling efficiency we selected geographical areas that had segments with various qualities of interest and randomly selected segments within those areas.

Table 1. Comparison of 2010 Enhanced Listings with 2010 DSF for All 48 Segments

	2010 Enhanced DSF		2010 DSF	
	Count	Percent	Count	Percent
Confirmed Lines	5797	83%	5797	90%
Lines Added in Field	1190	17%		
DSF Lines Deleted in Field			672	10%
Total Lines	6987	100%	6469	100%

A more detailed breakout of the match for the 46 segments with ratios in the .80 to .90 range of the total of 48 is shown in Tables 2 and 3. It appears that the DSF is a better match to reality when TEA=1, regardless of whether the DSF to Claritas ratio is greater than or less than .85. Perhaps this speaks to the error inherent in allocating Claritas block group data to blocks; that is, the error in assuming all blocks within a block group experience the same growth rate. One would also expect some degree of error in the Claritas housing unit estimates themselves, as they are based on models derived from market-research data. Therefore, the DSF to Claritas ratio may not by itself be a very precise measure of the suitability of the DSF in providing relatively complete coverage.

It is also important to re-emphasize that non-city-style addresses on the DSF are not geocodable[6], and so cannot be verified in the field. One assumes, then, that some quantity of lines will be added in the field in rural areas (non-TEA=1) as some housing units receive mail via non-city-style delivery.

Regarding the nature of the discrepancies, the enhanced listings may reflect the newest construction and vacant HUs not on the DSF, and may not have addresses of structures that were demolished since the DSF update. During review, we noted that some rural areas are less complete in the DSF; the DSF was also more likely to miss new HUs. In addition, geocoding error could cause discrepancies in either direction.

[6] Geocoding is the process of interpolating a longitude and latitude coordinate for an address so that they can be mapped. Non-city-style addresses generally geocode to ZIP code centroids.

Table 2. Comparison of the 2010 Enhanced Listings with 2010 DSF for 30 Urban Segments With 2009 DSF to Claritas Ratios of .80 to .90

	Urban (TEA=1)			
	Ratio ≥.85		Ratio <.85	
	Count	Percent	Count	Percent
Total 2010 DSF Lines	3072		1570	
Total Enhanced DSF Lines	3324		1605	
Confirmed Lines	2861		1514	
Lines Added in Field	463	14%	91	6%
DSF Lines Removed in Field	211	7%	56	4%

Table 3. Comparison of the 2010 Enhanced Listings with 2010 DSF for 16 Rural Segments With 2009 DSF to Claritas Ratios of .80 to .90

	Rural (TEA=0)			
	Ratio ≥.85		Ratio <.85	
	Count	Percent	Count	Percent
Total 2010 DSF Lines	847		811	100%
Total Enhanced DSF Lines	935		944	100%
Confirmed Lines	590		663	
Lines Added in Field	345	37%	281	30%
DSF Lines Removed in Field	257	30%	148	18%

For the two purposively selected segments with much lower DSF to Claritas ratios, the DSF was a very reasonable representation of the addresses in these segments. One segment had a 2009 DSF/Claritas ratio of .27 and a DSF/Census ratio of .33. After manual review, however, the determination was made to use the DSF anyway, which turned out to be much closer to current "truth." The second segment had a 2009 DSF/Claritas ratio of .45 and a DSF/Census ratio of 1.15 with four newly constructed homes (per Valassis) in the last few years. This is one case where the allocation of Claritas totals from block groups to blocks may have been misleading. The DSF/Census ratio proved a reasonable second test.

3. Review of Traditionally Listed Segments New to the Sample

This section addresses the second study question: Would it have been possible to use the DSF in segments that were traditionally listed for the 2009 RECS?

3.1 Methods

Some segments that were new to the 2009 RECS sample were traditionally listed in the field. We examined a small sample of 20 such segments that were close to the decision cutoff to determine whether the DSF could have been used successfully after all. We matched the 2009 listings to the August 2009 DSF and summarized as for the other segments. We note that none of the segments has a DSF to Claritas ratio very close to .85 because all segments with ratios greater than .80 used DSF frames. Thus the comparison was restricted to segments whose DSF to Claritas ratio was between .70 and .80.

3.2 Findings

The results of this matching are shown in Table 4 and 5, below. In these tables we consider addresses that are "on" or "not on" the two lists as being "matched" or "non-matched," respectively. The traditionally listed counts are significantly higher than the DSF counts, and the number of traditionally listed addresses not captured by the DSF is high in rural areas, even allowing for some subjectivity and error in the match process. This effect is to be expected as non-city-style addresses are fundamentally unmatchable.

Table 4. Comparison of 2009 Traditional Listings with 2009 DSF for 10 Urban Segments (TEA=1)

		Urban DSF Lines		
		On List	Not On List	Total
Trad. Lines	On List	1786	502 (22%)	2288
	Not On List	290 (14%)	0	-
	Total	2076	-	-

Table 5. Comparison of 2009 Traditional Listings with 2009 DSF for 10 Rural Segments (TEA=0)

		Rural DSF Lines		
		On List	Not On List	Total
Trad. Lines	On List	335	404 (55%)	739
	Not On List	74 (18%)	0	-
	Total	409	-	-

We make the following anecdotal observations about non-matches in the above segments:

1. If the housing numbers were unknown, it was more difficult to match between DSF and traditional listing. 140 out of 3027 lines (4.6 percent) do not begin with a digit (specifically, "NO#" or "NO.") in the traditional listings.
2. For some streets, DSF has odd house numbers where traditional listing has even numbers. This occurred in two segments. In the first, there were 26 addresses on the list and 10 addresses on the DSF that could not be matched because of the odd-even street numbers. In the second segment, there were 30 list addresses and 35 DSF addresses that could not be matched because of the same issue.
3. Traditional listings have more apartment units.
4. DSF has more streets (possibly new).

Without a more careful examination into the specifics of the addresses, it appears that where the DSF to Claritas ratio is less than .80, traditional listing is slightly more complete in urban areas and significantly more complete in rural areas. This is not surprising given that manual review in 2009 also determined that DSF was not adequate in these areas. It is also true, however, that the DSF does have some coverage in areas believed to require traditional listing and could serve as a very effective starting point for enhanced listing.

4. Review of Enhanced-Listing Segments from 2005

This section examines the final study question: Would it have been possible to use the DSF in segments that were enhanced listed for the 2009 RECS?

4.1 Methods

Some segments that were traditionally listed in 2005 were revisited in the field in 2009 for enhanced listing. That is, the listers were equipped with the previous 2005 field listings and asked to mark updates. We selected a small sample of such segments that were close to the decision cutoff to determine whether the DSF could have been used successfully after all. For this research we compared the enhanced listings to the August 2009 DSF. Again, because all segments with DSF to Claritas ratios greater than .80 used a DSF frame, these segments have ratios in the range of .70 to .80.

4.2 Findings

The results of this matching are shown in Table 4-6, below. In these tables we consider addresses that are "on" or "not on" the two lists as being "matched" or "non-matched," respectively. The traditionally listed counts are significantly higher than the DSF counts, and the number of traditionally listed addresses not captured by the DSF is high in rural areas, even allowing for some subjectivity and error in the match process. This effect is to be expected as non-city-style addresses are fundamentally unmatchable.

Tables 6 and 7 below summarize the results of the comparisons.

Table 6. Comparison of 2009 Enhanced Listings with 2009 DSF for 10 Urban Segments

		Urban DSF Lines		
		On List	Not On List	Total
Trad. Lines	On List	1398	599 (30%)	1997
	Not On List	257 (16%)	0	-
	Total	1655	-	-

Table 7. Comparison of 2009 Enhanced Listings with 2009 DSF for 10 Rural Segments

		Rural DSF Lines		
		On List	Not On List	Total
Trad. Lines	On List	278	357 (56%)	635
	Not On List	156 (36%)	0	-
	Total	434	-	-

We can make the following observations from the comparisons:

1. In the enhanced lists, 138 out of 2632 urban and rural lines (5.2 percent) have "NO#" in the street address, especially when the street names are "County Road" or "U.S. Highway" or other rural roads. In addition, the addresses that are on county roads and highways are more difficult to match to the DSF.

2. The DSF has more streets, and possibly contains new streets.

3. The apartment units and numbers are different in traditional listings and the DSF (e.g. Apt 1L vs. Apt 1). These were treated as matches where it made sense to do so. Although we matched units to the best of our ability within the context of the listings, some units remained as mismatches, as we wanted to err on the side of conservative matching.

4. There are some cases with different street names referring to the same street (e.g. State Road 227 vs. State Route 227, U.S. 84 HWY E vs. Highway 31). We matched where we could do so confidently, but were unable to confirm some differing street names as the same street, so some of these ended up as mismatches.

As noted above, we recognize that many of the mismatches will in fact be matches that are impossible to determine, so conclusions regarding coverage are approximate and conservative. Nevertheless, the results in Table 4.1 support the conclusions from Table 3.1 that segments with DSF to Claritas ratios below .80, and whose manual review indicates that DSF may be inadequate in these areas, probably are better covered by a field visit, especially in rural areas.

5. Conclusions

Probably most segments whose DSF to Claritas (or DSF to Census) ratios are less than .80 will have better coverage from field listings, whether traditional or enhanced. The DSF can be a starting-point for enhanced listing, even in rural areas.

A decision rule in the vicinity of .80 or .85 for the DSF to Claritas ratio seems reasonable at this time, especially for urban areas, but this one decision rule should not be the sole determinant for the method of producing a list. The DSF to Census ratio is also useful, especially if the census counts are relatively current.

Manual review using online resources is a valuable addition to the decision rules, but is feasible for only a limited number of segments. We recommend manual review especially where the DSF/Claritas and DSF/census ratios are in conflict or where the ratios seem low for an urban area. Manual reviews are time-consuming and costly, but less so than a field visit and so should be considered by survey managers.

Given the similarities in results in Sections 3 and 4, and the substantial difference in time and costs, we recommend enhanced listing over traditional listing in segments where the DSF alone is deemed inadequate.

Acknowledgements

This research was conducted by NORC for the Energy Information Agency (EIA) under contract DE-AM01-04EI45001, while all authors were employed by NORC at the University of Chicago. The authors gratefully acknowledge EIA's research guidance and support, and note that the conclusions are the responsibility of the authors.

References

Eckman, S. and Kreuter, F. 2010. Confirmation Bias in Housing Unit Listing. *Public Opinion Quarterly* (forthcoming).

English, N., O'Muircheartaigh, C., Eckman, S., Dekker, K., and Latterner, M. 2009. Coverage Rates and Coverage Bias in Housing Unit Frames. *Proceedings of the American Statistical Association, Survey Research Methods Section* [CD ROM], Alexandria, VA: American Statistical Association.

Groves, R. M., Mosher, W. D., Lepkowski, J. M., and Krigis, N. G.. 2009. Planning and Development of the Continuous National Survey of Family Growth. *Vital Health Statistics* Sep(1), 1-64.

Harter, R., Eckman, S. English, N., and O'Muircheartaigh, C. 2010. Applied Sampling for Large-Scale Multi-Stage Area Probability Designs. In *Handbook of Survey Research, Second Edition*, P. Marsden and J. Wright, eds. Emerald.

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(2), 202-210.

Iannacchione, V., Morton, K., McMichael, J., Shook-Sa, B., Ridenhour, J., and Stolzenberg, S. 2010. The Best of Both Worlds: A Sampling Frame Based on Address-Based Sampling and Field Enumeration. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Link, M. D. 2010. Address Based Sampling: What Do We Know So Far? Webinar sponsored by the Survey Research Methods Section of the American Statistical Association and the American Association of Public Opinion Research. November 30, 2010.

McMichael, J. P., Ridenhour, J. L., and Shook-Sa, B. E. 2008. A robust procedure to supplement the coverage of address-based sampling frames for household surveys. In *JSM Proceedings,* Survey Research Methods Section, Alexandria, VA: American Statistical Association. 4329-4335.

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

O'Muircheartaigh, C., English, N., Latterner, M., Eckman, S., and Dekker, K. 2009. Modeling the Need for Traditional vs. Commercially-Available Address Listings for In-Person Surveys: Results from a National Validation of Addresses. 2009 Proceedings of the American Statistical Association, AAPOR Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

Staab, J. M. and Iannacchione. V. G. 2003. Evaluating the Use of Residential Mailing Addresses in a National Household Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association.*

U.S. Census Bureau. 2006. Technical Paper 66: Design and Methodology, Current
Population Survey. Technical report.

U.S Energy Information Administration. 2011. *About the RECS.*
http://www.eia.gov/consumption/residential/about.cfm