



December 2008

## White Paper

# Address Based Sampling

---

*(Centris provides its clients with an Address Based Sampling solution in its US Communication and Entertainment Omnibus Survey. It relies on its sister company, Marketing Systems Group (MSG) to develop its sample frames. This White Paper provides background on the Address Based Sampling approach used by MSG as well as the rationale for using this sampling methodology.)*

### Overview

Increasingly, survey and market researchers are reverting back to address-based methodologies to reach the general public for survey administration and related commercial applications. Essentially, there are three main factors for this change: evolving coverage problems associated with telephone-based methods; eroding rates of response to telephone contacts along with the increasing costs of remedial measures to counter non-response; and on the other hand, recent improvements in the databases of household addresses available to researchers. This note provides an assessment of these three factors along with specific enhancements the Marketing Systems Group (MSG) can offer when developing an address-based protocol for survey and market research applications.

In particular, enhancements provided by MSG include amelioration of some of the known coverage problems associated with the addressed-based sampling frames as well as their augmentations with demographic, geographic, and other supplementary data items. While reducing bias due to under-coverage – particularly in rural areas where more households rely on P.O. Boxes and inconsistent address formats – such enhancements enable researcher to develop more efficient sample designs as well as broaden their analytical possibilities through an expanded set of covariates for hypothesis testing and statistical modeling tasks.

### Coverage Problems for Telephone Surveys

There is a growing body of literature revealing the many coverage problems researchers are facing when developing representative telephone samples for the general public. For instance, most Random Digit Dialing (RDD) samples are generated within the 100-series telephone banks that contain at least one listed number. For years, this method of list-assisted RDD has relied on a fundamental assumption that exclusion of 100-series telephone banks with no listed numbers – zero-listed banks – results in a small coverage bias. Brick et al. (1995) had estimated that only 3.7 percent of all telephone households were not covered when the sampling frame was confined to listed 100-series banks. However, recent investigation by Fahimi et al. (2008) suggests that the extent of this coverage bias is now approaching 20 percent.

On the other hand, with improvements in cellular coverage and ease of number porting an increasing proportion of households is forsaking landlines and relying entirely on wireless phones for voice communications. The latest estimates suggest that over 15 percent of all households are now cell-only. To make the situation more complicated, the rate of this switch is not uniform across all household types. For instance, more than 25 percent of households headed by someone 35 years of age or younger are only reachable by cell phones. More alarmingly, it is estimated that almost 50 percent of adults living with roommates are among the cell-only population. Moreover, an equally sizable and growing number of households are becoming cell-mostly, resulting in 3 out of every 10 adults in the U.S. receiving all or nearly all of their calls on cell phones (Blumberg and Luke 2007).



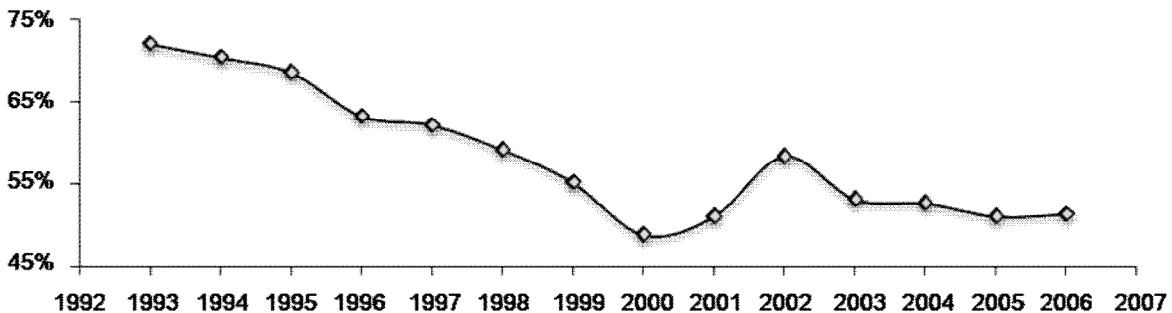
Also, there are many secondary problems that have resulted from the above telecommunication metamorphosis. For example, a growing portion of the cell-only households consists of those that have ported their existing landline numbers to a cellular device. When embedded in typical RDD samples these ported cellular numbers are indistinguishable from regular landline numbers, presenting hidden legal ramifications for call centers that utilize predictive dialing equipment. Conversely, cellular RDD samples that are selected from dedicated cellular exchanges will fail to include ported landline numbers.

While a large proportion of cell-only households that started as such tend to consist of younger individuals, those that became cell-only by porting their landline numbers tend to skew in the other direction of older and more established households. Consequently, developing sound sampling strategies for the conventional telephone-based method of RDD and producing reliable estimates from the resulting data are no longer straightforward challenges.

### Eroding Rates of Response to Telephone Surveys

Biener et al. (2004) and Curtin et al. (2005) point out that the rate of response to telephone surveys has been on a decline. More recent investigations by Fahimi et al. (2007a) suggest that the national rates of response to the Behavioral Risk Factor Surveillance System (BRFSS) survey, which is the largest RDD survey in the world, follow this trend as well. As shown in the following figure, BRFSS has suffered a drop of nearly 20 percentage points in response rates during the course of the past decade.

Figure 1. Response rate for the BRFSS survey from 1993 to 2006



Given that non-response is highly differential in nature and varies 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 non-response adjustment procedures are employed to reduce the incurred bias, it would be naive to assume such remedial procedures can reduce non-response bias to a tolerable and measurable level. Also, it should be noted that reducing non-response bias via weighting is always exercised at the expense of the precision of survey estimates, since weight adjustments inflate variance estimates (Fahimi et al., 2007b).



Beyond statistical techniques, many researchers have resorted 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 to 20 percentage points can still leave a survey with an overall non-response rate above 50 percent. Moreover, this marginal gain in response rate is often achieved at a high cost, as practical non-response 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 mentioned above, the overall cost of dealing with non-response 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 Delivery Sequence File (DSF) of the USPS is a computerized database that contains all delivery point addresses, with the exception of general delivery where carrier route<sup>1</sup> or P.O. Box delivery is not available and mail is held at a main post office for claim by recipients. With over 125 million records, DSF provides mailers with the following delivery features.

- Address validation and standardization;
- ZIP+4 and carrier route coding;
- Delivery sequence;
- Detection of addresses that are potentially undeliverable;
- Delivery-type code that indicates business or residential; and
- Seasonal delivery information.

The second generation of this database (DSF2) is the most complete address database available. With more than 135 million addresses on file, it is safe to assume that if an address cannot be matched against DSF2 it is probably undeliverable. While providing validation services for both correctness and completeness of addresses, DSF2 can determine delivery type and Locatable Address Conversion System (LACS) information. As the 9-1-1 address conversion process continues to change the rural-style delivery points to one of city-style, LACS allows mailers to obtain the revised versions of such addresses. It should be noted that in some instances this conversion may result in renaming or renumbering of existing city-style addresses, however, the Address Management System (AMS) of the USPS makes it possible to obtain the needed transfer information for affected addresses. By providing the most current address information and enhancements to address hygiene, this system helps reduce the number of undeliverable-as-addressed mailings, increase the speed of delivery, and reduce cost. Given daily feedback from tens of thousands of letter carriers, the database is updated on a nearly continuous basis.

## Using DSF for Sample Survey Purposes

Given the evolving problems associated with telephone surveys on the one hand, and the exorbitant cost of on-site enumeration of housing units in area probability sampling applications on the other, many researchers are considering the use of DSF for sampling purposes. Moreover, the growing problem of non-response – 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.

---

<sup>1</sup> A carrier route consists of 100 to 2,500 households served by an individual mail carrier within a [five-digit Zip Code](#) area. There are approximately 570,000 carrier routes in the U.S



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 addressed-based sample designs provide a convenient framework for an effective administration of surveys that employ multi-mode alternatives for data collection.

Cognizant of the potential implications of combining different modes of data collection, the emerging conclusion from many studies seem to suggest that different research modalities (CATI, in-person, or self-administered via web, IVR, or mail) 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). As for comparisons with in-person and mail-only modes of data collection, needless to say, the former is too costly to be practical for many survey applications while the latter (with notoriously low rates of response) requires expensive non-response 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.

Considering that through reverse-matching the telephone numbers for a large proportion of 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 DSF-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 via web, dial-in numbers for live interviewing, an IVR system, or other options. 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.

## Potential Issues When Using DSF for Sampling Purposes

As reported by a number of researchers, certain households have a higher likelihood of not being included as a delivery point on the DSF. Staab and Iannacchione (2003) estimate that approximately 97 percent of all US households have locatable mailing addresses, however, this prevalence diminishes with population density and approaches zero in areas where home delivery of mail is unavailable. Dohrmann and Mohadjer (2006) report that when comparing lists of on-site enumerated addresses to DSF generated listings of households for the same geography, in rural areas the rate of mismatches can be over 23 percent<sup>2</sup>. However, these researchers do indicate that as rural area addresses go through the 911 address conversion and acquire a city-style format, the coverage of DSF-based lists in rural areas is likely to improve in the future. Also, O'Muircheartaigh et al. (2003) point out another source of under-coverage for address-based samples when lists of households are purchased directly from commercial list compilers because households can request that their addresses not be sold.

Beyond coverage issues, when DSF generated samples are used in surveys that adopt 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

<sup>2</sup> We speculate this unusually high rate of mismatch is partly due to the method used for address comparisons. Fewer mismatches should result with more comprehensive matching techniques and resolution of undeliverable addresses.



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 might 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).

## Enhancements of DSF

As mentioned above, the current version of DSF can serve as a comprehensive sampling frame for address-based survey applications when complete coverage and proper representation of the target population are among the nonnegotiable features of the sample design. In addition to a near perfect coverage of the entire country, this versatile database can provide scientific samples for finely defined geographic sub-domains – a resilient point of contention for telephone-based surveys, particularly in light of the growing coverage problem resulting from landline number portability. In addition to the available information that can be accessed directly from DSF2, it is possible to append many ancillary data items to each address for use in complex surveys that require detailed information for stratification purposes. This is the crossroad where basic list suppliers, those that simply offer extracts from what the USPS provides, are differentiated from reputable statistical sampling companies that provide enhanced versions of DSF2.

As the survey research industry's preferred full-service sampling company and creator of the first and only in-house and Web-based sample design and generation systems, MSG is capable of providing significant enhancements for DSF-based samples. Beyond complete access to all the available data items on DSF2, a brief listing of key enhancements provided by Centris include:

- Detailed Geographic Information;
- Household Demographic Information;
- Name and Telephone Number Retrieval;
- Simplified Address Resolutions; and
- Voice, Video, and Data Usage Information.

**Detailed Geographic Information** can be appended by MSG by taking advantage of its comprehensive geographic correspondence databases that extend beyond what is provided by the USPS. 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 of aggregation, 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). MSG also maintains a complete set of telecommunications and data provision boundary layers, including those for Cable Systems (MSO's), Rate/Wire Centers, along with availability of Telco Fiber and DSL (by speed). In addition to sample design applications, such detailed geographic information can be particularly beneficial in the construction of area probability samples for in-person data collection, which in conjunction with the information available from DSF can virtually eliminate the need for costly on-site enumerations or pre-listings.

**Household Demographic Information** can improve the efficiency of a sample both with respect to design (stratification) and data collection (field work). While DSF2 can provide basic geographic details about an address, oftentimes, researchers rely on demographic data for sample design and allocation. By accessing several large databases that contain various demographic data items for each household, MSG can enhance DSF-based sampling frames for targeted sampling applications.



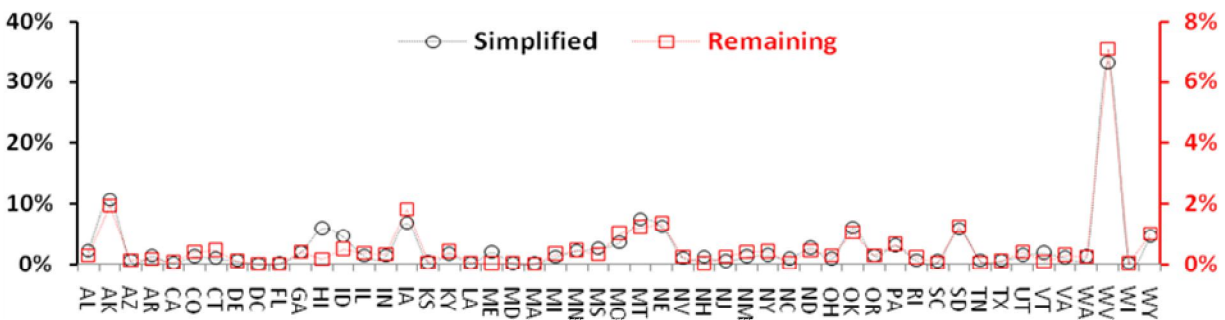
While many of such data items correspond to individual households, there are also modeled characteristics that are available at different levels of aggregation, such as block groups, telephone exchanges, counties, etc.

**Name and Telephone Number Retrieval** are known to improve response rates and reduce data collection cost by customizing the initial mailings to individual sample households. Given the plethora of junk-mail that households receive on daily basis where the packets typically carry generic contact names, research suggests that the rate of response can increase significantly 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 non-respondents to the mail survey through follow-up phone calls.

Based on years of experience with these activities and access to several external sources, MSG can enhance the DSF-based samples by successfully retrieving names and telephone numbers associated with many addresses. On average, about 85 percent of addresses can be name-matched and over 60 percent can be linked to a landline telephone number – match rates decrease with inclusion of P.O. Box addresses.

**Simplified Address Resolutions** are among the most important enhancements that can be added when selecting address-based samples; since the DSF only provides counts of these undeliverable addresses that are void of street numbers or other pertinent delivery information. While the number of such addresses is rapidly decreasing as they go through the 9-1-1 address conversion, currently there are about 1.5 million simplified addresses in the DSF. As seen from the following chart, the distribution of simplified addresses varies across states with West Virginia topping the rank with more than 30 percent of its addresses considered to be simplified. Again, by accessing several large databases that contain different information for households, MSG can obtain the missing information for almost 88 percent of simplified addresses. Subsequent to this resolution, all other informational data that exist for addressed households become available for sample design and data collection purposes.

Figure 2. Percent simplified addresses and remaining simplified addresses after MSG augmentation by state as of May 2008



**Voice, Video, and Data (VVD) Usage Information** includes exact and modeled characteristics for households – an expending wealth of proprietary information exclusively available to MSG through its CENTRIS® database. CENTRIS is the only national database that continuously collects household information on the choice and use of VVD as well as electronic products and services. Subsequently, these behavioral estimates are projected to various geographic levels starting from the Census block group on up.



Among other utilities, such estimates provide a unique framework for exploring market potential, tracking market share, and forecasting demand in the marketplace. These distinguishing features are of high utility for enhancing sampling designs that employ usage related measures of size for disproportionate sample allocations in consumer studies and related survey applications.

## Concluding Remarks

Due to ongoing changes in the U.S. telecommunications infrastructure, telephone-based surveys are facing challenges with respect to coverage and response rates as well as cost. In fact, all single-mode methods of data collection are subject to such difficulties on varying basis. Mail surveys often secure too low of a response rate to produce reliable results; in-person interviews are typically too costly to be practical in many instances; and telephone surveys suffer from both coverage and response rate problems. It is against this background that multi-mode methods of data collection are gaining popularity as alternatives that might reduce some of the problems associated with single-mode methods. As such, addressed-based samples provide a convenient framework for effective design and implementation of surveys that employ multi-mode alternatives for data collection. In this regard, the Delivery Sequence File of the USPS – once properly enhanced and prepared – provides a power tool for complex surveys. Enhancements provided by MSG can significantly improve the coverage of DSF and expand its utility for design and analytical applications.

## References

- Aquilino, W.S. (1994). Interview mode effects in surveys of drug and alcohol use: a field experiment. *Public Opinion Quarterly*, 58, 210-40.
- 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.
- Biemer, P.P. & Lyberg, L.E. (2003). *Introduction to Survey Quality*, New York: John Wiley & Sons, Inc.
- Blumberg, S. J. and Luke, V. J. (2007). "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey." <http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200805.htm>.
- Brick, J. M., J. Waksberg, D. Kulp, and A. Starer. 1995. "Bias in List-Assisted Telephone Samples." *Public Opinion Quarterly*, 59: 218-235.
- Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey non-response over the past quarter century. *Public Opinion Quarterly*, 69, 87-98.
- de Leeuw, E. & de Heer, W. (2002). Trends in household survey non-response: a longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge (Eds.), *Survey Non-response* (pp. 41-54). New York: John Wiley & Sons, Inc.
- Dillman, D. A. 1991. The Design and Administration of Mail Surveys, *Annual Review of Sociology*, 17, 225-249.
- 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.
- 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.
- Groves, R.M. (2005). *Survey Errors and Survey Costs*, New York: John Wiley & Sons, Inc.
- 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)*, Volume 5: No. 3.
- 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.



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.

Fahimi, M., Kulp, D. & Brick, J. M. (2008). Bias in RDD Sampling: A 21st Century Digital World Reassessment. Presented at the American Association for Public Opinion Research Annual Conference, New Orleans, LA.

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

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.

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.

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

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.

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

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.