When to Use Commercial Data for Improved Efficiency

Colm O'Muircheartaigh[1,2], Ned English[1], Holly Hagerty[1], and Carina Hoyer[1]

[1]NORC at the University of Chicago, 55 E. Monroe Street, Ste 3100, Chicago, IL, 60603

[2]University of Chicago Harris School of Public Policy, 1155 E. 60th Street, Chicago, IL, 60637

Abstract

Address-Based Sampling [ABS] designs based on extracts of the United States Postal Service computerized delivery sequence file (USPS CDSF) allow for the enhancement with information from supplementary lists designed to identify households belonging to specific demographic groups, including those that could be considered rare or hard-to-reach (H2R). We know from previous research, however, that such targeted lists can have reduced coverage for certain subgroups. Our paper uses area information from the American Community Survey and other sources to predict when a specific supplementary source is most appropriate for a specific target subgroup. We also describe the characteristics of survey respondents who are well-identified vs. poorly-identified from a specific targeted list source or combination of sources.

Key Words: ABS, commercial data, coverage, enhanced sample design

1. Introduction

Address-based (ABS) designs have increased in popularity in recent years as a platform for multiple mode data collection (Harter et al. 2016, Link et al. 2008). A significant additional potential advantage of ABS designs is the possibility of appending to the sampling frame location-based data at multiple scales and from multiple sources, including at the area, household or individual level (English et al. 2018, Harter et al. 2016, Smith and Kim 2013). However, it is documented that the data sources vary in quality and contain a variety of errors (Roth et al. 2018, Harter et al. 2016).

AmeriSpeak® is a multi-mode ABS panel designed to support NORC's mission to deliver reliable data to guide critical programmatic, business, and policy decisions (Dennis 2017, Montgomery et al. 2016). AmeriSpeak uses the continuously-updated NORC 2010 National Sampling Frame to create a nationally-representative sample with specific age and race/ethnic oversamples (Pedlow and Zhao 2016). At the stage of household selection the AmeriSpeak design incorporates vendor-provided demographic data to target households based on their expected race/ethnicity, age, or other factors. At issue is how the accuracy of such data might impact survey efficiency and the resulting data.

The purpose of our analysis is to understand the utility of vendor-provided data for targeting subgroups of interest to surveys, in the particular case of households containing Latino members. We explore a design that recruits only those households flagged as being "Latino" on an address list enriched by multiple vendors. We examine (i) the coverage of the Latino population in such a design and (ii) the characteristics of households incorrectly

flagged as Latino in the data sources. Finally we examine the whole frame to find Latino households that are not identified as such by the vendors. Our research is of interest to practitioners of ABS surveys or those interested in targeting rare or hard-to-reach populations.

2. Background

Sample members on the AmeriSpeak panel are first selected from NORC's National Sampling Frame, an area-probability frame funded and managed by NORC and used for national in-person studies at NORC including the General Social Survey and Survey of Consumer Finances (Pedlow and Zhao 2016). The NORC national frame is fundamentally based on an extract of the U.S. Postal Service (USPS) database called the Computerized Delivery Sequence File (CDS or CDSF), known to have very high coverage of US households, especially for those that use the mail-mode (Harter et al. 2016, Iannacchione 2011, Link et al. 2008, O'Muircheartaigh et al. 2006).

One advantage of address-based designs is the potential to append auxiliary data to the frame or samples (Harter et al. 2016), either from federal data sets at areal units of aggregation or from commercial frames identifying the households themselves. Common examples of household or member level appends include telephone number associated with an address (Olson and Buskirk 2015), the number of adults in a household (Roth et al. 2018), home tenure (Roth et al. 2018), the presence of children (English et al. 2014), or demographics such as age, race/ethnicity, or income (Roth et al. 2018, Pasek et al. 2014, DiSogra et al. 2010). It is clear from the literature that both the match-rate and accuracy of appended data vary depending on the variable of interest and specific geography of the households in question (Roth et al. 2018, Amaya, Skalland, and Wooten 2010, Buskirk et al. 2014, Pasek et al. 2014). One reason for such variability is how the data are modeled, compiled, and appended to an individual address (English et al. 2017).

3. Data and Methods

Our results focus on the two specific commercial vendors that AmeriSpeak licensed to enhance the ABS frame in 2015, which we will refer to as "Vendor A" and "Vendor B", and the effectiveness of these vendors in identifying Latino households before panel recruitment. Variables of interest chosen for analysis were collected from responding households in AmeriSpeak; these included age, race, Hispanic/Latino ethnicity, income level, marital status, educational attainment, presence of children in the household, and home ownership. These are respondent-reported variables which we analyze at the household level. We appended data from the American Community Survey (ACS) at the census tract level to each address for contextual information. Doing so allowed us to examine the quality of the vendor information in terms of neighborhood-level contextual information such as the percentage of households that were below poverty.

We focused on what is essentially a cross-tabulation of Latino / non-Latino by flagged / not flagged. "Latino" households are defined as households containing at least one Hispanic/Latino member and non-Latino households otherwise. We examined variables of interest for "Latino households" and "non-Latino households", analyzed by whether the household was flagged by either Vendor A or Vendor B as likely to contain a Latino individual, flagged by both vendors, or flagged by neither. Such an approach enables us to compare the characteristics of flagged or non-flagged households in each group with the

reference populations of all Latino households in the sample, or all non-Latino households in the sample. Additionally, one of the vendors provided a "Latino surname" flag at the household level, which we considered during our analysis.

We examined the coverage and hit-rate of the two vendors with respect to finding Latino households. "Coverage" (or sensitivity) is the proportion of the target subgroup matched by a given flag or flags, while "hit rate" (or precision) measures the accuracy of a specific flag. In addition to survey-reported variables, ACS tract-level information, and the vendor-provided Latino surname flag, we looked at information on whether the household contained any members of other race/ethnicity groups to provide context about the composition of households that were flagged incorrectly as Latino households.

4. Results and Discussion

Our first research question relates to how well the two vendors were able to identify Latino households, as measured by their coverage and hit-rate. Overall, we found that among those households flagged by either vendor as being "Latino" 62% were, meaning the hit-rate for either vendor was 62%. The coverage of either vendor was 64%, meaning 64% of the actual Latinos in our set of recruited households were correctly-flagged as such. About half of the households correctly identified were flagged by both vendors. Relatedly, approximately equal shares of Latino households were identified correctly by both vendors (33%) as were by neither (36%).

Table 1: Comparison of Households Correctly Flagged

Variable	*All Latino Households*	*Flagged by Either Vendor*	*Flagged by Both Vendors*
% Homes Owned	39.9	43.2	46.0
% Children in Household	51.0	52.9	39.7
% < $30,000 Household Income	31.3	29.8	30.7
% Latino by Tract	39.8	47.9	50.6
% Latino + African-American by Tract	50.3	57.1	59.2
% Below Poverty by Tract	16.8	17.7	18.3
% Latino Surname	62.1	88.4	96.7

Table 1 contrasts the characteristics of households that were correctly flagged by either (or both) vendor(s) as containing Latino members, with the same characteristics all recruited Latino households. The households flagged by the vendors broadly resembled Latino households in general, though they were overrepresented in Latino and Latino/African-American tracts. Latino surnames were also overrepresented in the vendor lists.

Table 2: Comparison of Households Misidentified

Variable	*Latinos, Correctly Flagged*	*All non-Latinos*	*Non-Latinos Flagged "Latino"*
% Homes Owned	43.2	58.1	45.2
% Children in Household	52.9	32.6	34.7
% < $30,000 Household Income	29.8	29.1	29.8
% Latino by Tract	47.9	12.7	27.4
% Latino + African-American by Tract	57.1	27.0	40.1
Any African-American Member	.6	21.2	19.6
Any White Member	2.5	69.5	64.5
Any Asian Member	.3	3.2	6.6

Table 2 describes households that were misidentified as being Latino, meaning they were flagged as "Latino" but contained no such members. Table two compares such households to all non-Latino households as well as correctly-flagged Latino households. As shown "misidentified" households were less likely to be homeowners, and more likely to reside in tracts with a higher incidence of African American or Latino households, than non-Latino households in general.

Table 3: Latino Households Not Identified as "Latino"

Variable	*All Latino Households*	*Latino, Not Flagged*	*All non-Latinos*
% Homes Owned	39.9	34.2	58.1
% Children in Household	51.0	47.8	32.6
% < $30,000 Household Income	31.3	34.0	29.1
% Latino by Tract	39.8	25.7	12.7
% Latino + African-American by Tract	50.3	38.4	27.0
% Below Poverty by Tract	16.8	15.1	13.0
% Latino Surname	62.1	16.6	5.4

Finally, table 3 describes households that were not flagged as being Latino, which would potentially represent under coverage if one depended on such lists. Non-flagged Latino households were less likely to be homeowners, tended to have fewer children, had a much lower incidence of identifiable Latino surnames, and were less likely to reside in tracts with a substantial Latino or African-American population. A sample design based on flagged lists would miss such households which would present considerable risk of bias.

5. Conclusions and Next Steps

Our research has shown that vendors employ a combination of household and member-level data with area-level characteristics to assign demographic flags. As such specific categories of households will be more likely to be flagged by a given vendor, based on

their visibility to market researchers and modeled characteristics. One conclusion for researchers is that it is necessary to consider different sampling frames depending on whether the desired outcome is hit-rate or coverage. We have also found that specific lists will favor particular subgroups at rates higher than random, with the impact on any study being domain-dependent.

Our recommendation is to use a combination of targeted lists plus a sample of unenriched addresses for both coverage and efficiency, knowing that the effort required isn't possible in all instances. Moving forward we will be pursuing modeling to understand correlates of coverage and hit-rate of attritional population groups. As such we will likely integrate area-level variables including those from the Census Planning Data Base (PDB), mail return rates, and low-response scores.

References

Amaya, Ashley, Ben Skalland, and Karen Wooten (2010), "What's in a match?" Survey Practice 3.

Buskirk, Trent D., David Malarek, and Jeffrey S. Bareham (2014), "From flagging a sample to framing it: Exploring vendor data that can be appended to ABS samples." Pp. 111-124 in Proceedings of the Survey Research Methods Section: American Statistical Association.

Dennis, J. Michael (2017), "Technical overview of the AmeriSpeak® panel NORC's probability-based research panel". White paper at http://amerispeak.norc.org/research/.

DiSogra, Charles, J. Michael Dennis, and Mansour Fahimi (2010), "On the quality of ancillary data available for address-based sampling." Pp. 4174-4183 in Proceedings of the Survey Research Methods Section: American Statistical Association.

English, Ned, Timothy Kennel, Trent Buskirk, and Rachel Harter (2018). The Construction, Maintenance, and Enhancement of Address-Based Sampling Frames. Journal of Survey Statistics and Methodology, https://doi.org/10.1093/jssam/smy003.

English, Ned, Colm O'Muircheartaigh, and Margaret Allen (2017). Using Commercial Data to Enhance Survey Eligibility: The AmeriSpeak Experience. 2017 Proceedings of the American Statistical Association, Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

English, Ned, Ying Li, Andrea Mayfield, and Alicia Frasier. The Use of Targeted Lists to Enhance Sampling Efficiency in Address-Based Sample Designs: Age, Race, and Other Qualities. 2014 Proceedings of the American Statistical Association, Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

Harter, R., M. P. Battaglia, T. D. Buskirk, D. A. Dillman, N. English, M. Fahimi, M. R (2016). Frankel, T. Kennel, J. P. McMichael, C. B. McPhee, J. M. DeMatteis, T. Yancey, and A. L. Zukerberg (2016), "Address-based Sampling." Prepared for AAPOR Council by the Task Force on Address-based sampling, Operating Under the Auspices of the AAPOR Standards Committee. Oakbrook Terrace, Il. http://www.aapor.org/getattachment/Education-Resources/Reports/AAPOR_Report_1_7_16_CLEAN-COPY-FINAL-(2).pdf.aspx Accessed March 1, 2016. 140 pages.

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

Link, Michael W., Michael P. Battaglia, Martin R. Frankel, Larry Osborn, and Ali H. Mokdad (2008), "A comparison of address-based sampling (ABS) versus random-digit dialing (RDD) for general population surveys." Public Opinion Quarterly 72:6-27.

Montgomery, Robert, J. Michael Dennis, and Nada Ganesh. (2016), "Response rate calculation methodology for recruitment of a two-phase probability-based panel: the case of AmeriSpeak". White paper at http://amerispeak.norc.org/research/.

Olson, Kristen and Trent D. Buskirk (2015), "Can I get your phone number? Examining the relationship between household, geographic and census-related variables and phone append propensity for ABS samples." in 70th Annual AAPOR Conference. Hollywood, FL.

O'Muircheartaigh, Colm, Ned English, Stephanie Eckman, Heidi Upchurch, Erika Garcia Lopez, and James Lepkowski. Validating a Sampling Revolution: Benchmarking Address Lists Against Traditional Field Listing (2016), 2006 Proceedings of the American Statistical Association, AAPOR Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

Pasek, Josh, S. Mo Jang, Curtiss L. Cobb, J. Michael Dennis, and Charles DiSogra (2014), "Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data." Public Opinion Quarterly 78:889-916.

Pedlow, S. and Zhao, J. (2016). Bias Reduction through Rural Coverage for the AmeriSpeak Panel. 2016 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association.

Roth, Shelly Brock, Andrew Caporaso, and Jill DeMatteis (2018), "Variables Appended to ABS Frames: Has Data Quality Improved?" in 73rd Annual AAPOR Conference. Denver, CO.

Smith, Tom W. and Jibum Kim (2013), "The Multi-Level, Multi-Source (ML-MS) Approach to Improving Survey Research.". GSS Methodological Report 121 at http://gss.norc.org/Documents/reports/methodological-reports/MR121%20MLMSmethreport121.pdf.