Marketing Systems Group

GENESYS SAMPLING

Recent Developments in Address-Based Sampling

Overview

Increasingly, survey and market researchers are reconsidering address-based sampling (ABS) methodologies to reach the general public for data collection and related commercial applications. Essentially, there are four main factors for this change:

- Evolving coverage problems associated with telephone-based sampling methods, particularly those relying on random digit dialing (RDD);
- Technical and operational complexities associated with dual-frame methods;
- Eroding rates of response to single modes of contact along with the increasing costs of remedial measures to counter nonresponse; and on the other hand
- Recent improvements in the databases of household addresses available to researchers.

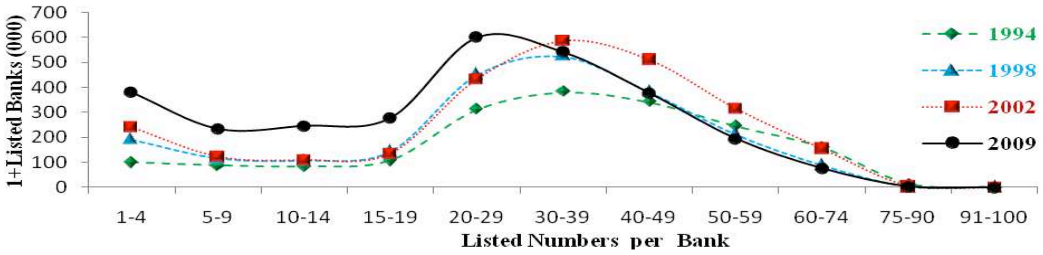
This note provides an assessment of the above factors, evaluates pros and cons of ABS as an alternative, and discusses specific enhancements that can establish this emerging methodology as a practical solution for many survey and market research applications. In particular, such enhancements include amelioration of some of the known problems associated with ABS frames through augmentations with geodemographic and other supplementary data items. While enabling researchers to develop more efficient sampling designs, such enhancements broaden their analytical possibilities by providing an expanded set of covariates for nonresponse bias analysis and weighting, as well as hypothesis testing and statistical modeling tasks.

Coverage Problems for Telephone Surveys

For the past decade and a half, a large portion of telephone surveys have been based on the list-assisted RDD methodology where samples of telephone number are generated within the 100-series telephone banks that contain at least one listed number (1+listed banks). During the intervening years, however, this method has overlooked the many fundamental changes in the U.S. telephony and relied on a convenient assumption that elimination of 0-listed telephone banks – those with no listed numbers – from the sampling frame amounts to exclusion of a small percentage of landline households, hence resulting in an ignorable coverage bias. However, recent investigations suggest that the extent of undercoverage for landline RDD frames that only include the 1+listed banks has been growing (Fahimi et al. 2009). While the exact magnitude of this undercoverage is difficult to estimate due to definitional and operational challenges, there have been several studies supporting the stated concern (Weiss et al. 2009 and Baron et al. 2010).

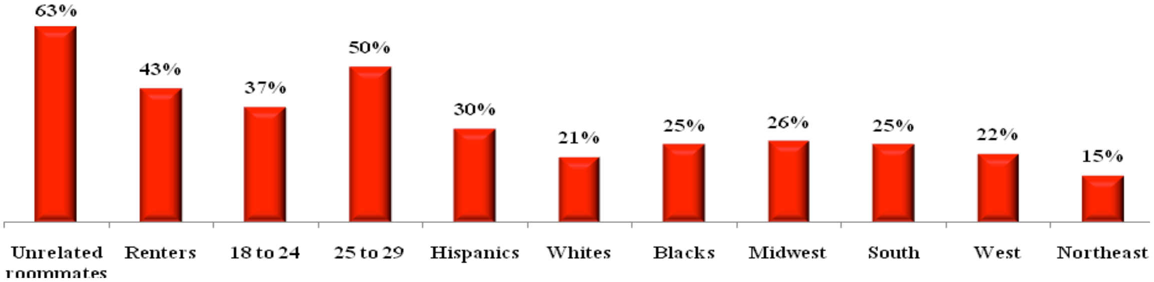
There are several interrelated reasons germane to the above degradation. For instance, Tucker and Lepkowski (2008) report that the proportion of residential numbers in the 1+listed banks has been on a decline. The following figure shows how the residential distribution of such banks has indeed changed, both with respect to the location and scale parameters, across the years. While the residential density of 1+listed banks was much flatter in 1994 with an average of about 35 listings per bank, this distribution has become more peaked in recent years with a 2009 average of about 26 listings per bank and a sharp increase in the number of banks with lower residential density. Listed banks with low residential density are particularly vulnerable to losing their listed status because in many instances it takes only a handful of listed numbers becoming unlisted for their corresponding banks to get demoted to a 0-listed status and removed from the traditional RDD frame.

Distribution of 1+Listed 100-Series Banks by Residential Density



Another source of undercoverage for landline RDD samples has to do with the growing number of households that are abandoning their landline services and come to rely entirely on cellphones for their telecommunication needs. As depicted in the following chart, an alarming percentage of the US adults are becoming what's called cell-only. Combined with the coverage problems associated with the traditional RDD frames, landline RDD samples can fail to cover a non-ignorable proportion of the US households – a problem that becomes even more pronounced when surveys target special sub-populations such as younger adults. It is worth mentioning even with dual-frame alternatives, whereby both landline and cellular RDD samples are combined, the problem of undercoverage persists. Moreover, there are many operational as well as technical complications that further reduce the appeal of such interim solutions. These include problems associated with contacting individual on cell phones and estimation issues related to the unavailability of reliable estimates for cell-only households at smaller levels of geography.

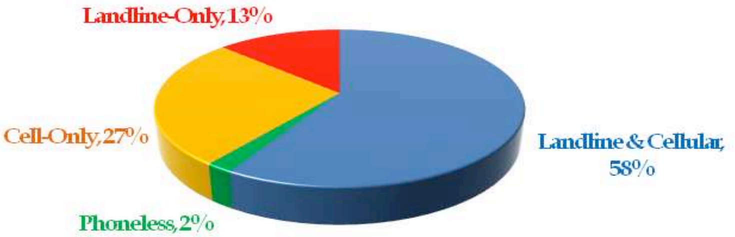
Percent cell-only adults by geodemographic indicators (Source CDC 2010)



Technical and Operational Complexities of Dual-Frame Methods

In light of the coverage problems associated with the traditional method of RDD, virtually all creditable sample surveys that employ this methodology now rely on dual-frame techniques that include both landline and cellular telephone numbers. Given that there are no reliable estimates for the number of cell-only households at various levels of geography, however, this alternative is subject to inconsistencies for sample allocation and weighting applications. That is, there are no consistent guidelines for determining the proper mixture of landline and cellular telephone numbers in a dual-frame RDD sample – an ambiguity that has further implications for computing survey weights. The resulting arbitrariness is further compounded by the fact that cellular telephone numbers can reach both cell-only households as well as those relying on landline services as well. While somewhat dated, the following table provides an approximated distribution of the US households based on their telephone status.

Telephone status of the US households (Source CDC 2010)



It should be noted that in addition to sampling and estimation complexities, there are operational inefficiencies due to the difficulty of geographic targeting of cellular RDD samples. This becomes of particular concern for smaller geographic locations, since many cellular telephone subscribers reside outside of the area corresponding to the billing location of their services. The following link (http://aapor.org/Cell_Phone_Task_Force.htm) provides access to a comprehensive document developed by the Cell Phone Taskforce of the AAPOR, covering the above issues in more details.

Eroding Rates of Response to Single Modes of Contact

Biener et al. (2004) and Curtin et al. (2005) point out that the rate of response to telephone surveys has been on a decline, which is also the case for well-funded government surveys such as the Behavioral Risk Factor Surveillance System (BRFSS), Fahimi et al. (2007a). Given that nonresponse is highly differential and can vary significantly across different demographic subgroups, it is of a great concern when over half of the sample households opt not to respond to a survey. Even when sophisticated nonresponse adjustment procedures are employed to reduce the incurred bias, it might be farfetched to assume such remedial procedures can reduce nonresponse bias to a tolerable and measurable level. Furthermore, reducing nonresponse bias via weighting is always exercised at the expense of reduced precision of survey estimates, since weighting inflates variance of survey estimates (Fahimi et al., 2007b). The inflation due to weighting can be approximated by the following index, in which W_i represents the final weight of the i^th^ respondent and $\bar{W}$ denote the average weigh for all survey respondents.

$$\delta = 1 + \frac{\sum_i \frac{\left(W_i - \bar{W}\right)^2}{n-1}}{\bar{W}^2}$$

Beyond statistical techniques, many researchers resort to other tactics to improve response rates to surveys. As reported by Fahimi et al. (2004) the offer of incentives can significantly increase response rates, however, even an increase of 10 percentage points can still leave a survey with an overall response rate well below 50%. Moreover, marginal gains in response rate are often achieved at a high cost, as practical nonresponse conversion strategies are labor intensive and require exceedingly larger amounts of incentives to be effective. Coupled with the non-monetary cost due to loss of precision, the overall cost of dealing with nonresponse can be prohibitive.

Improvements in Databases of Household Addresses

Recent advances in database technologies along with improvements in coverage of household addresses have provided a promising alternative for surveys and other commercial applications that require contacts with representative samples of households. Obviously, each household has an address and virtually all households receive mail from the U.S. Postal Service (USPS). The Computerized Delivery Sequence File (CDSF) of the USPS is a database that contains all delivery points, with the exception of general delivery where carrier route or P.O. Box delivery is not available and mail is held at a main post office for claim by recipients.

With more than 135 million delivery points on file, the latest generation of the CDSF is the most complete address database available. As such, it is safe to assume that if an address cannot be matched against the CDSF it is most likely an undeliverable address. What is more, by providing validation services for both correctness and completeness of addresses the CDFS can significantly enhance the address hygiene. Consequently, this system helps reduce the number of undeliverable-as-addressed mailings, increase the speed of delivery, and reduce cost. Also, with daily feedback from tens of thousands of letter carriers the database is updated on a nearly continuous basis. The following table provides counts for the main groups of delivery types.

Distribution of the CDSF delivery type points as of March 2011

Delivery Point Type	Count
City Style/Rural Routes	114,990,949
Traditional PO Box	14,085,124
OWGM PO Box	1,435,992
Seasonal	857,958
Educational	99,375
Vacant	3,916,261
Throwback	285,309
Drop Points	779,399
Augmented addresses (by MSG)	126,368
Total	**136,576,735**

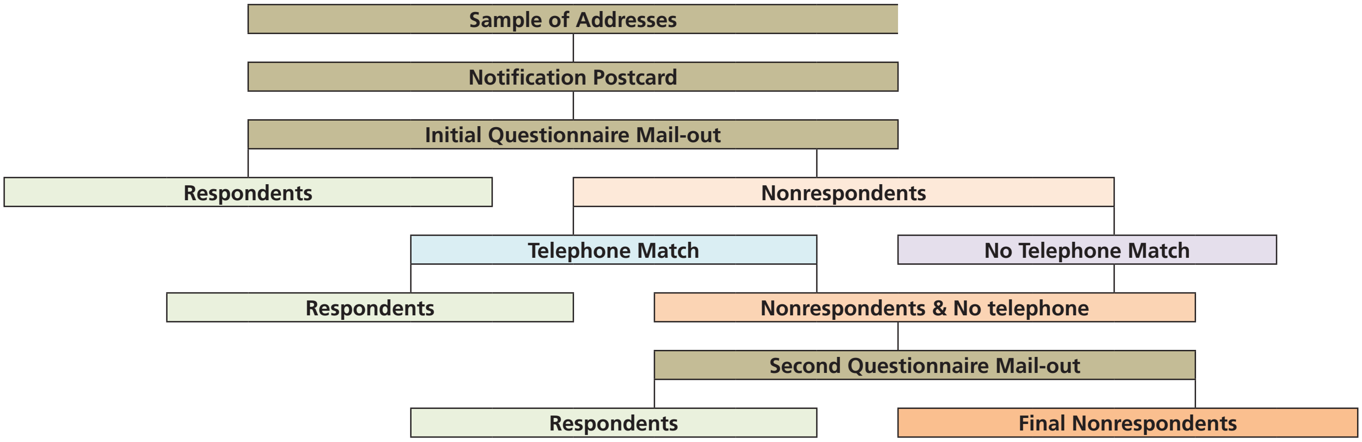
Using CDSF for Sampling Purposes

Given the evolving problems associated with telephone-based methods of sampling and data collection, many researchers are now considering the use of CDSF for sampling purposes. Moreover, the growing problem of nonresponse – which is not unique to any individual mode of survey administration or country (de Leeuw & de Heer 2002) – suggests that more innovative approaches will be necessary to improve survey participation. These are among the reasons why multi-mode methods for data collection are gaining increasing popularity among survey and market researchers. It is in this context that ABS designs provide a versatile framework for creative methods of survey administration that employ multi-mode alternatives for data collection.

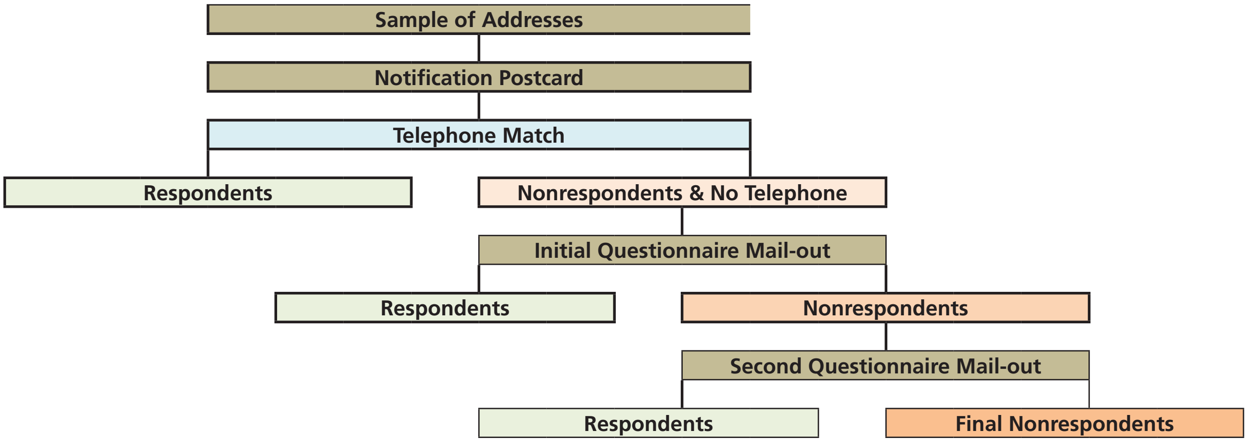
Considering that through reverse-matching the telephone numbers for many addresses can be obtained, different strategies for a multi-mode survey administration can be developed to accommodate the timing, budgetary, and response rate needs of a survey. One such strategy could start with the selection of a CDSF-based probability sample of households in the geographic domain of interest. This sample may be selected across the entire domain, or clustered in an area probability fashion if in-person attempts are contemplated as part of the design. Initial contacts can be by phone and/or mail and can include attempts for survey administration at the same time. Alternatively, this first contact can serve as a recruitment effort to invite potential respondents to participate in the survey using one of the many options, including web, dial-in numbers for live interviewing, an IVR system, or even a regular mail.

Once the nexus of contact modes has been developed for each respondent, further contacts and reminders for survey completion can take place in any order or combination of modes that meets the project needs and is best suited for the given respondent. The following schematics display two options for multi-node survey administration under an ABS design.

First possible multi-mode protocol for survey administration under an ABS design



Second possible multi-mode protocol for survey administration under an ABS design



Cognizant of the potential implications of combining different modes of data collection, the emerging conclusions from many studies seem to suggest that different contact modalities can often be combined effectively to boost response rates (Gary 2003). In comparison to an RDD-only approach, in particular, an address-based design using multiple modes for data collection can provide response rate improvements, cost savings, as well as better coverage for households that are completely uncovered by landlines (Link 2006). In comparisons with in-person and mail-only modes of data collection, needless to say, the former is too costly to be practical for many applications while the latter (with notoriously low rates of response) requires expensive nonresponse follow-up efforts to produce creditable data (Groves 2005). What seems critical, however, is for researchers to minimize differences between survey instruments associated with each mode. Moreover, effective weight adjustment techniques might be needed post data collection to account for the observed differences in the profile of respondents to each mode.

Potential Issues When Using the CDSF for Sampling Purposes

Fundamentally, the CDSF is a database for mail delivery and not a sampling frame. As such, the raw CDSF needs refinements in several aspects before it can qualify as a credible tool for survey sampling. First and foremost, this database does not include geodemographic indicators for effective sample stratification – an issue of critical importance for complex designs. Moreover, certain households have a higher likelihood of not being included as a delivery point on the CDSF. Staab and Iannacchione (2003) estimate that approximately 97% of all US households have locatable mailing addresses, however, this prevalence may diminish with population density and in areas where home delivery of mail is not readily available. Dohrmann and Mohadjer (2006) report that when comparing lists of on-site enumerated addresses to the CDSF generated listings of households for the same geography, in rural areas the rate of mismatches can be over 23%. However, as rural area addresses go through the 9-1-1 address conversion and acquire a city-style format the coverage of CDSF-based lists in rural areas is rapidly improving. As will be discussed later, in 2004 more than 7% of all addresses were undeliverable (simplified) yet today this percentage has dropped to less than a third of a percent.

Beyond coverage issues, when CDSF generated samples are used in surveys with a multi-mode approach for data collection one has to be prepared to address concerns about mode effects. While somewhat academic in nature, concerns have been raised about systematic differences that can be observed when collecting similar data using different modes (Dillman 1996). On the one hand, several studies have shown a greater likelihood for respondents to give socially desirable responses to sensitive questions in interviewer-administered surveys than in self-administered surveys (Aquilino 1994). On the other hand, the rate of missing data is often significantly higher in self-administered (mail or web) surveys as compared to interviewer-administered (telephone or in-person) surveys (Biemer et al., 2003). While roots of differences in data quality and response rates between various modes of data collection deserve further investigations, some solace may result when surveys are administered without confining data collection to any single mode. Arguably, certain shortfalls of one method can be mitigated when other methods of data collection are made available to the respondents as well. Ultimately, however, it might be impossible to untangle the immeasurable interactions between the mode, the interviewer, the respondent, and the survey content (Voogt & Saris 2005).

Lastly, there are survey situations where data are to be collected in-person. In such cases reliance on delivery information may not be adequate, as the exact location of all sample dwellings must be known. This is of particular complication when a P.O. Box is the only means of delivery for a household. On the other hand, there are households that have both residential addresses as well as P.O. Boxes. Ignoring this problem leads to frame multiplicity, since such households will have multiple chances of selection. These are additional refinements that may be added to the CDSF before it can evolve from a delivery database into a sampling frame.

Available Enhancements for the CDSF

As mentioned above, the CDSF can be used to select probability-based samples of addresses in finely defined areas down to ZIP+4. Since there is not a one-to-one correspondence between the USPS and Census geographic definitions, unfortunately, this creates a problem as in most surveys the Census geographic definitions are used due to the availability of household and person counts for sampling and weighting applications. This gap, however, can be bridged by geo-coding each address to a unique Census block. Subsequently, one can append many ancillary data items to each address, including those available from the Census and commercial sources. This is the crossroad where basic list suppliers – those that simply offer raw extracts from the USPS – are differentiated from Marketing Systems Group that provides enhanced version of the CDSF by supplementing this basic delivery database with:

- Detailed geodemographic information;
- Name and telephone number;
- Simplified address resolutions;
- Indicators for areas with potential coverage problem; and
- Frame multiplicity reduction.

Detailed Geodemographic Information

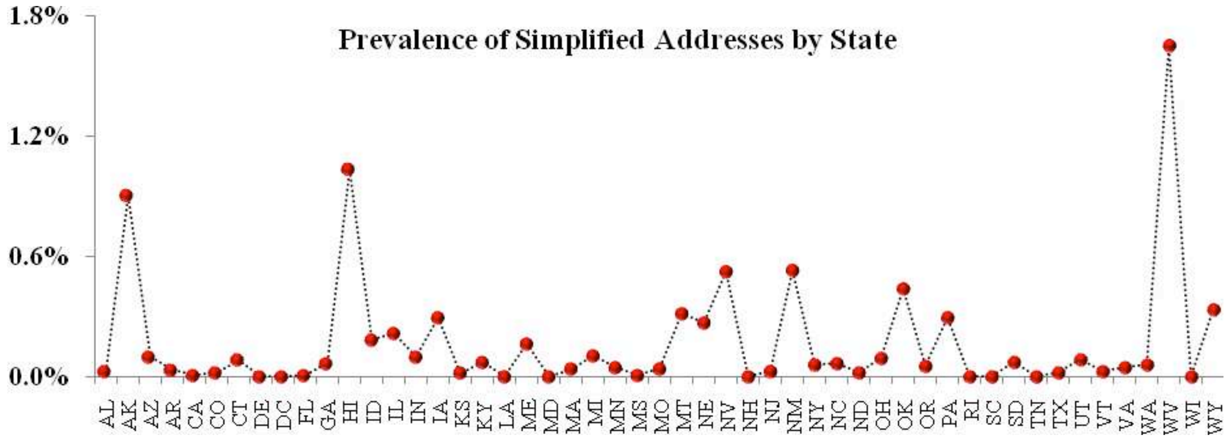
While the CDSF can provide basic delivery details about an address, oftentimes, researchers require detailed geodemographic data for sample design and allocation. Such needs can be accommodated by appending Census-based localities and population figures at the available geographic levels. Moreover, by accessing several commercial databases that contain various data items for households it is possible to enhance the CDSF for targeted sampling applications. While many of such data items correspond to individual households, there are also modeled characteristics that are available at different levels of aggregation. Starting from the ZIP+4 level, which typically consists of only a handful of households, the resulting information can then be rolled up to higher levels, including all Census geographic domains (Block, Block Group, Tract, County, MSA, State, and Region); marketing geographic domains (Media Markets, ZIP Areas, etc.); as well as custom areas (Retail Trading Areas and specific geographies based on distance or radius).

Name and Telephone Number Retrieval

Customizing the initial mailings to sample households is known to improve response rates and reduce cost. Given the plethora of junk-mail that households receive on daily basis where the packets typically carry generic contact information, research suggests that the rate of response can increase when the name of survey recipients appear on the mailed material (Dillman 1991). Moreover, with multi-mode survey applications one can reduce the number of nonrespondents to the mail survey through follow-up phone calls. Taking advantage of several databases, MSG makes it possible to retrieve names and telephone numbers associated with many of the CDSF addresses. Depending on the type of addresses, up to 85% of addresses can be name-matched and half can be linked to a landline telephone number, although match rates decrease with inclusion of P.O. Boxes.

Simplified Address Resolutions

Since the CDSF only provides counts of undeliverable (simplified) addresses that are void of street numbers or other pertinent delivery information, resolution of such cases provides an important enhancement for sampling purposes. While the number of such addresses is rapidly decreasing as they go through the 9-1-1 address conversion, currently there are just over 100,000 simplified addresses in the CDSF. As seen from the following chart, the distribution of simplified addresses varies across states with West Virginia topping the rank with less than 2% of its addresses considered to be simplified. Again, by accessing several large databases that contain different information for households MSG obtains the missing information for virtually all simplified addresses. Subsequent to this resolution, all other informational data that exist for addressed households become available for sample design and data collection purposes.



Indicators for Areas With Potential Coverage Problem

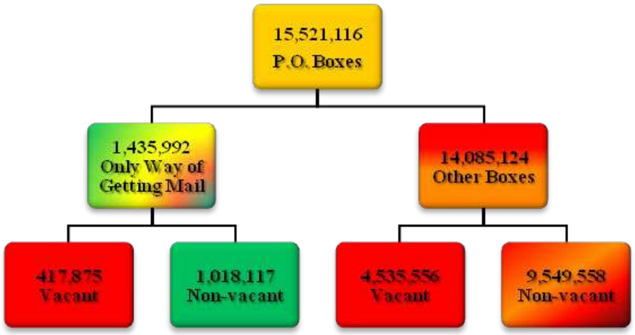
There are delivery points that are reachable only via P.O. Boxes. Also, in certain areas there are newly constructed dwellings that are currently not registered with the Postal Service. In such cases, the physical location of the corresponding households maybe unknown and not included in the CDSF. Yet, there are surveys for which visits to sample households is part of the data collection protocol for in-person interviewing or gathering of physical measurements. For such instances, contacting households by mail or telephone (if obtainable) may not be a viable alternative for survey administration – although, mail or telephone can always be used to recruit such households and obtain their residential addresses. When the physical location of a household is unviable, it might become necessary to resort back to the traditional method of onsite enumeration.

Given the significant cost of such an endeavor, researchers have developed creative options for assessing the need for onsite enumerations so that only those areas poorly covered by the CDSF may require onsite enumeration. For example, using regression models based on specific characteristics of area segments it is possible to predict the quality of the CDSF coverage (McMichael et al. 2010 and Montaquila et al. 2010). Relying on various commercial and public databases MSG can provide the needed covariates for use in such models.

Frame Multiplicity Reduction

As mentioned earlier, there are CDSF delivery points that are reachable only through a P.O. Box. While currently there are over 1.4 million of such delivery points, about 70% are non-vacant. Aside from these, there are about 14 million additional P.O. Boxes that are not the only means of delivery. In all likelihood, these correspond to households that are represented in the CDSF multiple times: once as a residential address and one or more times via a P.O. Box. Consequently, by eliminating such P.O. Boxes and those that are the only means of delivery but vacant, it is possible to remove virtually all duplicate listings in the CDSF. Before selection of samples, MSG provides the counts of these and all other delivery types so that researchers can determine the exact composition of the sampling frame for their surveys.

Composition of P.O. Boxes in the CDSF



Concluding Remarks

All single-mode methods of data collection are subject to growing rates of coverage and participation difficulties. Surveys that rely on telephone for data collection, in particular RDD-based surveys, suffer from pronounced coverage problems. In-person surveys are typically too costly to be practical as the only mode of data collection in many instances, and mail surveys alone are often too slow and secure too low of a response rate to produce reliable results. It is against this background that multimode methods of data collection are gaining popularity as alternatives that can reduce many of the problems associated with single-mode methods. In this regard, address-based sampling provides a convenient framework for development of effective sampling designs and creative protocols for implementation of surveys that employ multi-mode alternatives for data collection.

The Computerized Delivery Sequence File of the USPS can provide a powerful tool for sample surveys, however in its raw form the CDSF is simply a database for delivery of mail. It is only through proper enhancements that the CDSF can evolve into an effective sampling frame for selection of probability-based samples with surgical precisions. Enhancements provided by MSG aim to achieve this critical objective by significantly improving the coverage of the CDSF and expanding its utility for complex sampling designs and analytical applications.

References

1. Aquilino, W. (1994). Interview mode effects in surveys of drug and alcohol use: a field experiment. Public Opinion Quarterly, 58, 210-40.

2. Barron, M., J. Kelly, R. Montgomery, J. Singleton, H. Shin, B. Skalland, X. Tao, and K. Wolter. (2010). More on the Extent of Undercoverage in RDD Telephone Surveys Due to the Omission of 0-Banks. Survey Practice, April 2010.

3. Biener, L., Garrett, C.A., Gilpin, E.A., Roman, A.M., & Currivan, D.B. (2004). Consequences of declining survey response rates for smoking prevalence estimates. American Journal of Preventive Medicine, 27(3), 254-257.

4. Biemer, P.P. & Lyberg, L.E. (2003). Introduction to Survey Quality, New York: John Wiley & Sons, Inc.

5. Blumberg, S. and Luke, V. (2007). "Early Release of Estimates from the National Health Interview Survey."

6. Brick, J. M., J. Waksberg, D. Kulp, and A. Starer. 1995. "Bias in List-Assisted Telephone Samples." Public Opinion Quarterly, 59: 218-235.

7. Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. Public Opinion Quarterly, 69, 87-98.

8. de Leeuw, E. & de Heer, W. (2002). Trends in household survey nonresponse: a longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge (Eds.), Survey Nonresponse (pp. 41-54). New York: John Wiley & Sons, Inc.

9. Dillman, D. A. 1991. The Design and Administration of Mail Surveys, Annual Review of Sociology, 17, 225-249.

10. Dillman, D., Sangster, R., Tanari, J., & Rockwood, T. (1996). Understanding differences in people's answers to telephone and mail surveys. In Braverman, M.T. & Slater J.K. (eds.), New Directions for Evaluation Series: Advances in Survey Research. San Francisco: Jossey-Bass.

11. Dohrmann, S., Han, D. & Mohadjer, L. (2006). Residential Address Lists vs. Traditional Listing: Enumerating Households and Group Quarters. Proceedings of the American Statistical Association, Survey Methodology Section, Seattle, WA. pp. 2959- 2964.

12. Groves, R.M. (2005). Survey Errors and Survey Costs, New York: John Wiley & Sons, Inc.

13. Fahimi, M., M. W. Link, D. Schwartz, P. Levy & A. Mokdad (2008). "Tracking Chronic Disease and Risk Behavior Prevalence as Survey Participation Declines: Statistics from the Behavioral Risk Factor Surveillance System and Other National Surveys." Preventing Chronic Disease (PCD), Vol. 5: No. 3.

14. Fahimi, M., D. Creel, P. Siegel, M. Westlake, R. Johnson, & J. Chromy (2007b). "Optimal Number of Replicates for Variance Estimation." Third International Conference on Establishment Surveys (ICES-III), Montreal, Canada.

15. Fahimi, M., Chromy J., Whitmore W., & Cahalan M. Efficacy of Incentives in Increasing Response Rates. (2004). Proceedings of the Sixth International Conference on Social Science Methodology. Amsterdam, Netherlands.

16. Fahimi, M., Kulp, D. & Brick, J. M. (2009). A Reassessment of List-Assisted RDD Methodology. Public Opinion Quarterly, Vol. 73, No. 4, pp. 751–760.

17. Gary, S. (2003). Is it Safe to Combine Methodologies in Survey Research? MORI Research Technical Report.

18. Iannacchione, V., Staab, J., & Redden, D. (2003). Evaluating the use of residential mailing addresses in a metropolitan household survey. Public Opinion Quarterly, 76:202-210.

19. Link, M., M. Battaglia, M. Frankel, L. Osborn, & A. Mokdad. (2006). Addressed-based versus Random-Digit-Dial Surveys: Comparison of Key Health and Risk Indicators. American Journal of Epidemiology, 164, 1019 - 1025.

20. McMichael, J., V. Iannacchione, B. Shook-Sa, J. Ridenhour, K. Morton, and J. Chromy (2010). Predicting the Coverage of Address-Based Sampling Frames Prior to Sample Selection. Joint Statistical Meetings, Vancouver Canada.

21. Montaquila, J., S. Shore, and V. Hsu (2010). An Investigation of the Presence or Absence of Households at Addresses Obtained by Field Listing and from USPS Lists. Joint Statistical Meetings, Vancouver Canada.

22. O'Muircheartaigh, C., S. Eckman, and C. Weiss (2003). Traditional and enhanced field listing for probability sampling. Proceedings of the American Statistical Association, (CD-ROM), Alexandria, VA, pp.2563- 2567.

23. Staab, J.M., & Iannacchione, V.G. (2004). Evaluating the use of residential mailing addresses in a national household survey. Proceedings of the American Statistical Association, Survey Methodology Section (CD-ROM), Alexandria, VA, pp.4028- 4033.

24. Tucker, C. Lepkowski, J. M. (2008). Advances in Telephone Survey Methodology. New York: John Wiley & Sons, Inc.

25. Voogt, R. & Saris, W. (2005). Mixed mode designs: finding the balance between nonresponse bias and mode effects. Journal of Official Statistics. 21, 367-387.

26. Weiss, A., M. Battaglia, J. Boyle, A. Hyon, and D. Kulp. 2009. "Unlisted Banks in New York City: Coverage Error and Bias in Urban Areas from RDD Samples Based on Hundreds Banks with Listed Numbers." Presented at the American Association for Public Opinion Research, Hollywood, FL.

27. Wilson, C., Wright, D., Barton, T. & Guerino, P. (2005). "Data Quality Issues in a Multi-mode Survey" Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Miami, FL.