

RECENT DEVELOPMENTS IN ADDRESS-BASED SAMPLING (ABS)

Overview

Increasingly, survey and market researchers are reconsidering address-based sampling (ABS) methodologies to reach the general public for survey administrations as well as commercial applications. Essentially, there are three main reasons for this emergence:

- Evolving coverage problems associated with sampling methods that employ random digit dialing (RDD);
- Eroding rates of response to single modes of contact, along with the increasing costs to counter nonresponse; and
- Recent improvements in the databases of household addresses available to researchers.

This note provides an assessment of the above, evaluates pros and cons of ABS as an alternative, and introduces specific enhancements provided by Marketing Systems Group (MSG) that can establish this emerging methodology as a practical solution for survey and market research applications. In particular, such enhancements include amelioration of some of the known coverage problems of the ABS frames as well as their augmentations with geodemographic and other ancillary data items. While reducing bias due to undercoverage, such enhancements enable researchers to develop more efficient sample designs as well as broaden their analytical possibilities through an expanded set of covariates for hypothesis testing and related inferential tasks.

Coverage Problems for RDD Samples

For the past two decades, a large portion of telephone surveys have been based on the list-assisted RDD methodology where sample telephone numbers are selected from the 100-series telephone banks that contain at least one listed number (1+listed banks). During the intervening years, however, this methodology has overlooked the many fundamental changes in the U.S. telephony and relied on a convenient assumption that elimination of other telephone banks – those with no listed numbers – from the sampling frame amounts to exclusion of a small percentage of landline households. However, recent investigations suggest that the extent of undercoverage for landline RDD frames that only include 1+listed banks is growing for several reasons, including the emergence of alternative providers of telecommunication services through Voice over Internet Protocol (VoIP) and cable. What further deepens this coverage problem is the increasing number of households that are cutting the cord by relying only on cellular devices for voice services.

Eroding Rates of Response to Single Modes of Contact

It is becoming progressively more difficult to secure respectable rates of response for all surveys, particularly those relying on single methods of contact. Given that nonresponse is highly differential and can vary significantly across different demographic subgroups, it is of great concern when a notable percentage of sample units opt not to respond to a survey. Even when sophisticated nonresponse adjustment procedures are employed to reduce the incurred bias, it would 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 as weighting inflates the variance of survey estimates. As such, single-mode methods of contact are rendering inadequate and becoming an outdated protocol for survey administration.

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 that require contacts with representative samples of households. The Computerized Delivery Sequence File (CDSF) of the USPS with more than 135 million delivery points on file is the most complete address database available in the US. What is more, by providing validation services for both correctness and completeness of addresses the CDSF can significantly enhance the address hygiene to reduce the number of undeliverable-as-addressed mailings and increase the speed of delivery. Also, with daily feedback from tens of thousands of letter carriers the database is updated on a nearly continuous basis. The following table provides summary counts for the main types of delivery points.

Delivery Point Type	Count (March 2011)
City Style/Rural Routes	114,990,949
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 undercoverage and poor rates of response problems undermining the viability of the conventional methods of survey design and administration, researchers are now considering the use of CDSF for sampling purposes. This is becoming of particular popularity because ABS designs provide a versatile framework for creative methods of survey administration while virtually eliminating the problem of undercoverage. Considering that through reverse-matching telephone numbers for many addresses can be obtained, under an ABS survey protocol different multi-mode strategies for survey administration can be developed to accommodate the timing, budgetary, and response rate needs of surveys. At the same time, indeterminable coverage problems of landline RDD sampling methods as well as those related to the exclusion of cell-only households become moot issues at once.

Potential Issues When Using 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, for surveys where data are to be collected in-person reliance on delivery information may not be adequate as the exact location of all sample dwellings must be known. This is of particular concern 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. In what follows we provide a brief overview of how many of these shortfalls may be avoided through CDSF enhancements provided by MSG.

Available Enhancements for the CDSF – the MSG Difference

Since there is not a one-to-one correspondence between the USPS and Census geographic definitions, it is not possible to append many of the ancillary data items to the CDSF address that are available through the government and commercial sources. However, by geo-coding each address to a unique Census block this divide can be bridged to allow appendage of valuable information to all US addresses. This is the crossroad where basic list suppliers – those that can simply offer raw extracts from the CDSF – are differentiated from MSG. By providing the following suite of enhancements, MSG can help evolve the CDSF from a basic delivery database into a bona fide sampling frame suitable for complex surveys.

Detailed Geodemographic Information

By mapping each address to Census localities and accessing several commercial databases, MSG can append a host of household attributes to each address for stratified or targeted sampling applications. Starting from the ZIP+4 level, the resulting information can then be rolled up to higher levels of aggregation, including all Census geographic domains (Block, Block Group, Tract, County, MSA, and State); marketing domains (Media Markets and DMAs); as well as custom areas (retail trading areas and specific geographies based on distance or radius).

Name and Telephone Number Retrieval

Relying on multiple database passes, MSG makes it possible to retrieve names and telephone numbers for CDSF-based sample addresses. This is why on average more than 85% of our sample addresses can be name-matched and 50% can be linked to a landline telephone number. Obviously, match rates depend on several factors and can decrease with inclusion of P.O. Boxes.

Undeliverable Address Resolutions

The CDSF contains records that are void of delivery information, commonly referred to as simplified addresses. While the number of such addresses is rapidly decreasing as they go through the 911 address conversion, currently there are just over 100,000 simplified addresses. MSG can resolve the majority of simplified addresses and make it possible for all ancillary data items available for addressed households to be appended to the resulting addresses as well. Moreover, MSG has entered into an agreement to receive data on addresses identified by the USPS as No-Stat. Such addresses include vacant or new housing units, as well as delivery points within gated communities. This supplement adds to our ABS frame new delivery points and previously vacant units that have become occupied, without waiting for the next version of the CDSF to become available.

Identification of 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. Given the significant cost of on-site enumeration that can become necessary for such areas, MSG has developed procedures for predicting the quality of the CDSF coverage based on specific characteristics of area segments obtained from commercial and public databases.

Frame Multiplicity Reduction

As mentioned above, there are CDSF delivery points that are reachable only through a P.O. Box. While currently there are over 1.4 million such delivery points, about 70% are non-vacant. Aside from these, there are 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 P.O. Boxes. 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 from the CDSF. Before selection of samples, MSG provides the counts of these and all other delivery types in the geography of interest so that our clients can determine the exact composition of the sampling frame for their surveys.

Selection of Samples Based on Complex Designs

In addition to the above enhancements, MSG is the only sample generating company with in-house sampling statisticians for design and selection of samples based on complex designs. Under the supervision of our Chief Methodologist, Dr. Mansour Fahimi, we offer sampling design consultations to clients with a wide range of market and survey research needs. Proper selection of such samples often requires applications of advanced procedures to accommodate multi-stage sampling in conjunction with selection methods with probabilities proportional to various size measures.

Virtual Platform for Address Based Sampling

In order to accommodate our clients' sampling needs on a 24/7 basis, MSG is the only company that offers virtual sampling – both for RDD and ABS – through our secure client-server architecture. Virtual ABS makes it possible to select probability-based sample addresses from our enhanced frame based on several selection parameters, such as geographic domains down to ZIP+4 as well as various attributes unique to each address. As with custom orders, with Virtual ABS our clients have access to our advanced process for name and telephone number retrieval that result in the highest match rates possible.

For a full version of this whitepaper please visit our website at: www.m-s-g.com or contact us at 215-653-7100.