Using Commercial Data to Enhance Survey Eligibility: The AmeriSpeak Experience

Ned English[1], Margaret Allen[1], and Colm O'Muircheartaigh[1,2]
[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
The AmeriSpeak® Panel is a multi-mode address-based (ABS) panel designed to support NORC's mission to deliver reliable data to guide critical programmatic, business, and policy decisions. AmeriSpeak uses the continuously-updated NORC 2010 National Sampling Frame to create a nationally representative sample of all Americans, with specific age and race/ethnic oversamples. The ABS design allows for the enhancement of addresses from an extract of the United States Postal Service Computerized Delivery Sequence file (USPS CDSF) with lists designed to flag households as being members of specific age, race/ethnicity, or other targeted groups. AmeriSpeak has employed such "vendor-provided" lists to increase the sample size for specific demographics while not undermining the probability basis of the design. Our paper examines the utility of such lists in this context as well as any trade-offs between efficiency, expressed as a "hit rate", and the coverage of the target population. This research is relevant to survey practitioners interested in improving design efficiency for particular domains.

Key Words: ABS, commercial data, coverage

1. Introduction

The use of address-based samples (ABS) has become relatively common in recent years for surveys with in-person, telephone, mail, and web components (Harter et al. 2016, Link et al. 2008). Address-based sampling has the benefit of being able to target specific geographies and demographics based on location and the further potential to append data sourced from vendors at the household level. In the latter case it is clear that any appended household or individual-level demographic data can carry error that should be considered by survey practitioners (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 more about the qualities of vendor-provided demographic flags, using data collected during Amerispeak recruitment. First, we would like to know the accuracy of such flags for specific demographic categories; we do so by exploring both the hit-rate (precision) and coverage (sensitivity) of each list. Second, we consider the best source for a given demographic class where there is a focus on a particular subset of the population.

2. Background

AmeriSpeak households are selected initially from NORC's National Sampling Frame, an area-probability sample funded and managed by NORC and used for national in-person studies at NORC (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), shown to have very high coverage of US households (Iannacchione 2011, Link et al. 2008, O'Muircheartaigh et al. 2006). Survey practitioners often target special subpopulations, for which even area stratified samples can be inefficient.

One advantage of address-based designs is the potential to append auxiliary data to the frame or samples (Harter et al. 2016). Common examples of auxiliary data include telephone number associated with an address (Olson and Buskirk 2015), the presence of children (English et al. 2014), or demographics such as age, race/ethnicity, or income (Pasek et al. 2014, DiSogra et al. 2010). All such data appends are provided by vendors who employ proprietary data sets compiled from diverse sources, including public records, consumer databases, and others. 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 (Amaya, Skalland, and Wooten 2010, Buskirk et al. 2014, Pasek et al. 2014).

3. Data and Methods

Our results focus on the two specific commercial vendors that Amerispeak employed to enhance the ABS frame in 2015, which we will refer to as "Vendor A" and "Vendor B". The main focus of this paper is to compare *a priori* vendor flags to actual demographics from our recruiting experience in order to evaluate how well each vendor captures a given subgroup, as described by "coverage" and "hit-rate". "Coverage" (or sensitivity) is the proportion of the target subgroup matched by flags. Thus, if a given vendor were able to successfully flag half of the actual members of a given target subgroup as determined through screening, we would say their coverage was 50%.

The "hit rate" (or precision) measures the accuracy of a specific flag by dividing the number of successful matches or "hits" by the total flagged for the target. For example, if we learned that half of the time a specific flag was correct in identifying a target subgroup, we would say the hit rate was 50%. Together both coverage and hit-rate generally describe how effective a vendor would be for our purposes. For example, we may be concerned with a specific variable characterized by low coverage due to the potential risk of bias if unflagged households different from those who were flagged, while a poor hit-rate would affect survey efficiency.

Variables of interest chosen for analysis included age, race, income level, marital status, educational attainment, presence of children in the household, and home ownership. It is important to acknowledge that not all vendors publish data in identical units of measurement. For example, while Vendor A provides four different age categories, Vendor B provides five. It was thus necessary to combine results into equivalent groups to calculate the comparable hit and coverage rates among different commercial lists.

We then calculated the hit rate and coverage for each variable of interest, and conducted logistic regression modeling to understand covariates with success. Our logistic regression focused on understanding which variables from the American Community Survey (ACS) would predict a better hit-rate, including demographics, urbanicity, and population density. Our assumption in so doing would be that list accuracy varies based on social environment.

4. Results and Discussion

Table 1: Summary of Coverage Rates for Variables Examined

Variable	*Vendor A*	*Vendor B*	*Either*	*Both*
Young (18-24)	30.7%	24.1%	42.4%	12.5%
Older (65+)	71.7%	57.3%	86.8%	52.3%
Hispanic/Latino	50.3%	45.3%	62.9%	32.7%
African-American	55.6%	49.2%	70.6%	34.3%
Lower Income (<$30,000)	67.0%	60.6%	80.7%	46.8%
Higher Income (>$125,000)	38.5%	26.7%	49.7%	15.5%
High School Graduate	21.4%	30.9%	45.4%	6.9%
Graduate Degree	5.6%	16.0%	21.3%	0.3%
Home Owner	81.3%	76.9%	91.0%	67.2%
Renter	68.1%	6.4%	69.4%	5.1%

Table 2: Summary of Hit-Rates for Variables Examined

Variable	*Vendor A*	*Vendor B*	*Either*	*Both*
Young (18-24)	33.6%	40.2%	35. 6%	38.8%
Older (65+)	31.9%	49.4%	35.1%	55.6%
Hispanic/Latino	66.9%	68.8%	63.2%	78.8%
African-American	68.5%	75.4%	66.5%	84.9%
Lower Income (<$30,000)	39.9%	35.7%	35.4%	42.8%
Higher Income (>$125,000)	26.0%	27.0%	25.1%	32.0%
High School Graduate	22.3%	27.1%	25.0%	24.3%
Graduate Degree	7.3%	42.6%	19.1%	11.1%
Home Owner	85.6%	77.2%	76.5%	88.8%
Renter	79.6%	73.8%	78.9%	81.9%

Tables 1 and 2 above show the coverage and hit-rates respectively for each vendor. In general, we can observe that Vendor A showed superior coverage with Vendor B having a superior hit rate. That said, there were exceptions to this overall pattern, but it speaks to differences in how "conservative" each might be in assigning a specific flag. To illustrate the rates shown in the above tables, the "young" flag in the first row had a 30.7% coverage from Vendor A and a 33.6% hit-rate for the same. We can interpret these results to say that Vendor A could identify 30.7% of the actual "young" people identified in the screened sample, while this flag was accurate 33.6% of the time in doing so.

Coverage, as shown in table 1, appears to be somewhat haphazard across variables. That said, combining lists tended to optimize coverage by roughly 30%, with the trade-off being expense and effort. Overall, we can observe that lists are generally better at covering older populations than younger ones, due to the higher quantity of information available for older individuals.

Table 2 shows hit-rates by variable, illustrating which variables are more easily captured by vendors correctly than others. In general we can observe that hit-rates depend both on prevalence, and thus "luck" as well as the ability to model a specific characteristic. For example, "white non-Hispanic" appears to be easier for vendors to correctly assign rather than "high-school graduate" due to racial/ethnic segregation in the US. We can observe that characteristics that are clustered or are associated with predictive surnames tend to have higher hit-rates than those that are dispersed.

Table 3: Distribution of Incorrect Flags

Actual Race/Ethnicity		*Flagged Race/Ethnicity*				
		African-American non-Latino	*Asian non-Latino*	*Latino*	*White non-Latino*	*Other non-Latino*
	African-American non-Latino		17%	20%	39%	*
	Asian non-Latino	2%		7%	7%	*
	Latino	23%	17%		27%	*
	White non-Latino	58%	51%	63%		*
	Other non-Latino	17%	15%	11%	27%	
		100%	*100%*	*100%*	*100%*	

Table 3 shows the distribution of incorrect outcomes for race/ethnicity flags, i.e., what actual race/ethnicity an individual was if the vendor-provided flag was wrong. As shown, outcomes are not random and reflect the nature of segregation in the United States, in addition to hinting towards how vendors predict race/ethnicity. For example, most incorrect African-American non-Latino flags were White non-Latino and vice-versa, due to neither group having predictive surnames. In the above table the "other" outcome was often multi-race.

Table 4: Hit-Rate for "African American non-Latino" Flag by Tract-Level Demographics

Demographic Quartile	*% Below Poverty*	*% African-American or Latino*	*Median Household Income*
Lowest	58.6%	68.3%	70.3%
Second	64.0%	63.2%	69.5%
Third	64.1%	60.2%	61.6%
Top	72.8%	70.7%	59.1%

Overall: 66.5%

We may begin to hypothesize that the social environment can influence the success or failure of vendors in correctly assigning flags. Table 4 shows how the hit-rate for assigning "African-American non-Latino" varies based on the nature of the Census tract an individual resides in, specifically based on the proportion below poverty, the proportion African-American or Latino, and the median household income. For example, the first row shows that for the lowest quartile of proportion below poverty the hit-rate for "African American non-Latino" it was 58.6%. As shown, flags were more successful in the lowest income/high-poverty tracts, due to the correlation with our outcome variable (in this case proportion African-American non-Latino). Vendors would tend to be "luckier" in such environments.

Table 5: Hit-Rate for "Latino" Flag by Tract-Level Demographics

Demographic Quartile	*% Below Poverty*	*% African American or Latino*	*Median Household Income*
Lowest	65.4%	62.9%	60.7%
Second	63.5%	66.7%	70.4%
Third	62.1%	61.0%	60.1%
Top	62.1%	62.8%	61.4%

Overall: 63.2%

The same isn't the case for those flagged as "Latino" as shown in table 5, with more success in tracts of moderate income and proportion African-American non-Latino. Again such outcomes could be due to "luck", meaning there are more of the target population in tracts of moderate income and proportion minority.

Table 6: Predicting Latino Hit Rate (Precision)

Variable	*β*	*Odds Ratio*
Highly Latino	.52	1.7
High Poverty	-.34	.72
Moderate Poverty	-.26	.77
Mod Poverty; High Minority	1.6	3.2
Urban; High Poverty	.81	2.2
Urban; High Minority	.66	2.0
Low Poverty; Mod Minority	.41	1.5
Non-Urban; High Poverty	-.49	.62

Overall: 63.2%

Table 6 shows the results of a regression analysis predicting the hit-rate of being correctly flagged as "Latino", with the first three rows showing single variable outcomes and the remainder showing interactions. In all instances the reference category is the "lowest" for each variable, and so a tract in the top quartile for "proportion Latino" would be compared to a tract in the bottom quartile for the same. As indicated in table 6 the most successful tracts for being correctly flagged as "Latino" would be of moderate poverty and high minority, with the least successful non-urban, high-poverty tracts. Areas of highest incidence are not-necessarily the most successful, however, pointing towards a tension between effective construction of marketing databases and background density or “luck”.

5. Conclusions and next steps

Our preliminary research has begun to further explore the nature of targeted lists for identifying rare or hard-to-reach populations in social science research. We were able to use multiple concurrent sources successfully to increase coverage, while finding that individual lists may be more successful for specific purposes. We did find that flags may have secondary predictive powers, meaning if they incorrectly flag a specific population of interest they may be more likely to identify another population of interest. Such data could be used in sample designs if one were able to understand better the nature of respondents who originate from a particular source, as well as the impact on effective sample size. Looking ahead we will be enhancing our model through the integration of the Census Planning Data Base (PDB) as well as studying characteristics of respondents who are present or absent from specific lists.

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.

Harter, R., M. P. Battaglia, T. D. Buskirk, D. A. Dillman, N. English, M. Fahimi, M. R. 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. 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.