

Residential Address Lists vs. Traditional Listing: Enumerating Households and Group Quarters

Sylvia Dohrmann, Daifeng Han, Leyla Mohadjer
Westat, 1650 Research Blvd., Rockville, MD 20850

Keywords: Area samples, USPS, address lists, HOI procedure

1. Introduction

In area household surveys, a multi-stage sampling approach is frequently used to create a nationally representative sample. Such sample design involves four stages of selection: 1) the formation, stratification and selection of primary sampling units (PSUs) consisting of counties or groups of counties; 2) the formation and selection of secondary sampling units (SSUs) consisting of Census blocks or block groups, which are often referred to as segments; 3) the listing and selection of dwelling units (DUs) within segments; and 4) the enumeration and selection of eligible individuals within DUs.

Traditionally, before the selection of DUs in the third step above, enumerators are sent to the SSUs to record the addresses lying within the boundaries. The enumerators are assisted by maps created using the Census boundaries. These lists are used as the frame for the selection of DUs.

Recently there has been much interest in using address lists originating from the United States Postal Service (USPS) as area housing unit sampling frames in place of the more costly on-site enumerated lists. It is still unclear as to whether these lists are adequate as substitutes for on-site enumeration. For example, there are issues with obtaining information on special units such as dorms or other group quarters as well as concerns with undercoverage of these lists. In this paper we will discuss the sources (section 2.1), coverage (section 2.2), cost and comparability of the lists from different vendors (section 2.3). We will also present practical aspects of using these lists as sampling frames (section 3) for area samples including ways to handle missed units on the address lists (section 4).

2. Characteristics of the Address List

2.1 List Sources

Residential address lists cannot be purchased directly from the USPS. Rather an organization having a list of

residential addresses can get the delivery information for those addresses, and thus confirm that those addresses are correct, after qualifying and purchasing a license with the USPS. These organizations are usually vendors dealing with direct mail or other marketing agencies. The better vendors will have one of two licenses with the USPS: a license to the Delivery Sequence File (now in its second generation and known as the DSF2) and the Computerized Delivery Sequence (CDS) File. The two files are both “built from the Address Management Services (AMS) database. The AMS database contains the USPS’s official record of mailing addresses.”¹ However, there are a couple major differences.

The DSF2 is a computerized file that contains information on all delivery point addresses serviced by the USPS, with the exception of general delivery. (In cases which carrier route or PO Box delivery is not available, general delivery mail is held at a main post office for recipients to claim within 30 days.) Each address record submitted by the vendor that matches the file is assigned the ZIP+4 Code, carrier route code, delivery sequence, delivery type, and seasonal delivery information. The USPS does not correct or add addresses to the vendors’ lists during this process. However, any erroneous addresses will be indirectly identified since they will lack delivery information after this process. The DSF2 is updated monthly.

The CDS is similar to the DSF2. It is a 5-digit ZIP Code-based file that provides the same delivery information as the DSF2. The major difference is that the USPS will not only attach the delivery information for the units the vendor and USPS have in common, but the USPS will also update the vendor’s list in the process – adding or removing records as necessary, as well as make other corrections. Vendors must first qualify for CDS information within a 5-digit ZIP Code by already having at least 90 percent but not more than 110 percent of all the addresses in a given ZIP Code. If the vendor does not have this level of coverage, their files will not be updated; note, however, that the vendor will continue to sell the list even though it is not up-to-date. The CDS file is updated every two

¹ Excerpted from 2004 CDS User Guide published by the USPS.

months. The cost of obtaining a CDS license is more than that for the DSF2 license.

The cost of the lists will vary by vendor. Some vendors “add value” to the lists by attaching geocodes, Census information, or other data; their lists tend to be more expensive. Prices range from as high as \$25 per 1,000 addresses to as little as \$8 per 1,000 addresses. ZIP Code is the smallest level of geography for which lists can be purchased. Given that the size of ZIP Codes can vary greatly (see Table 1), the cost of the address lists can also vary.

Table 1. Average number of housing units per ZIP Code by population density

County population density (number of persons per square mile)	Estimated average number of DUs in a ZIP Code*
> 10,000	15,821
5,000-9,999	12,017
1,000-4,999	8,734
400-999	6,612
200-399	4,812
70-99	3,307
50-69	2,467
30-49	1,933
10-29	1,571
< 10	688

* Data calculated from the 2000 Census SF1 file using Census ZIP Code Tabulation Areas (ZCTAs) as an approximation for ZIP Code.

2.2 Sources of Undercoverage

There are several sources of potential undercoverage of these lists: DUs receiving mail via PO Boxes only, DUs along rural routes, and group quarters. In rural areas, the first two sources are more pervasive.

In the literature, there has been much discussion of undercoverage of these lists in terms of PO Boxes and Rural Routes. All companies we investigated can provide the physical address of people with PO Boxes since in most cases people who have a PO Box also have mail delivered to their homes. DUs for which the PO Box is the only method of mail receipt are more of an issue. Staab and Iannacchione (2003) estimated these cases to be about 1.3 percent of the households nationwide.

Rural routes are another source of undercoverage. In such cases, the address is simply in terms of a route number (not a recognizable street number) and a box number. Even if the route was identifiable, the box for the route may not be near the residence. In such cases it is not possible to locate the physical location of the DU based on the mailing address. Staab and

Iannacchione (2003) estimated that 3.9 percent of the households nationwide are unlocatable rural routes. One address vendor we investigated, Compact Information Systems, (CIS) has estimated that they are missing somewhere between two and three million rural addresses. The DSF2 and CDS are replacing the rural route numbers with the street-style addresses as they are available. The conversion of addresses to the street-style addresses in rural areas for purposes of 911 location will eventually diminish this issue. However, until that time, in rural areas, on-site enumeration may still be necessary.

The absence of PO Box-only and rural route DUs is not an issue for mail surveys, but can be an issue in area sample surveys.

Many area sample surveys are household surveys, but there are several that cover the civilian noninstitutional population which includes group quarters. Noninstitutional group quarters include dorms, assisted living facilities, halfway homes, and shelters. Group Quarters are not identified on the USPS lists. On the CDS file, there is a flag which may be used to identify educational units (i.e., dorms). However, the presence of these units on the file depends on how residents of the educational facility receive their mail. Some facilities operate their own post office, and thus the USPS does not have information on individual mailing addresses of the residents.

Other facilities, such as assisted living facilities, halfway homes, and shelters may be operated by a business or charitable organization. If residents’ mail is not delivered to the residents’ individual dwelling units, but instead to the business unit, the facility will not be included on a purchased residential address list. In order to include such group quarters on the list a research organization might include businesses in the address list purchase. However, this would require a prohibitive amount of screening to locate such organizations.

2.3 Evaluation

In 2005, we performed an evaluation on an area sample of the civilian noninstitutionalized population including group quarters. Three areas for which addresses were collected via traditional listing within secondary sample units (formed using Census geography) were considered for the evaluation: a moderately urban/suburban area (with 4,000 enumerated households); a very urban area (also with 4,000 enumerated households); and a rural area (with 3,000 enumerated households). We performed two types of comparisons. The first was a comparison

among vendors' lists for the urban/suburban area. The second was between the enumerated and purchased lists for all three areas.

Address lists from Donnelley Marketing, ADVO (part of the American List Council), and CIS were purchased for the urban/suburban area. A fourth vendor, Anchor Computer, was also considered.

Table 2 shows the licenses used by the four vendors and how closely their reported counts of units (given to us prior to purchasing) compared to the 2000 census. Since Anchor Computer reported fewer units than the 2000 census, despite reported growth in the area since 2000, a file was not purchased from that vendor.

As can also be seen in Table 2, the file received from Donnelley Marketing had the largest number of records reported before purchasing. However, the Donnelley list contained several duplicate records, erroneous apartment numbers, and a large number of addresses with incorrect ZIP Codes.

It was found that the CIS and ADVO lists were identical once duplicate records (pertaining to multi-drop sites such as high rise apartment buildings) were removed from the ADVO list. Since the ADVO list was much more expensive than the CIS list, and the Donnelley list deemed unacceptable, CIS was the primary vendor used for the comparison with the enumerated lists for the three areas.

Table 2. Summary of vendor comparison

	ADVO*	CIS†	Donnelley Marketing‡	Anchor Computer§
License type	CDS	CDS	DSF2	DSF2
Reported units on file	227,040	226,884	237,381	209,461
Difference from Census 2000	4%	4%	7%	-4%
Residential units (after file manipulation)	226,884	226,884	234,724	-
Cost per 1,000 records	\$25	\$12	\$9	\$8

* ADVO includes duplicate records for drops.

† CIS includes one record for each unique address. The number of records in each drop is indicated in a separate field. Current cost is \$15/thousand.

‡ The purchased file included PO Boxes and duplicate records.

§ A file was not purchased from Anchor computer, so the number of residential units could not be verified.

We matched the enumerated lists with the purchased lists for each of the three areas in several ways. First, the enumerated and purchased lists were merged together by all address fields (house number, street

name, pre- and post-direction, unit number, and ZIP Code). Any enumerated addresses that failed to match on all fields, were merged again to the purchased list by the same fields excluding the unit number/designator. Any enumerated addresses failing this match were merged a third time by geocoded latitude and longitude (obtained by our internal GIS), where possible. These second and third matching steps were conducted to overcome any differences in unit designations (such as apartment A, B, C vs. 1, 2, 3), or spelling (such as Ft. Meyer Blvd vs. Fort Meyer Blvd). Any remaining unmatched records were investigated manually.

The same team of highly qualified field enumerators created the lists for each area. The match results for the enumerated addresses in the three areas are shown in Table 3. The percentages in the table assume that the addresses on the enumerated lists are closer to the truth and are expressed in terms of the number of enumerated addresses that match to the vendor list.

In the urban areas the match rate was quite good when excluding group quarters. The urban/suburban area included two college campuses which were not available on the purchased lists (neither CIS nor ADVO), so the match rate "with group quarters" included in the denominator is quite low.

Table 3. Percent of enumerated addresses matched to vendor lists

	Match rate		
	Urban/suburban	Highly urban	Rural *
Without group quarters	99.1%	97.2%	76.8%
With group quarters	79.1%	94.8%	76.8%

* There were no group quarters identified in this county.

For the rural area, CIS was unable to provide addresses in three of the ZIP Codes covered by the sample so an additional list (of all sampled ZIP Codes) was purchased from ADVO. ADVO was able to provide addresses in all ZIPs, but the coverage in the ZIPs common to CIS was slightly lower than that for CIS. Regardless, the match rate was less than adequate in the rural area.

3. Two Sampling Approaches Using Address Lists

When using address lists for area sampling, two approaches can be used for selection of first and second stage units. The first approach is sampling by ZIP Codes (Staab and Iannacchione, 2003), where the

PSUs and the SSUs are respectively 3-digit and 5-digit ZIP Codes. Census demographic information used for sampling would be at the ZIP Code tabulation area (ZCTA) level. ZCTAs are ZIP Code equivalent areas on the Census files as defined in 2001, but they do not reflect any new ZIP Codes or changes in ZIP Code boundaries since 2001.

Another sampling approach utilizes Census geography, where the SSUs (or segments) consist of adjacent or nearby Census blocks within area PSUs. The objective, though, is to obtain all addresses within the segments. Most vendors do not geocode their address lists into Census blocks; those that do cannot guarantee the accuracy. Therefore, we recommend purchasing the address lists by ZIP Code as the finest geography. After segments are selected, geographic information systems (GIS) can be used to find the 5-digit ZIP Codes containing the sampled segments. Then address lists based on these ZIP Codes could be purchased. Finally, through geocoding, the purchased addresses could be positioned into or outside the sampled segments, creating a frame for within-segment sampling.

Both approaches have their advantages and disadvantages. Sampling by ZIP Codes has the advantages of simplicity and potentially less clustering of the sample. On the other hand, ZIP Codes usually cover very large areas, which could result in higher traveling cost and more burden on field staff. The inconsistency between ZCTAs and ZIP Codes can also be a problem. Since ZCTAs are rooted in the 2001 ZIP Code definitions, the samplers (survey statisticians) will need to somehow determine the newly developed ZIP Codes or they will be missed. Even if the newly developed ZIP Codes can be determined, it may be difficult to incorporate them into the measure of size calculation since no Census data can be obtained for these newly developed ZIP Codes.

When sampling by Census geography, the PSUs are generally counties or groups of counties. The segments are adjacent Census blocks which are small enough to make field work operationally feasible. Even if an area is selected for which the address lists are found inadequate, it is still possible to send enumerators to list the area since the segments are formed in an efficient manner for field listing. However, this approach usually imposes an extra cost associated with purchasing addresses outside the sampled segments. More importantly, after the list is purchased, each address has to be geocoded to determine which ones fall into the sampled segment. In addition, there are issues with those addresses that are not geocodable. (In one rural county, we found that 15 percent of the

addresses were not geocodable; this percentage was closer to 2 percent in the suburban and urban sites we examined.) Unless we can make a good speculation about their Census geography, these nongeocoded cases have to be given a chance to come into the sample through coverage improvement measures. We will discuss more about the issue of coverage improvement in the next section.

At the time that we started this research, we had selected counties as the PSUs for our survey. Thus, the discussions in the later sections will focus on sampling by Census geography.


4. Coverage Improvement When Using Address Lists for Area Sampling

Coverage improvement has always been an important issue in area sampling. Address lists have coverage problems, especially in rural areas and areas with rapid growth. To improve their coverage, the samplers should investigate ways to improve the purchased lists. After the sample is fielded and before the screening interview, a quality check should be conducted to pick up any addresses that were missed on the geocoded lists used as the frame for DU selection. This quality check is often referred to as the “missed structure procedure.”

4.1 Missed Structure Procedure Used for Sampling Frames Created Through Geocoding

As mentioned above, the frame file for within segment sampling is created through geocoding. Since the geocoding process is not completely accurate, small discrepancies may occur between geocoded address lists and Census geography. In this case, we suggest using the geocoded address lists as both the sampling frame and the comparison base for the missed structure check. Since entire ZIP Codes were purchased and geocoded, a much larger area beyond the sampled segments is included on these lists. As long as each valid address geocodes into one and only one segment, the inaccuracy of geocoding will not cause any sampling error.

Figure 1 is an illustration of the missed structure procedure. Suppose the solid line represents segment 1 boundaries based on Census geography and the geocoded addresses fall in an area represented by the dashed line. These are the possible scenarios:

1. The address indicated by  is on the geocoded list for segment 1, so this address should be kept in the sample even though it falls outside the segment boundaries on the map.

- The address indicated by ☼ is not on the geocoded list for segment 1 but will be found within the segment boundaries. In this situation, staff should search through the entire geocoded list purchased for the PSU (not just those addresses in the sampled segments, but in the list of all purchased ZIP Codes). If the address cannot be found on the list, then they would include it into segment 1. If the address is found anywhere on the list, then it was geocoded into another Census geography, and already had a chance of being selected through other segments. In this case, it should not be included into segment 1.

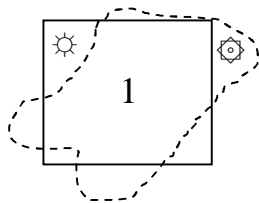


Figure 1. Scenarios for the Missed Structure Check

This principle guarantees that each DU has a probability of being sampled and no DU has a probability of being selected more than once.

4.2 Concerns with Large Numbers of Missed Structures

One of the concerns with any missed structure procedure is finding pockets of large numbers of missed units, where the samplers are faced with a dilemma. If all the missed units found were included into the sample, it would not only increase field burden, but also cause higher clustering of the sample. If instead, a subsample of units was brought into the sample, weighting adjustments would be necessary, creating large design effects.

This phenomenon is more likely to occur when address lists are used for sampling, due to the coverage problems discussed above. The challenge here is to achieve efficiency of the sample while maintaining a stable sample size for field work. As a possible solution to this problem, we introduce the Waksberg approach and a proposed enhancement.

4.3 Waksberg Approach and Proposed Enhancement

In this section we briefly compare two approaches for the missed structure procedure. The half-open interval

(HOI) frame-linking procedure (Kish, 1965) is based on a subsample of addresses across all the sampled segments. The other method, which we call the “Waksberg approach,” was developed by Joe Waksberg in the early 1970’s and has been used at Westat since that time. This approach involves selecting a subsample of the segments and then conducting a thorough listing check in the subsample.

Previous work by Iannacchione, Staab, and Redden (2003) recommends the HOI procedure to check for units missed on the purchased list. This method is not preferable for national-level surveys mainly because:

- It entails taking a simple random sample of intervals across all sampled segments leading to high field labor costs; and
- There is high home office cost in preparing maps for each segment with routes highlighted for each carrier route. This may not be realistic in terms of a continuous survey, though it might be feasible for a one-time survey.

In contrast, the Waksberg approach sends field staff to only a small number of segments in each PSU. Particularly, this approach, with the proposed enhancement below, offers a possible solution to the problem of encountering large numbers of missed DUs during the missed structure procedure.

The proposed enhancement to the Waksberg approach involves several steps for sample selection:

- Use Census data or another source of information to obtain an indication of growth for sampled segments;
- Categorize the segments based on their growth;
- Select a sample of the segments for the missed structure check, using higher probabilities in higher growth categories; and
- Within the segment, only include a subsample of the missed units found. Ideally, the selection rate should be the inverse of the sampling probability used in step 3. Generally, apply lower subsampling rates for including missed units in higher growth segments.

This proposed enhancement can solve the problem to some extent. If a new apartment complex with 300 dwelling units is found through missed structure check, only including a fraction of the units would still increase the sample size by a large number. Further

research is needed to arrive at a feasible approach for sampling and retaining missed structures.

5. Conclusions and Future Research

Using address lists in place of traditional listing could reduce the cost of a survey to a great extent. As described above, this new approach also brings challenges and complications, mostly associated with the coverage of the purchased lists, the geocoding process, and the missed structure procedure. Our research suggests that when considering purchasing address lists from a vendor, the samplers should 1) choose a vendor licensed for CDS updates in the desired ZIP Code; 2) ask when each ZIP Code was last updated; and 3) before purchasing, review counts of addresses by ZIP Code and compare them to Census information.

At this time it is still difficult to make a general conclusion as to whether these address lists are adequate enough as substitutes for on-site enumeration for area sampling. Samplers should examine the specific requirements of different studies and make case-by-case decisions. It is very important to ensure a level of adequate coverage before we can implement the use of these lists in area sampling throughout the nation. What is promising is that as more areas are converted to city-style addresses for purposes of 911

location, the coverage of these address lists in rural areas is more likely to be improved in the future.

Our future work will focus more on the practical challenges of using address lists. Particularly, we will look for sources to augment the purchased lists, find methods to handle nongeocodable addresses, and evaluate, with the goal of further improving, the proposed enhancement to the Waksberg approach in the field.

6. Acknowledgements

The authors want to acknowledge Michael Giangrande and Tom Hankins whose combined technical expertise was indispensable in this research.

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, L. (1965). *Survey Sampling, Inc.*, New York: John Wiley & Sons.
- Staab, J.M., Iannacchione, V.G. (2003). Evaluating the Use of Residential Mailing Addresses in a National Household Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 4028-4033.