Address Based Sampling Frames for Beginners

Sylvia Dohrmann[1], Trent D. Buskirk[2], Ashley Hyon[2], Jill Montaquila[1]
[1]Westat, 1600 Research Blvd., Rockville MD, 20850
[2]Marketing Systems Group, 755 Business Center Drive, Suite 200, Horsham PA 19044

Abstract

Address lists originating from the United States Postal Service (USPS) have been used as sampling frames in various survey modes for more than a decade. Surveys that might otherwise have opted for random digit dial (RDD) sampling with telephone interviewing are now using these address lists as sampling frames. Multi-stage area probability sample surveys with in-person data collection are using these address lists in place of (or to enhance) the traditional address listing process. The lists are only available to survey researchers through third-party vendors who license the lists from the USPS. Some of these vendors append auxiliary information to the lists which can be used in the sampling process in addition to the delivery information from the USPS. Together, these data have become known as address-based sampling (ABS) frames.

This paper gives an overview of ABS frames including the USPS- and vendor-provided information available and its utility for survey research.

Key Words: ABS, USPS, CDS, RDD sampling, Area sampling

1. Introduction

Historically, national household registers have not been available to survey researchers for household surveys in the United States. Therefore, organizations undertaking national household surveys have had to rely on either random digit dial (RDD) sampling of telephone numbers or multistage area sampling. The latter is expensive and requires a considerable amount of time to arrive at the final sample of households. The former is increasingly subject to undercoverage (particularly if a landline-only RDD frame is used), or complexity (if cell phones are included in a dual-frame design) as well as low (and declining) response rates.

When survey researchers began investigating the use of address lists made available from the United States Postal Service (USPS), there was some hope that these could be considered a proxy for a household register since in addition to the postal addresses, third-party vendors supplied auxiliary information about the household and the area in which it resides. More than a decade later, however, it is clear that cannot be considered a panacea. Rather, to use these combined data as sampling frames, known as Address Based Sampling (ABS) frames, researchers must understand their contents and scope.

The following sections provide an overview of ABS frames by starting with the information taken from the USPS (Section 2) and then discussing information added by the ABS frame vendor (Section 3). Section 4 describes various ways these frames have been used in survey research, and Section 5 provides a summary. Additionally, in the Appendix, some common postal service acronyms and abbreviations are provided.

2. Information Available from the USPS

ABS frames begin with addresses from the United States Postal Service (USPS) Computerized Delivery Sequence file (CDS):

> "The CDS file is a snapshot of the Address Management System (AMS) … and contains the official United States Postal Service® (USPS®) record of mailing addresses. The primary purpose of the database is to provide delivery and distribution information necessary to support processing of mail for delivery to the citizens and businesses of the United States." (USPS, 2013).

Table 1 contains counts of addresses on the CDS file, by type of address, as of July 2014. The AMS includes all postal customers at approximately 151 million U.S. addresses, both residential and business, in the 50 states and the District of Columbia. The addresses on these lists include those along mail carrier routes, and all Post Office (PO) boxes located in official USPS offices or establishments acting as private mail box operators (such as "UPS Stores"). Addresses not included on these lists are military and government addresses, and approximately 200,000 residential addresses known as simplified addresses[1].

Table 1. Number of Addresses on the CDS File as of July 2014.

	All Addresses		*Non-PO Box*		*PO Box Addresses*	
	Number	*%*	*Number*	*%*	*Number*	*%*
Total	152,407,466		133,152,312		19,255,154	
Residential Only	139,505,996	91.54%	123,884,116	93.04%	15,621,880	81.13%
Primarily Residential (some Business)	157,248	0.10%	157,222	0.12%	26	0.00%
Primarily Business (some Residential)	21,709	0.01%	21,637	0.02%	72	0.00%
Business Only	12,722,513	8.35%	9,089,337	6.83%	3,633,176	18.87%

The CDS file contains one record per address, and includes additional delivery information for each record. The address information is organized with each address component stored in a separate data field. Table 2 shows examples of different types of addresses.

[1] Simplified addresses are ones for which the only delivery information held by the USPS is the city, state, ZIP Code, and the number of postal customers receiving mail at that ZIP Code. Mail sent to customers via a simplified address need only include their name, city, state, and ZIP Code to reach the intended customer.

Table 2. Types of Addresses Available on the CDS File.

House number	*Pre direction*	*Street name*	*Street suffix*	*Post direction*	*Secondary unit descriptor*	*Apt number*	*City*	*State*	*ZIP*	*ZIP+4*
101	N	Main	St	S	APT	1	Anywhere	MA	12345	6789
900		PO Box					Anywhere	MA	12345	6789
105		RR 15					Anytown	MA	67891	1234
136A		HC 68					Anytown	MA	67891	5678

The first address in Table 2 is an example of a typical city-style address for a unit in a multi-unit apartment building; one would address a letter to this unit using the address fields in the order they are presented in the table. For the other addresses in the table, the USPS is using the House Number field to hold a box number for the post office, rural route, and highway contract boxes. These would be correctly addressed with the Street Name first, then the House Number (and "BOX" in between the two fields for the latter two addresses).

For each address the CDS file contains the delivery route information and information about the delivery point. Each delivery route is assigned its own ID, known as the Carrier Route ID, the first digit of the ID provides further information:

- 'B' indicates that all the addresses with this ID are PO Box addresses;
- 'H' indicates that the addresses are delivered as part of a highway contract route, a route that is contracted out to independent carriers (such as the last address in Table 2 above);
- 'R' indicates that the address is part of a mail delivery route in a rural area. These include, but are not limited to, addresses with rural route box numbers (such as the third address in Table 2 above).
- 'C' indicates that the address is part of a city route (such as the first address in Table 2).
- 'G' indicates that the address is delivered via general delivery. Addresses in this type of route are intended as a temporary means of mail delivery for transients and customers not permanently located, and customers who want PO Box service when boxes are not available.

Within each Carrier Route, addresses are given a Delivery Sequence Number which indicates the position of each address along the Carrier Route, or the order in which the mail is delivered. Within a 5-digit ZIP Code, the Carrier Route ID and Delivery Sequence Number together uniquely identify an address as illustrated in Table 3 below.

Table 3. Illustration of Addresses by Carrier Route ID and Delivery Sequence Number within a ZIP Code.

House number	*Pre direction*	*Street name*	*Street suffix*	*Post direction*	*ZIP*	*ZIP+4*	*Carrier route ID*	*Delivery sequence number*
400	N	Main	St	S	12345	6789	C011	1
402	N	Main	St	S	12345	6789	C011	2
404	N	Main	St	S	12345	1234	C011	3
406	N	Main	St	S	12345	1234	C011	4

The CDS also includes information about each delivery point including seasonal and vacancy codes, and delivery point usage and type codes. If the delivery point is a PO Box, there are variables indicating whether it is the only way the postal customer receives mail and whether they are eligible for mail delivery at a non-PO Box address. There is also a code indicating whether the delivery point serves multiple businesses or housing units and, if so, how many. These are each discussed in turn below.

Seasonal Code – The Seasonal Code indicates whether the delivery point has seasonal delivery ('Y'), is used for seasonal educational use ('E'), such as a student housing on or near a college campus, or neither ('N'). Note that the USPS does not have delivery point information (including individual addresses as well as all other fields described in this section) for all student housing units in the U.S. since many colleges and university campuses have their own post offices that deliver mail to students living on campus.

Vacant Code – The Vacant Code indicates whether the delivery point has been unoccupied for over 90 days ('Y') or not ('N').

Delivery Point Usage Code – The usage code indicates whether the delivery point is purely residential ('A'), purely business ('B'), primary residential with some business ('C'), primary business with some residential ('D'), or general delivery ('G'). As shown in Table 1, more than 90 percent of the addresses on the CDS file have a usage code of 'A', versus a little over eight percent that have a usage code of 'B.'

Delivery Point Type Code – The type code indicates how mail is delivered at the address and the type of service. Codes for city, rural, and contract delivery service routes include the following:
'A' - Curbline – mail receptacle is located at the curb
'B' - Centralized Box Unit (CBU) – mail receptacle is located within a cluster box
'C' - Central – mail receptacle is located within a centralized unit
'D' - Other – mail receptacle does not fit into one of the above categories

PO Box routes have their own set of Delivery Point Type Codes which designate whether the box is within a USPS branch or another location, as well as whether the box is assigned to a college or university, used for payments to an institution, rebate, coupon, or other operations.

OWGM (PO Box records only) – The OWGM (an abbreviation for "Only Way to Get Mail") flag indicates whether the PO Box is the only way the customer receives mail

('Y') or not ('N'). High concentrations of PO Box addresses with this code set to 'Y' will appear in ZIP Codes where home mail delivery is not available.

PO Box Throwback Indicator (PO Box records only) **–**This will be set to 'T' if the box holders have a street address that is eligible for mail delivery, but the residents or business owners choose to have their mail delivered to a PO Box address instead (i.e, "thrown back" to the post office). In this event, the postal customer is essentially on the CDS file twice, once at the street address and again at the PO Box address.

Delivery Point Drop Indicator – A ***drop point*** is an address that serves multiple businesses or housing units. Drop point addresses have a value of 'Y' for this indicator; if the address is not a drop point, this indicator will have a value of 'N'. Values of 'C' indicate that the address is for a private mail box operator (known as a commercial mail receiving agency or CMRA). The only information available for the individual housing units served at this delivery point (known as ***drop units)*** is the number served by that address. Less than 1% of all addresses on the CDS file are drop point addresses, but this type of addressing tends to be clustered by type of address (see Table 4 below) and geography.

Delivery Point Business Family Served Count - The number of potential drop units for a drop point. If the Delivery Point Drop Indicator equals 'Y', the Business Family Served Count field contains the number of businesses or families served at that drop site. If the Delivery Point Drop Indicator equals 'N', this field will be zero. If the Delivery Point Drop Indicator equals 'C', this field will be greater than zero.

Note that the AMS is not updated in real time. The postal carriers are responsible for conveying any corrections or updates to the fields described above and the method of the subsequent update is not automatic. As such, the CDS file may not accurately reflect the delivery information at the point in time the AMS "snapshot" was taken. So, for example, a record with a value of 'Y' for the Vacant Code may actually be occupied at the time the CDS file is created. Additionally, there is variation among vendors as to how often they receive updates from the USPS, as discussed in the next section.

Table 4. Number and Distribution of Drop Point Addresses on the CDS File as of July 2014.

	Drop Point Addresses	
	Number	*%*
Total	830,042	0.62%
Residential Only	715,410	0.58%
Primarily Residential (some Business)	4,112	2.62%
Primarily Business (some Residential)	2,456	11.35%
Business Only	108,064	1.19%

3. Vendor Processing of the Lists to form ABS Frames

The address records and delivery point information described in the previous section are not directly available from the USPS. Rather, they are only available through vendors holding a license to the CDS file. To qualify for a license, the vendor must already own a base file of addresses within each ZIP Code for which they wish to qualify. This license allows the vendor to receive updates and corrections to that base file on a bimonthly or weekly basis. These vendors then create an ABS frame by combining these data with other address data (Section 3.1) and/or with demographic and geographic information (Section 3.2) available from commercial sources. As a result, the final ABS frame may include data from multiple time periods (Section 3.3).

3.1 Additional Addresses

The ABS frame vendor may be able to add additional addresses not included in the CDS file. For example, the vendor may be able to identify the individual addresses in carrier routes containing only simplified addresses or individual unit numbers for housing units contained in drop points.

Additionally, the vendor may have access to the CDS No Stat file containing additional USPS addresses. The No Stat file compliments the CDS file and contains roughly 10 million additional inactive addresses. Addresses contained on the No Stat file are currently unable to receive mail for a variety of reasons. Types of addresses contained on the No Stat file include vacant housing units in rural areas, demolished or damaged structures, and new or planned construction. No Stat addresses are not mailable, but may be used in conjunction with CDS addresses for compiling a full list of available housing units.

3.2 Demographic and Geographic Information

The ABS frame vendor may also have access to one or more commercial data sources containing household level demographics, telephone numbers, names of residents, and other variables. Examples of such data include a flag indicating whether the surname of householder is Hispanic or Asian, the highest level of education achieved by any household member, ethnicity of the householder, age of the householder, household income, and whether the home is rented or owned, among others.

Buskirk and Malarek (2014) discuss record append rates for a battery of demographic variables. Vendors may also have access to behavioral and consumer related activity data that can also be appended to the sampled address. Note that in any case, the appended information is generally being matched to an address and represents attributes of at least one person who was linked to that address. In the case of multiple adults in the household, it is possible that information appended to an address from one vendor may not match that of another since the reference persons for a particular address may vary be vendor (Buskirk and Malarek, 2014).

The only “geographic” references on the USPS lists are the addresses themselves, including the ZIP and ZIP+4 codes associated with the address. Some vendors will append approximate latitude and longitude coordinates based on the address using geocoding. (See Section 3 of Dohrmann, et al. 2013 for more information on geocoding.) Based on these coordinates, U.S. Census Bureau geographic codes (e.g., Census tract, block group and block) can also be appended and with them a multitude of area-level

demographic characteristics available from the Census Bureau's publicly available data files.

3.3 Timing

Depending on the vendor's license with the USPS and the number and type of data sources appended to the available addresses, the ABS frame may include data captured at varying points in time across one or more years. For example, the USPS address update may be from the last week of July (with individual addresses fields updated at indeterminate points), but the name and phone number may come from a database last updated in March, the demographic data from commercial databases updated six months prior to then, and decennial census data captured at the start of the decade. Knowing the possible lag times and update schedules for appended information could help determine when to order an ABS sample and when to have information appended to it, especially if the sample field period is several months out from sample procurement. This issue is especially important if the appended information is to be used by field staff to prioritize or manage cases.

4. Utility of the Lists for Survey Research

For decades, organizations conducting telephone surveys have used address lists for operational reasons. Reverse matching processes have been used to append addresses to sampled telephone numbers, and these appended addresses have been used, for example, as the basis of confirmation of location (e.g., city, state, ZIP Code, cross-street, etc.) and to identify the time zone associated with the telephone number for the purpose of restricting hours during which calls to the number may be attempted.

Survey research organizations are increasingly turning to ABS as an alternative to RDD or in place of (or to enhance) traditional listing to construct housing unit frames in multi-stage area probability samples. While the aim of this paper is to provide information about the ABS frames themselves, several other articles have discussed methods and considerations for surveys that use the ABS frames. (See Iannacchione 2011 for a review.) Here, we briefly discuss various uses of the ABS frames.

As telephone survey response rates and landline coverage rates continue to decline (Curtin, Presser, and Singer 2005; Blumberg and Luke 2014), survey researchers are increasingly turning to ABS as an alternative to RDD. ABS frames are also well-suited—even more so than RDD frames in general these days, due to telephone number portability and lack of geographic specificity in numbers assigned to cellular telephone—for selecting samples for geographically restricted surveys (e.g., surveys of residents of a particular county). Mail may be the sole mode of data collection, or telephone the primary mode for addresses with a telephone number match.

ABS frames are also attractive for use with in-person area surveys. The addresses can be used to assist with or replace the traditional "listing" process. Traditional listing requires that field staff canvass the sampled areas in-person months before data collection begins in order to list the addresses of all the housing units observed in those areas and to allow adequate time for housing unit sample selection. The operation can be both costly and time consuming. However, ABS frames could be the starting point of this canvass (so that field staff simply add any housing units missing from the frame), or replace the canvass if there is some sort of frame quality control procedure is conducted to ensure that all housing units have a chance of selection. (Kalton, Kali, and Sigman, 2014).

In-person area surveys often rely on decennial census data to determine the geographic boundaries and measure of size, usually the number of housing units, of the secondary sampling units. Yet, late in the decade housing or demographic data from the last decennial census are likely to be inaccurate in local areas with considerable growth or demographic shifts since the census taking and intercensal estimates are not available at the required level. Address counts from an ABS frame can be used to estimate the measure of size of secondary sampling units late in the decade with reasonable effectiveness and at minimal cost. (Dohrmann, Li, and Mohadjer, 2011).

The cost of screening for an in-person area survey can be considerable if the target population is one possessing a rare characteristic. However, with an ABS frame screening may be conducted via mail, and then in-person data collection can be pursued for those screener respondents possessing the characteristic of interest.

The previous section discussed the variables vendors are able to append to addresses from third-party sources. These variables may be used for targeting subgroups of the population, although it is important to consider the effects of undercoverage due to misclassification (Valliant et al., 2014). They may also be considered for use in stratification of the sample (Roth, Han, and Montaquila 2013; Valliant et al. 2014); in that case, a two-phase sampling approach would be used, and the appended variables would be used for stratification at the second phase of selection. A third use of the appended variables is to tailor survey materials based on characteristics associated with the address, such as sending Spanish language materials to addresses associated with Hispanic surnames (Brick et al. 2012).

5. Summary

While ABS frames are not a panacea, they do offer an abundance of information to the survey researcher. The files originate with the approximate 151 million residential and business addresses in the United States as recorded by the United States Postal Service (USPS) on its Computerized Delivery Sequence (CDS) file. With these addresses comes delivery information which can inform the survey researcher as to whether the address is one to be included in any particular sample frame. It is important to remember, however, that how a household or business is represented on the CDS file depends entirely on how they receive their mail.

Commercial vendors who have access to these addresses create the ABS frames by appending additional information including address detail not provided by the USPS, demographic and geographic information about the household, as well as demographic information about the area in which the address resides. The appended information may originate from several sources, each updated on separate schedules. As such, it is important for the survey researcher to communicate with their ABS frame vendor to understand the origin of the information on which their design may rely, be it a mail survey, a multi-mode design including mail and RDD methods, or an in-person area survey.

References

Blumberg, S. J., & Luke, J. V. (2014). Wireless substitution: early release of estimates from the National Health Interview Survey, July-December 2013. *National Center for Health Statistics*. Available at http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201407.pdf [Last accessed September 15, 2014].

Brick, J. M., Montaquila, J. M., Han, D., & Williams, D. (2012). Improving response rates for Spanish speakers in two-phase mail surveys. *Public opinion quarterly*, 76(4), 721-732.

Buskirk, TD and Malarek, D. (2014). From Flagging a Sample to Framing It: Exploring Vendor Data Appended to Address-Based Samples. Paper presented at the 2014 Joint Statistical Meetings, Boston, MA. Paper forthcoming.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. *Public opinion quarterly*, *69*(1), 87-98.

Dohrmann, S., Kalton, G., Montaquila, J., Good, C., and Berlin, M. (2012). Using Address Based Sampling Frames in Lieu of Traditional Listing: A New Approach. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 3729-3741.

Dohrmann, S., Li, L., and Mohadjer, L. (2011). Updating the Measures of Size of Local Areas Late in the Decade Using USPS Address Lists. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2891-2901.

Iannacchione, V.G. (2011). The changing role of address-based sampling in survey research. *Public Opinion Quarterly*, *75*(3), 556-575.

Kalton, G., Kali, J., and Sigman, R. (2014) Handling Frame Problems when Address-based Sampling is Used for In-person Household Surveys. *Journal of Survey Statistics and Methodology*. doi: 10.1093/jssam/smu013Graham.

Roth, S. B., Han, D., & Montaquila, J. M. (2013). The ABS Frame: Quality and Considerations. *Survey Practice*, *6*(4).

United States Postal Service (USPS) (2013). *CDS March 2013 User Guide*. Available at *https://ribbs.usps.gov/cds/documents/tech_guides.*

Valliant, R., Hubbard, F., Lee, S. and Chang, C. (2014) Efficient use of commercial lists in U.S. Household Sampling. *Journal of Survey Statistics and Methodology* 2: 182-209.

Appendix: Table of Some Common Acronyms or Abbreviations Related to Address Based Sampling.

Abbreviation/Acronym	*Definition/Description*
USPS	United States Postal Service
ABS	Address Based Sampling
CDS	Computerized Delivery Sequence (used in reference to the CDS file)
AMS	Address Management System
CBU	Centralized Box Unit (a type of delivery point)
OWGM	Only Way to Get Mail (an attribute of PO Boxes)
CMRA	Commercial Mail Receiving Agency
PO BOX	Post Office Box