Using Targeted Lists for Studies of Rare Populations: The Super Wealthy

Ned English[1], Steven Pedlow[1], Lee Fiorio[1], Catherine Haggerty[1]
Benjamin I. Page[2], Jay Seawright[2]
[1]NORC at the University of Chicago, 55 E. Monroe St, Chicago, IL, 60603
[2]Northwestern University, Scott Hall, 601 University Place, Evanston, IL, 60208

Abstract

The Survey of Economically Successful Americans (SESA) is designed to understand the influence of exceptionally wealthy individuals on the American political process. Our pilot study had the goal of targeting the top 1/10 of 1% of households, estimated at $20-40 million in net-worth. One challenge was the absence of a sampling frame that efficiently captured such high-wealth individuals, and limitations in publically-available sources such as the American Community Survey. We created a composite frame from market-research sources, including lists of business executives and "wealthy" individuals. Most sources carried limitations in data resolution i.e. top-coding, as well as inconsistent accuracy. Our current research uses external data sources to enhance our results with the goal of improving both the coverage and hit-rate of our methodology. Examples of newly available data sources include estimates of total liquid assets, home value, and stock-sales that were not available during our pilot phase design. We present models that outline the most efficient approach for conducting nationally-representative household surveys of very wealthy populations.

Key Words: Hard to reach, frame construction, list matching, survey methodology

1. Introduction

There has been recent interest in the theme of wealth disparities in Sociology and Political Science, with a focus on the social views and potential influence of very wealthy individuals (Piketty and Saez 2003). For example, considerable media attention has been given to how wealthy individuals such as George Soros, Sheldon Adelson, and David and Charles Koch might influence political campaigns and results during the recent 2012 US national election. At the same time we have witnessed a general increase in the difference between the top and bottom income tiers in the U.S.; in 1950 0.1% of income-earners had 3.2% of total income, while by 1995 the same percentage of income-earners had 11% of total income (Bartels 2008). The intersection of wealth and elections is therefore a timely issue.

At question is how one can measure the political and civic engagement, attitudes, and opinions of very wealthy Americans about issues facing all Americans through an academic survey. The Survey of Economically Successful Americans, or "SESA", has been designed to consider such themes, including how very wealthy Americans view government programs, markets, equal opportunity, and even the meaning of "America"

(English et al. 2012). The SESA Pilot Study of 2011 targeted very wealthy households in these respects, as measured by total net worth. Specifically, our original intention was to have half of our completed interviews be with households worth of at least $20 to $40 million[1], representing the top 1/10 of 1%, and the other half optimally worth at least $5 million, roughly representing the top 1% to allow for comparison. However, certain challenges specific to very wealthy households prevented such a straightforward study design.

First, very wealthy households by our definition of a small fraction of total households are found at low incidences even in the most concentrated parts of the United States. Second, the prevalence of multiple residences and gate-keepers and can reduce cooperation even when eligible households are found. To complicate matters further, "wealth" itself is a difficult concept to measure, requiring numerous questionnaire items (savings, home value, retirement, other assets, etc.) that are not always available from controls such as the American Community Survey. Consequently, the possibility of geographic stratification is limited, thus raising the question as to the availability of alternative list frames.

This paper details the experience of the pilot test for the SESA project as it relates to targeting very wealthy households. In so doing we describe the construction of a composite list targeting very wealthy households. Second, we discuss results of regression modeling to predict wealth using variables from our list frame, the American Community Survey, and other extant data. Third, we describe lessons-learned to inform our national survey design in the near future. Our paper applies directly to those researchers interested in targeting rare populations using composite lists.

2. Background

The Survey of Consumer Finances (SCF), conducted for the Federal Reserve Board by NORC at the University of Chicago, does target similarly wealthy households by using a proprietary list provided by the U.S. Internal Revenue Service (Johnson 2013, Haggerty and Kennickell, 2012). Because the IRS-provided list may not be used for any other purpose, our study needed to know what for-profit list and telephone sample vendors could provide. Such compilers can create household lists containing demographic "flags", including age, gender, and race-ethnicity. Most demographic information is modeled, using information related to surnames, linked consumer activity, and Census data (Bilgen et al., 2012, English et al., 2012).

One could theoretically use lists of targeted households to enrich a global housing unit list, with disproportionate selection depending on inclusion or exclusion in the lists. While we have previously evaluated using proprietary lists to target members of race/ethnicity or age groups as cited above at NORC, we had not done so for very wealthy households. Such an approach would involve licensing lists from vendors containing likely households, and then using probabilistic matching software to merge variables of interest with a list of all addresses in an area. For example, NORC licenses a

[1] Different sources (Chicago Tribune, something else) have indicated the minimum threshold of wealth for the top 1/10 of 1% to be somewhere between $20 and $40 million.

version of the United States Postal Service Delivery Sequence file (DSF, CDS, or CDSF) from the Valassis vendor, essentially representing all mailable addresses in the United States. The advantage of using targeted lists for stratification as opposed to as a primary frame is to limit the potential for inherent coverage biases in depending solely on the list frame.

3. Methods

Our initial pilot design was based on using a list of very wealthy households provided by the InfoUSA vendor, known as "WealthFinder". "WealthFinder" is designed to rank the net worth of each known household in the United States by modeling attributes such as income, investment activity, philanthropic behavior, and other behavioral and lifestyle characteristics. The top category on the WealthFinder index, known as "rank A", contains households estimated to have a median net worth of $2.695 million; this subset of households was the basis for much of our frame construction for SESA. We then geocoded all WealthFinder Rank A households and selected a subset of them from four affluent pilot study areas in the Chicago Metropolitan Statistical area, these being the towns of Hinsdale, Lake Forest, and Winnetka, and the "Streeterville" neighborhood in the City of Chicago. Finally, we scored all households based on a regression equation and fielded a sample of 200 cases which had the highest estimated "wealth" scores, based on household characteristics such as income.

It became clear during production that limitations in the source-data impeded sample efficiency and thus necessitated an alternate approach. Two limitations reduced our ability to isolate the most-wealthy households present on WealthFinder; annual household income was both imperfect and top-coded at $500,000, and home values were top-coded at $5,000,000. While both measures are indicative of high net worth households, the caps of income and home value are too low to efficiently isolate those households likely to be in the top 1/10 of 1% in net worth. We needed uncapped income and home value estimates. Secondly, the models used by InfoUSA to create the list of WealthFinder rank A households are proprietary, which limited our ability to acquire parameters that might have enhanced our approach. In a sense, we were relying on the rank A rating of households having high coverage and high accuracy, but to WealthFinder, high coverage is more important than high accuracy since clients would like to reach as many wealthy households as possible, even if some less wealthy households are contacted.

Such challenges associated with the original pilot resulted in a revised design, based on refining the WealthFinder rank A list with additional data. The first additional data source is known as "ExecuReach", also provided by InfoUSA. ExecuReach is a database of business executives with their home address, containing information about their title and their firm's number of employees and sales volume. Second, we acquired an additional source of estimated home value from Marketing Systems Group (MSG); while this home value was also top-coded at $5,000,000, we felt that having two separate sources would be helpful. The third source of enrichment was an estimate of household income-producing assets, also provided by MSG. Income-producing assets would describe any source of investment income, including stocks, bonds, bank accounts, certificates of deposit, and mutual funds; it was top-coded at $2 million. We then matched the relevant fields from the three new data sets to our initial WealthFinder list. Following enrichment, we fielded households as described in Table 1 below, with those households considered

likelier to be in the targeted wealth category selected at higher rates. Of the 472 sampled cases, 85 resulted in a completed questionnaire with a non-missing wealth estimate.

Table 1- Sample Design for Revised Pilot

Source	*Description*	*Frame Size*	*Selected/ Complete*
WealthFinder Only	Rank A only; household income at least $500,000; home value at least $1,000,000 and income-producing assets at least $2,000,000	24,661	336 56
WealthFinder in ExecuReach	WealthFinder (as above) and in ExecuReach with "top" titles and sales	75	75 20
WealthFinder in ExecuReach	WealthFinder (as above) and in ExecuReach without top condition	1237	6 0
ExecuReach Only	ExecuReach with "top" titles and sales not in WealthFinder Rank A	55	55 9
ExecuReach Only	ExecuReach without top condition and not in WealthFinder Rank A	1704	0 0
		27,732	*472* *85*

At this point, we were interested in determining whether additional data, either from publicly available sources or from market research companies, could be used to identify high-wealth households or, alternatively, allow us to isolate low-wealth households on the frame. For the 85 completed cases, we pursued data from a number of sources, including Security and Exchange Commission filings; job title and company information from Manta.com; and non-top-coded home values from Zillow.com. At the same time, we were made aware of two additional market research data files from InfoGroup, an Automated Valuation Model (AVM) for property values and a Total Liquid Assets (TLA) index. We also selected a number of relevant variables from the American Community Survey including the median household income of the tract, the median number of rooms in housing units contained by the tract, the percent of households in the tract receiving at least some income from dividends, and the percent of households in the tract receiving at least some income from retirement funds. Together with the frame source indicators, the new variables listed above were merged with questionnaire-derived reported wealth to create a dataset for the purposes of modeling reported wealth. All variables contained in the dataset can be found in Table 2.

Given that we were working with a limited number of observations, we decided to take a varied approach to model selection and thereby attempted to identify the most predictive and influential variables across multiple modeling approaches. Using the logarithm (log) of reported wealth as our dependent variable, we fit two generalized linear models. The first strategy was to use stepwise regression, while the second strategy was a full model including all variables. Next, we fit a classification tree to predict whether a respondent would be in the target wealth category of $20 million or more. The classification tree was created using Recursive Partitioning and Regression Tree (RPART) methodology implemented in the R statistical package. The RPART package in R returns only binary trees (i.e., no more than two children per node), where each leaf node represents a decision based on a predictor of membership according to the conditions represented in the internal nodes all the way back to the root.

Figure 1. Histogram of Reported Wealth

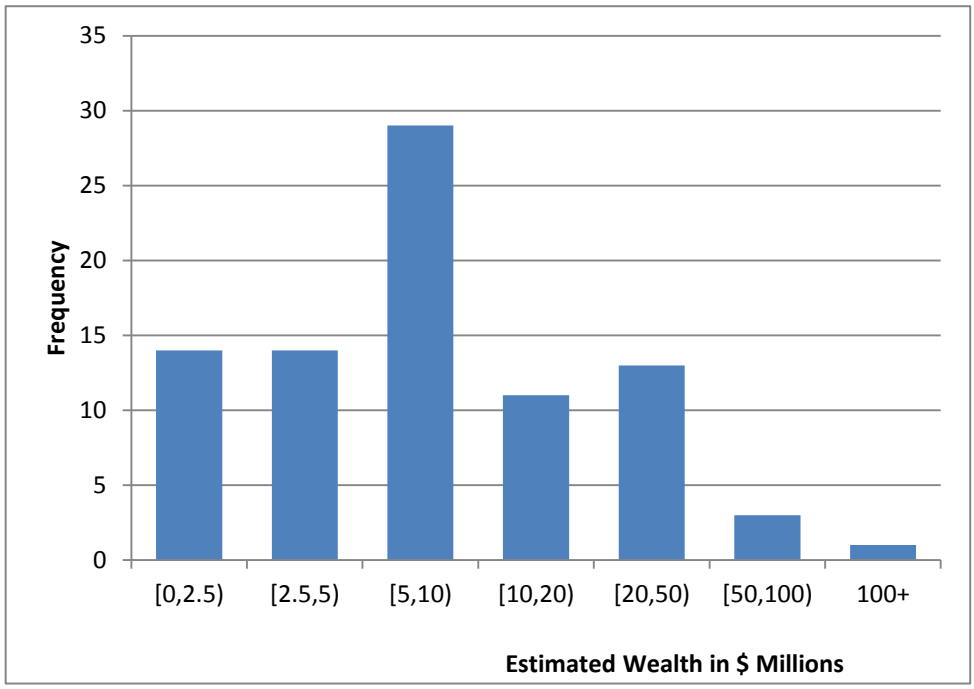


Table 2- Summary of Variables used in Regressions

Variable	Type	Source	Description
Target_wealth	Categorical	Questionnaire	1 if final reported wealth greater than or equal to 20 mil; 0 if not
log_wealth	Continuous	Questionnaire	log of final wealth reported
ER_Only	Categorical	Frame	1 if matched to ExecuReach and not to WealthFinder; 0 if not; reference =0
ER_and_WF	Categorical	Frame	1 if matched to ExecuReach and also to WealthFinder; 0 if not; reference =0
log_zillow	Continuous	Auxiliary	log of Zillow Home Value
manta_com_result__y_n	Categorical	Auxiliary	1 if Manta.com result returned; 0 if not; reference = 0
SEC_filing__y_n	Categorical	Auxiliary	1 if SEC filing found; 0 if not; reference = 0
log_AVM	Continuous	Auxiliary	log of Automated Value Model (from InfoUSA)
log_TLA	Continuous	Auxiliary	log of Total Liquid Assets (from InfoUSA)
medianhhincome_100k	Continuous	Census	Median HH income of census tract represented in $100,000s
medhhrooms	Continuous	Census	Median rooms in households of census tract
pcthhdividends	Continuous	Census	Percent of households in census tract that receive some income from investments or dividends
pcthhretirement	Continuous	Census	Percent of households in census tract that receive some income from retirement

4. Results

We completed 104 total interviews during our original and revised pilot, achieving a net worth estimate for 85 households. Figure 1 shows the distribution of final wealth for the 85 cases with wealth estimates; while most were below $20 million in net worth, 17 were at or above that threshold. In interpreting Figure 1, we may ask two questions related to households in different wealth categories. First, we would like to know how the different frame sources listed in Table 1 performed alone and in comparison with each other. It would be valuable to know the differential efficiency of each list, as well as the kinds of households that tend to be included by each component. Second, we would like to know what factors might be the best predictors of net worth, both at the area level from the American Community Survey and from our various ancillary data sets. It would be encouraging to know, for example, that it were possible to predict very wealthy

households from publicly-available area information a priori instead of having to rely on vendor-provided lists with known coverage limitations.

Table 3 presents results comparing the different frame sources in terms of their effectiveness. In the below table rates for ER only and the intersection of WealthFinder and ExecuReach were statistically different at the $p < .05$ level.

Table 3- Results by Frame for Pilot Study

Source	*High Wealth*[2]	*%*	*Low Wealth*	*%*	*Total*
WealthFinder Only	10	18%	46	82%	56
ExecuReach Only	0	0%	9	100%	9
Both	7	35%	13	65%	20
Total	*17*	*20%*	*68*	*80%*	*85*

As described above, we constructed three models to identify the most significant and explanatory covariates of wealth. Straightaway, we determined variables derived from Manta.com and SEC filing data o not be helpful in our first round of modeling -- including both variables resulted in a singular matrix and non-unique estimates. For this reason, we excluded manta.com from all models going forward. Also because of the high cost of AVM data as well as its high correlation with Zillow home values, we excluded log AVM from all three models. Some missingness in the remaining independent variables reduced the number of useable observations. TLA had a particularly low match rate of 60% for the 85 cases with non-missing wealth. Consequently, we used AVM, Zillow, and respondent wealth data to model the missing TLA values. This brought the number of useable observations for our models to 71, 16 of which are over $20 million in net worth.

Significant variables from the three modeling approaches are presented in Table 4. The log of the Total Liquid Assets index (log TLA) was the most important variable in both generalized linear models followed by the frame type (as shown by the ExecuReach in WealthFinder intersection flag and the ExecuReach Only flag. Households with higher TLA and that are on both ExecuReach and WealthFinder lists were generally wealthier than their counterparts in our dataset. The stepwise model also identified the log Zillow variable and the tract-level ACS variable on percentage of households with income coming from dividends as significant variables, but these variables were not significant in the full model.

The decision tree offers us a slightly different perspective than the regressions, as illustrated in Figure 2. The first split in Figure 2 is between respondents with a log of their Zillow home value at or above 14.2, which corresponds with approximately $1.5 million. A much higher percentage of respondents with valuable homes are also in our target wealth category of $20 million or more. Within those with more valuable homes, we can further isolate our target group by examining tract-level ACS data on the percentage of households receiving income from dividends. Within our dataset, nine out of ten respondents living in homes worth more than $1.5 million in tracts where 60% or

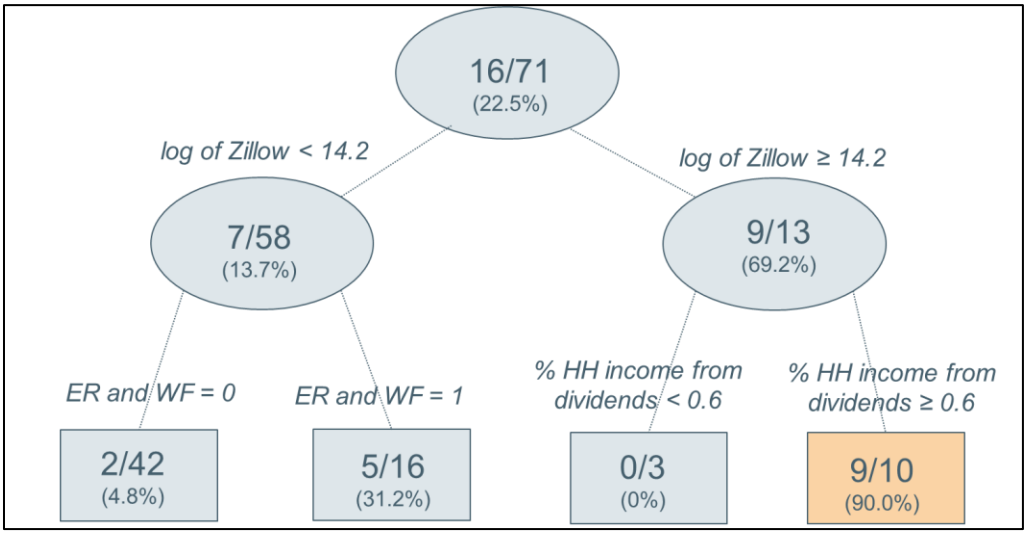
[2] We define "high" wealth for purposes in Table 3 as >= $20,000,000

more of the households receive some income from dividends were also in our target category. On the other branch of the tree, additional information becomes relevant. Of those living in homes valued at less than $1.5 million, only two out of forty-two are in the target category if they are also missing from either the ExecuReach and WealthFinder lists.

Table 4 Summary of Regression Results[3]

Stepwise GLM (In Selection Order)	*Full Model (significant variables)*	*Classification Tree*
Log Total Liquid Assets ***	Log Total Liquid Assets ***	
In ExecuReach and WealthFinder **	In ExecuReach and WealthFinder **	In ExecuReach and WealthFinder
Log Zillow Home Value **		Log Zillow Home Value
ExecuReach Only ***	ExecuReach Only **	
% Household Income from Dividends °		% Household Income from Dividends

Figure 1- Classification Tree of Households with At Least $20 Million in Net Worth



5. Discussion and Conclusions

It is important to note a few caveats with our analysis thus far. First, our discussion is based on relatively few interviews in specific locations in metropolitan Chicago, Illinois. Interviewing very wealthy households is challenging, and we would expect the median

[3] *° p < 0.1, * p < 0.05, ** p < 0.01, *** p <0.001*

wealth to increase with response-rate. So, we view our pre-test early respondents as likely to be less-wealthy than the remaining non-interviewed cases. In addition, it is important to note that many of our independent variables were top-coded, and so our results might have been different in an environment with non-top-coded income, for example. It is true that not having a single list for executing the study requires the creation of a composite, and there are likely several options from multiple vendors in addition to the ones we examined. We did find that the most efficient sample included those cases that were present on multiple lists, with the highest home-value and presence of dividend income being the most important.

Our pre-test has been able to inform the ultimate design of our national study, which we expect will be a nested area-probability design with some implementation of composite targeted lists.

References

Bartels, Larry M. 2008. Unequal Democracy: The Political Economy of the New Gilded Age. Princeton, N.J.: Princeton University Press.

Bilgen, Ipek, Ned English, and Lee Fiorio. Coverage and Data Quality Association in Enhanced Address-Based Sample Frames. 2012 Proceedings of the American Statistical Association, Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

English, Ned, Ipek Bilgen, and Lee Fiorio. Coverage Implications of Targeted Lists for Rare Populations. 2012 Proceedings of the American Statistical Association, Survey Research Methods Section [CD ROM], Alexandria, VA: American Statistical Association.

Haggerty, Catherine and Arthur Kennickell. 2012. *The Survey of Consumer Finances: Creating a Consistent Culture of Quality*. Proceedings of the First International Conference for Surveying and Enumerating Hard-to-Reach Populations, October 31-November 3rd, 2012, New Orleans, LA.. Available online at http://www.eventscribe.com/2012/ASAH2R/assets/pdf/49950.pdf.

Johnson, Barry. 2013. *An Enduring Partnership: Incorporating Administrative Data Into Sample Design for the Survey of Consumer Finances*. Presented at the Joint Statistical Meetings, August 2013, Montreal, Canada.

Piketty, Thomas, and Emmanuel Saez. 2003. *Income Inequality in the United States, 1913-1998*. Quarterly Journal of Economics 118 (1): 1-39.