From Flagging a Sample to Framing It:
Exploring Vendor Data that can be appended to ABS samples

Trent D. Buskirk[1], David Malarek[1], Jeffrey S. Bareham[1]
[1]Marketing Systems Group, 755 Business Center Drive, Suite 200, Horsham, PA 19044

Abstract

Address based sampling and random digit dialing designs are two of the most commonly used probability based sampling methods to obtain data aimed at answering questions in a broad spectrum of fields including social science, public health, and medicine and beyond. Enhancements to samples can be made in order to facilitate better frame construction including the appending of information on home ownership or presence of children or age of head of household to create strata that could be meaningful for a given study. These variables can also be used as part of a tailored design or in part for nonresponse and weighting adjustments after data collection is completed. The degree to which these appended variables can be used for various aspects of the survey process from sampling, to recruitment to weighting depends in part on their availability for all sampled or population units and on their accuracy at the time of appending. In this presentation we provide a broad overview of the types of variables that can be appended to address based samples with particular focus on vendor related, rather than public use, variables. We describe appending rates for a key set of socioeconomic and household variables, and discuss their variability across vendors. We discuss how survey researchers can properly evaluate the quality of these data and make recommendations about their potential uses for ABS sampling designs.

Key Words: RDD, Dual Frame Samples, Address-based samples, Vendor Data, Appended Data, Frame Variables, Record Linkage

1. Introduction

Address based sampling offers survey researchers with an enormous array of household, census and other related variables that can be linked to each sampled address via their geo-location coordinates (i.e. latitude and longitude), their zip code or census block group. Several papers in the research literature have explored the various uses of census related variables that are appended to address based samples. The uses of these appended variables have ranged from tailored survey design and *a priori* response propensity grouping and prediction (Link and Burks, 2013 and Buskirk, et al., 2014) to nonresponse bias evaluation and adjustment (Biemer and Peytchev, 2012, 2013). Some work has also used a combination of census-related variables along with variables available from proprietary sources such as Claritas or Nielsen in addition to variables available from various Address Based Sample vendors such as Marketing Systems Group, Valassis and Survey Sampling International (Burks, et al., 2012). For clarity, we will refer to these three types of appended variables as CS (appended from CensuS), PS (appended from Proprietary Source or Service) and SV (appended from sample vendor). Generally speaking, CS variables that are appended to addresses from ABS samples refer to block-group statistics or larger levels of geography; PS data appended to addresses within ABS samples represent statistics aggregated at either the zip code level, the zip+4 level or in some cases the zip+6 level, but in some cases though, PS variables may in fact represent household level measures directly. SV data appended to addresses within an ABS sample generally refers to statistics or characteristics of the dwelling unit corresponding to the address that was sampled. The SV can also source appended variables from vendor partners (VPs).

While each of these three types of appended variables have their merit in propensity models, stratification, tailoring for ABS sample designs, our focus in this paper will be on the SV related variables. In section two we describe ABS sampling frames and describe the composition of each included record as well as discuss specific types of SV variables that can be appended to addresses sampled from an ABS sample frame. In section 3 we will discuss the overall append rates for these variables across three data vendors and then explore variability in append rates for a core set of demographic variables as a function of household and census block-group specific variables. We conclude the paper with a discussion about how survey researchers can evaluate the accuracy of appended data based on the findings of this study.

2. Background on Address Based Samples and Variable Appending

2.1 Overview of Address Based Sampling Frames and Records

Address based sampling designs offer several advantages to researchers primarily in the areas of population coverage and the availability of frame data attached to address based sampling frames (Link, 2010). However, the ABS sampling frames are certainly not a panacea for survey researchers. Iannachione, et al. (2010) noted that the ABS frame may not have adequate coverage for areas with new construction, highly rural areas and group quarters. ABS designs may also not be compatible with shorter data collection field periods and in some cases, costs can be more expensive per complete compared to other survey modes. In any case, the effectiveness of the ABS methodology hinges on the availability of a master database of (address) records that provide adequate coverage of a target geography such as a city, state or nation (Link, 2010).

In general, in the U.S., the master file for ABS sampling frames constructed from sample vendors comes from the United States Postal Service's (USPS) Computerized Delivery Sequence (CDS) file. The CDS file consists of over 151 million residential and business addresses and is updated regularly by the USPS and these updates are made available at regular intervals to licensed vendors like MSG. In addition to the CDS file, sample vendors can further refine ABS sampling frames by adding information contained in the USPS NOSTAT file which contains approximately 10 million additional inactive address listings. While these inactive records cannot receive mail, their presence in a frame could assist survey researchers with field enumerations or other listings required by an address-based sampling design. Sample vendors can also add simplified addresses as well sourced from other third party vendors.

Variables available for each address in ABS sampling frames are described in more detail elsewhere (Dohrmann et al., 2014) but generally consist of the physical address, route type, drop code, drop count, seasonal code, dwelling type and vacant and only way to get mail flags. For more specific details on any of these variables consult the USPS CDS user guide (available: https://ribbs.usps.gov/cds/documents/tech_guides/CDS_USER_GUIDE.PDF). While many of these frame variables are helpful for sorting and processing ABS samples (e.g. route type gives information on whether or not the address is a PO BOX, a Rural Route, etc.), the key to linking addresses to data available from the U.S. Census relies on additional geocoding of addresses in the sampling frame to append the census tract, block and block group. Attaching these additional geographical variables to each address on the frame allow researchers a whole host of additional variables that can be appended for stratification, weighting adjustments or tailoring (see Burks and Buskirk, 2012; Buskirk et al., 2013; Link and Burks, 2013 or Biemer and Peytchev, 2013).

2.2 Appending Sample Vendor Related Variables

While the number of variables that can be appended to each address in an ABS sample via the census geography key variables is extensive, most of information contained in these types of variables resides at the census block level. Household level and within household level can be directly appended by SVs based on their own internal data or data sourced from one or more VPs. The number and nature of these household level variables span from demographic to behavioral. Generally when these types of variables are appended to an address from an ABS sample the results represent either the actual appended data (or category in the case of categorical variables) or a blank/missing value. The missing values are important to understand as they convey two types of information: either the information was not available for a given address (i.e. append failed) or the particular household does not have a given attribute. So for example, if interest was given to appending Hispanic ethnicity, the appended variable would likely assume the levels of 1 – indicating Hispanic ethnicity or Blank/Missing – indicating either no Hispanic ethnicity information available or NOT of Hispanic Ethnicity. We also remark that personal-level information such as gender and education are provided for a primary reference person (PRP) that is identified separately for each VP data source. So if a SV uses multiple VPs as data sources for appending, it is possible that personal demographic information from one VP source will not match that of another VP – even for the same address because each different VP identified a different PRP for the same address. This is especially likely for addresses that have more than one adult, for example. Differences in other household level variables such as own/rent, income, number of adults in the household or presence of children may also not match across appending data sources because different SVs and VPs update their databases at different intervals and with potentially different frequencies for each variable.

3. Objectives and Sample Data

This study had three specific goals related to appended information available from a SV and their VPs including: (a) estimating the append rates for specific demographic variables at the person and household level; (b) determine the degree of consistency of the appended information across three VP data sources and (c) to determine variables related to variability in append rates for addresses within an ABS sampling frame.

To accomplish these objectives, we randomly selected an ABS sample of 1,000,000 residential addresses from the SV's ABS Residential Frame. The sample vendor had three vendor partner sources available for demographic data that could be appended to addresses in the frame. We attempted to append each of the variables listed in Table 1 to each sampled address from each of the three VP data sources. More explicitly, for each variable in Table 1 we created three versions of appended information depicting the results of appending from VP1, VP2 and VP3.

Table 1: Partial Listing of Specific Demographic Variables that can be appended to an ABS sampled record

Annual Household Income *
Presence of Adults in HH 65+ *
Gender of PRP *
Number of Adults in HH *
Presence of Adults in HH ages 18 to 24 *
Presence of Adults in HH ages 25 to 34 *
Presence of Adults in HH ages 35 to 64 *
Surname of PRP *
Given Name of PRP*

Age of PRP
Marital Status
Education of PRP
Ethnicity of PRP
Presence of Children in HH
Own / Rent
Ages of Additional Adults in HH
Telephone Number

PRP= Primary Reference Person for the Household. Some vendors refer to this adult as the "Head of Household"
** indicates that the variable was part of a common core of demographic variables that were all appended together (i.e. all or none) from the three vendor partners.

To facilitate analyses of the stated objectives, we created an analytic data file comprised of a total of 983,186 addresses. Specifically, from the 1,000,000 sampled records 12,998 on rural routes or identified as the only way to get mail were removed because none of these had appended information on any of the variables for which append attempts were made (i.e. constant append rate); an additional 3,013 sampled addresses were identified as drop points that were revised/updated or expanded by the SV during frame construction and this expansion used VP appended information *a priori* and would have artificially biased upward the measure of append rates and confounded the signal in the drop indicator predictor; finally, we removed another 805 cases that exhibited anomalous and inconsistent vendor append data.

For each of the variables listed in table 1 we measured the append rates from each of the VPs using append results from each of the addresses in the analytic file. Initially for this study we anticipated being able to model append rates for each VP source using some type of Poisson regression model or similar, but after more careful examination of append rates across the three VP data sources we discovered that the appending frequencies were 0, 9, 10, … 14 (representing the first 14 variables in Table 1). The 9 variables that were matched together across the three VP databases are denoted by asterisks in Table 1. We refer to these as "the common core". We also measure the distribution of the appended information for a subset of the common core across the three VPs. Data from the 2013 CPS serve as a reference point for these comparisons.

To understand the consistency of appended information across the three SV's VP data sources, we computed concordance measures for each of the 9 common core variables as well as for phone number, presence of children, home ownership, middle initial and marital status using addresses that had information for the specific variable appended from at least two of the three VP data sources. More specifically, for each sampled record in the analytic data file that had information appended from at least two of the three VP data sources for a given variable, concordance for this record on this variable was assigned one of three possible outcomes: 1 – if appended information from all three VPs matched; 2 – if information for the variable from two of three VPs matched and 3 – if no information from any of the VP sources matched for the variable. To clarify, addresses that had information from a given variable appended from only two VPs were categorized as a level 2 concordance if both sources matched and 3 if neither matched. We note that it was possible for addresses to have missing information from the core variables from one vendor or another since the common core append rates varied across the three VP sources. In total, there were approximately 730K addresses used for computing concordance rates for the common core set of variables. The number of addresses used for estimating concordance for the other variables varied from about 132K to 684K as the underlying append rates for the non-common core variables varied and were generally smaller across the three VPs. For the purposes of this study we define the **concordance rate** (cr) for a given variable to be the *proportion of the addresses with a concordance rating of at least 2 among all addresses for which a concordance value was computed for that variable.*

For the final objective, we explore the relationship between variables available for every record from the SV and the CDS file and selected census block group variables thought to be associated with information availability and whether or not the common core set of variables was appended from VP1 using a conditional random forest. Conditional random forests generate unbiased measures of variable importance (Strobl et al., 2007) and require two parameters – the number of trees used in the forest and the number of variables selected randomly at each tree node split. In this application, we used 250 trees and the default value of three variables selected per node split.

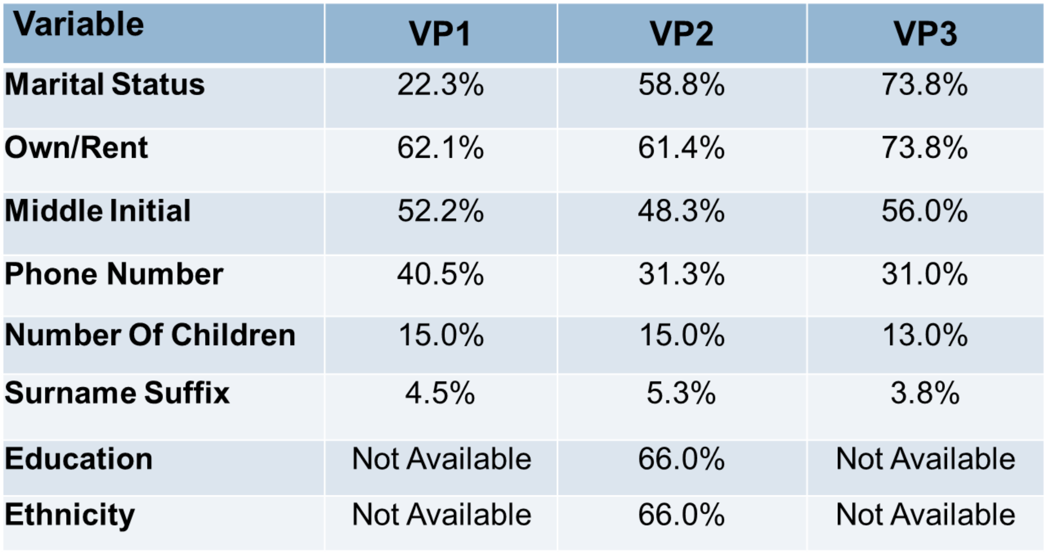
A total of 17 predictor variables were used in this investigation and the forests were constructed using a 15% subsample (148,112 cases) of the analytic data file due to computational limitations of the conditional random forest at the time of the analysis. Of the 17 variables considered, 5 came from the CDS file (route type, vacant indicator, seasonal flag and a drop point indicator); 4 came from the SV (census division based on state of address, time zone and suburban ring – a five level urbanicitiy variable and a level of geocoding accuracy) as well as 8 Census block-group variables (percent owner, median household income, average number of cars per household, the median age of the block group population, the median years households have resided in their current home, the percentage of adults in the block group who obtained at least a Bachelor's degree, the percentage of households within the block group with children under 18 in the home, and the percentage of the civilian workforce within the block group who are employed. Measures of importance were derived from the conditional random forests and bivariate effect size measures were computed between each of the predictors and the binary core appended from the VP1 data source variable using eta-squared for the continuous predictors and Cramer's V for the categorical predictors (Bethlehem et al., 2011).

3. Empirical Results

3.1 Append Rates for Demographic Variables and Distributions of Appended Information

The core append rate for the three VP data sources were 74.6%, 66.0% and 73.8% and the core variable. The append rates for variables not in the common core are presented in Table 2. We reiterate that the base of numbers used in generating these estimates was 983,186. There is a fair amount of consistency in append rates for number of children the home and surname suffix and middle initial. The smaller append rates for surname suffixes isn't surprising, though given the prevalence of such suffixes in the general population and that the information. There is also marked variability noted in append rates for marital status and phone number. Optimistically, approximately 40% of sampled addresses had phone matches available on the high end and just under one third on the low end. We note that phone append rates have been declining in part due to increases in the prevalence of cell phone households and the overwhelming predominance of telephone numbers that are appended to addresses are landline telephone numbers.

Table 2: Append Rates for Non-Common Core Variables across Three VP Data Sources



Variable	VP1	VP2	VP3
Marital Status	22.3%	58.8%	73.8%
Own/Rent	62.1%	61.4%	73.8%
Middle Initial	52.2%	48.3%	56.0%
Phone Number	40.5%	31.3%	31.0%
Number Of Children	15.0%	15.0%	13.0%
Surname Suffix	4.5%	5.3%	3.8%
Education	Not Available	66.0%	Not Available
Ethnicity	Not Available	66.0%	Not Available

We compared the distribution of appended information on select household variables across the three VP sources. Due to the sizes of the samples, many of these comparisons were statistically significant (type-I error rate of 5%), even after adjusting for multiple comparisons – however, not all differences are of practical significance (i.e. differences within 1 to 2 percentage points for some levels of categorical variables, for example). We present the results here for reference purposes and leave the evaluation of practical significance to the discretion of the reader. We also include population estimates obtained from the 2013 Current Population Survey (CPS) for reference.

The distribution of the appended number of adults to addresses within the analytic file are fairly consistent across the three VP data sources but don't match the overall distribution of the population estimated by 2013 CPS as shown in Figure 1. We note that the proportion of households with 3 and 4 or more adults from vendor one matches the CPS statistics, but consistently, the appended data overestimate the number of one-person households and underestimate two-person households. We don't have direct hypotheses as to why the overestimation occurs, but indirectly, we suspect this could be related to how vendors link information from multiple reference persons that live within the same households, treating these as two one-person households rather than one two-person household. This could especially be true in households where adults do not share the same surname.

Using the four separate presence of adults in age group variables listed in Table 1 we constructed the overall distributions depicted in Figure 2. In general, the appended presence of adults in the four age group flags reflect national CPS estimates for presence of adults 65 to 80+; however, the degree to which vendor appended distributions match the CPS totals for the other age groups varies by VP source. Systematically, the appended information for the youngest age group (18-24) underestimates the true proportion of households with a member of this age group as estimated by CPS 2013. For this age group, the underestimation is directly tied to how vendors source this information and the fact that many 18-24 year olds within the U.S. do not own their own homes and often are tied to a permanent address as well as a seasonal one like an apartment, dormitory or other similar quarters.

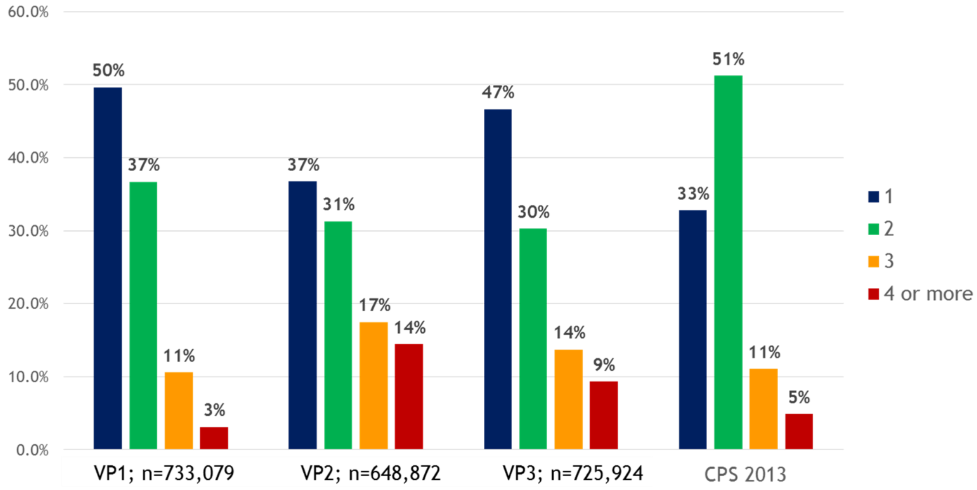


Figure 1: Distribution of the number of adults in the household based on data appended from the three VP data sources with the 2013 CPS distribution included for reference.

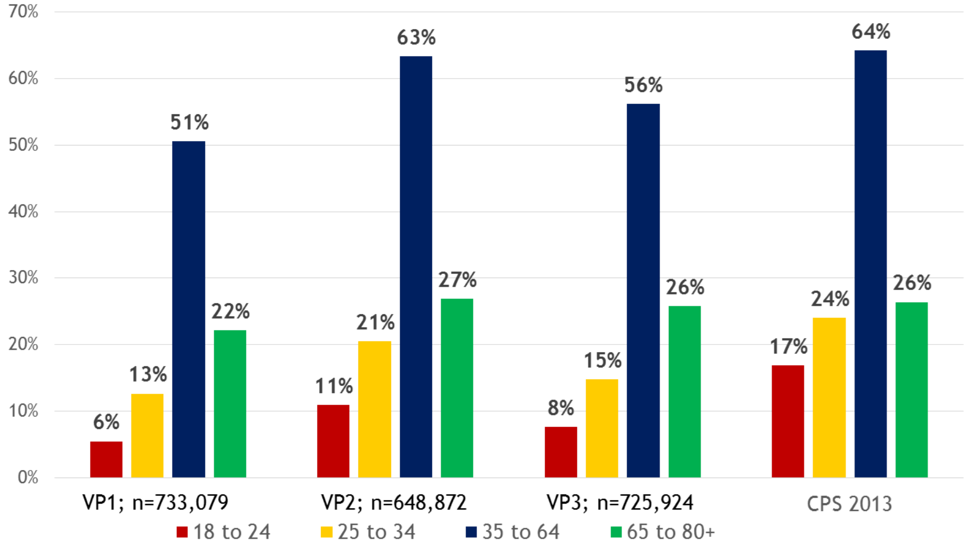


Figure 2: Distribution of the presence of adults in the household within a given age group across for each of the three VP data sources along with the 2013 CPS distribution included for reference. Each bar in the VP graphs represents the data from four separate appended variables corresponding to the four age-specific range variables listed in Table 1.

The distribution of home ownership based on appended data seems to be overestimated consistently across the three TP sources as shown in Figure 3. This result is likely related to the underestimation of adults between 18 and 24 in households as well as an underestimation of the heads of householders between 18 and 34 years of age. If a larger proportion of adults in either of these age ranges are renters, then one might expect the underestimation of renters that is depicted in Figure 3.

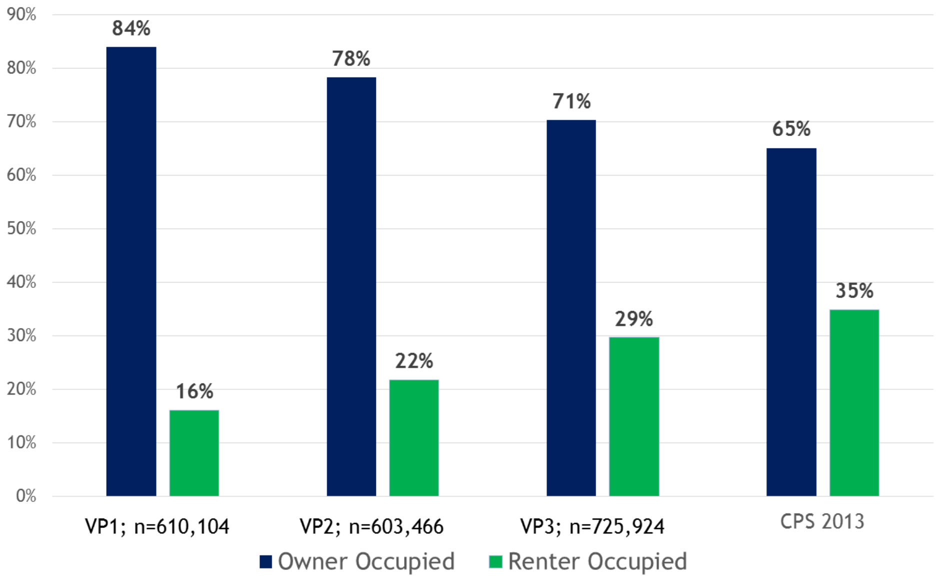


Figure 3: Proportion of owner/renting households based on appended renter status data from each of the three VP sources along with the population proportions estimated from CPS, 2013.

3.2 Concordance of Appended Information across Vendor Partner Sources

Using the concordance measures defined in the previous section, we saw the highest concordance rates for appended phone numbers (cr >97%) and home ownership status (cr > 92%). The concordance rates for the four targeted age grouping variables with the percentage of addresses for which at least two sources matched each exceeded 90%. A related variable that is not shown here is the age of the primary reference person. The match rates for this appended data across the three VP sources hovers at around 80%. While the appended information for the general presence of adults in various age groups had high consistency across the vendors, we saw a bit lower concordance rate of approximately 80% for the actual number of adults within household. Similar concordance statistics were also noted for Gender (of the reference person) and Household Income. Differences in gender can be directly tracked back to households with more than one adult of each sex and two different VP data sources uses each of these two adults as the respective primary reference person for the sampled address.

We saw slightly lower concordance rates for both given name (cr ≈ 79%) and marital status (cr ≈ 78%). Differences in given name across TP data sources could be tracked back to either differences in primary reference person or in name spelling. Longer given names are especially prone to differences across sources as each TP has potentially different tolerances for field lengths and potentially different rules for abbreviating and/or truncating such name fields. The lower concordance rate for marital status was tracked back to a nearly mirrored distribution of married estimates derived from appended data from VP1 and VP2 as shown in Figure 4. We note that the distribution of marital status for VP3 nearly matched the estimates from CPS 2013 but have no explanation for the nearly opposite matches from VP1 and VP2. We note that the operational definitions for married and single were standardized across the data sources prior to appending so one possible explanation for these differences could be related to update frequency or lag time.

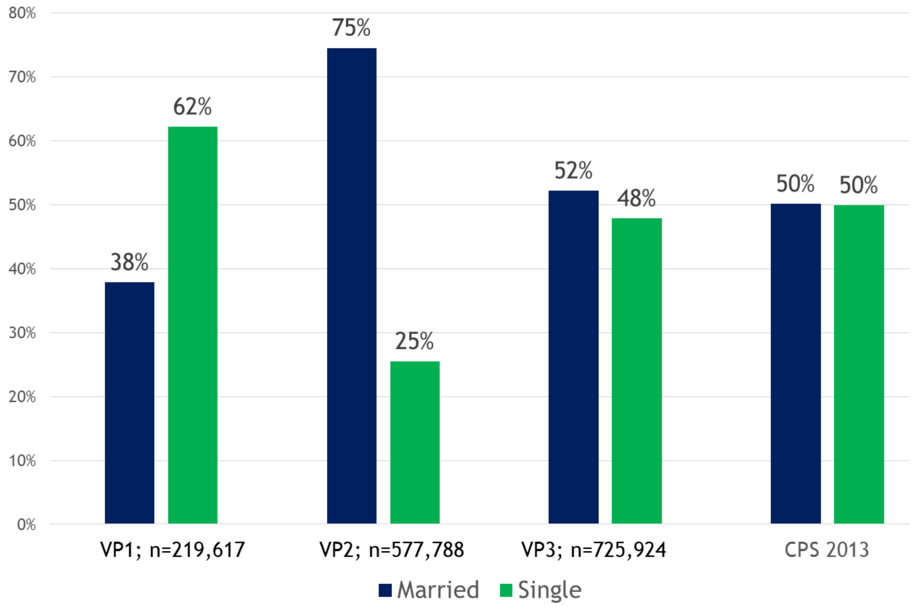


Figure 4: Distribution of married head of householders based on appended data from the three VP data sources along with the overall population proportion estimates derived from the 2013 CPS.

3.3 Variability in Core Appended Variables

Approximately 75% of the addresses in the analytic file had core variables appended from VP1. Factors related to the core append from this particular TP were explored using a conditional random forest model based on 17 census block group, SV and CDS related variables as outlined in section 2. The overall importance measures based on the conditional forest model constructed using a 15% random subsample of the analytic file are given below in Figure 5.

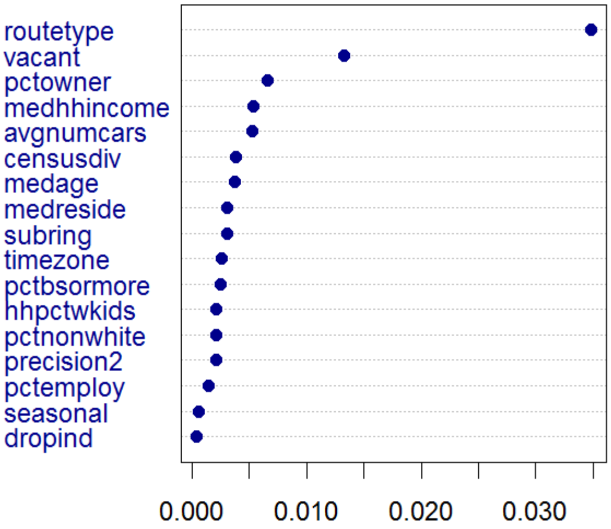
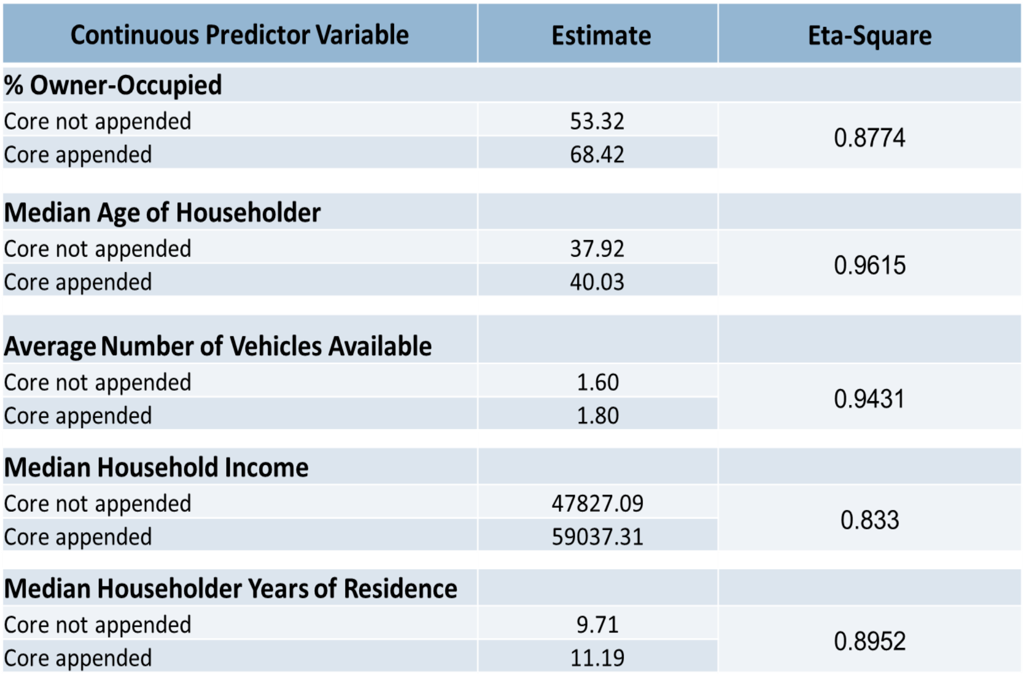


Figure 5: Variable importance measures derived from a conditional random forest model predicting core variable append rates from VP1 as a function of census block group, SV and CDS variables.

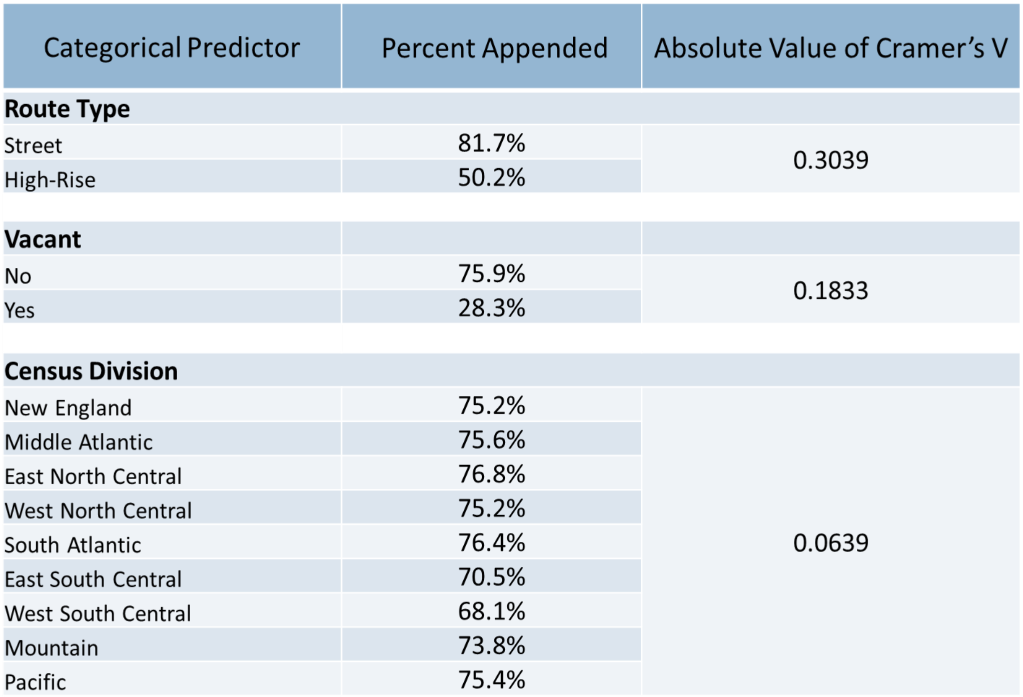
The most influential variables for classifying addresses as having core variables appended or not from VP1 appear to be route type, vacant status as well as the percent of households in the address' block group that are owned, the median household income for the block group and the average number of cars for households within the address' block group. The A bit more exploration of these most influential variables revealed that addresses with core variables appended (from VP1) were in census block groups with a higher percentage of owner occupied units with slightly higher median household incomes and slightly more cars per household. Because we removed rural routes from the analytic file prior to this analysis, the route types used for the conditional forest model were either street addresses or high rise addresses. A closer look at this variable in relation to core append rates for VP1 reveal that those addresses with street routes had an append rate of just over 1.6 times that of addresses with a high-rise route type (82% versus 50% append rates, respectively). High rise addresses may contain a larger share of rental units and so this difference is consistent with the differences we saw with the distribution of owned versus rented households based on appended information relative to population estimates derived from the 2013 CPS (see Figure 3). The summary statistics and effect size measures for addresses with and without core variables appended from VP1 for the 5 most influential continuous predictors are given below in Table 3 followed by the append rates and effect size measures for the top three categorical variables are given in Table 4.

Table 3: Summary statistics for the five continuous variables with the largest importance measures for predicting core variable append rates from VP1 along with effect size measures.



Continuous Predictor Variable	Estimate	Eta-Square
% Owner-Occupied		
Core not appended	53.32	0.8774
Core appended	68.42	
Median Age of Householder		
Core not appended	37.92	0.9615
Core appended	40.03	
Average Number of Vehicles Available		
Core not appended	1.60	0.9431
Core appended	1.80	
Median Household Income		
Core not appended	47827.09	0.833
Core appended	59037.31	
Median Householder Years of Residence		
Core not appended	9.71	0.8952
Core appended	11.19	

Table 4: Summary measures for the three categorical predictors with the highest importance measures for predicting core variable append rates from VP1 along with effect size measures.



Categorical Predictor	Percent Appended	Absolute Value of Cramer's V
Route Type		
Street	81.7%	0.3039
High-Rise	50.2%	
Vacant		
No	75.9%	0.1833
Yes	28.3%	
Census Division		
New England	75.2%	0.0639
Middle Atlantic	75.6%	
East North Central	76.8%	
West North Central	75.2%	
South Atlantic	76.4%	
East South Central	70.5%	
West South Central	68.1%	
Mountain	73.8%	
Pacific	75.4%	

4. Conclusions and Discussion

During our investigation of variable append rates across three vendor data sources we discovered that variables don't append independently of one another. In fact, we identified a common core of 9 demographic variables that tended to be appended in a chunk/all-or-nothing fashion. Moreover, having core variables appended was a necessary, but not sufficient, condition for further variable appends. We noted moderately high append rates for this common core of demographic variables across three sample vendor partner data sources and in general, the proportion of addresses that had the common core of variables appended was between 66% to 75% across the three vendor partner data sources. Differences in append rates for the common core from the first vendor partner source were also observed most notably between dwellings that have street addresses versus those in high rises. Addresses with core appended information were generally located in census block groups with higher median household incomes, higher owner occupied units and in block groups with a slightly higher number of cars per household, on average compared to addresses without the core demographic variables appended.

Beyond the common core of demographic variables, we also noted phone number append rates varied across the three vendor partner data sources and ranged from about one third to about 40 percent for sampled addresses in our analytic file of over 983K records. But where they were able to be appended, the actual phone numbers were highly consistent across the vendor partner data sources. We also noted high concordance rates across the vendor partner data sources for appended information about home ownership, presence of adults in various age groups, and surname. We also observed several differences in the actual data appended from three VP data sources for many variables including gender, age, given name and middle initial. We note that these differences do not necessarily imply inaccuracies in the appended data. Differences in appended data could certainly be represented by inaccuracies coming from administrative sources, but another real possibility that is also likely and has some empirical basis as a result of this study lies in how vendors append information to the household, especially for the primary reference person. In households with more than one adult, for example, it is completely possible for one vendor to refer to one adult in the household as the PRP but a second vendor could refer to another adult in the household as the PRP. In households with a man and a woman, differences in appended gender could easily occur if each of these adults served as the PRP for the two vendors. Any person level variable that may vary for adults within a household also presents with the same possibilities for differences across vendor partner data sources. Age related variables and other variables that one might consider "time varying" might also vary across vendors simply due to differences in variable update frequencies or lag times between updates across vendors.

Taken together this study has implications for researchers who use vendor appended information for sampling planning or evaluation. For planning purposes, the primary variable of interest should be noted – perhaps the variable will serve as a stratification variable or perhaps used for determining eligibility or for tailoring the survey protocols. Whatever the case, the variable or variables should be prioritized and researchers should inquire about the variability in append rates for these key variables across multiple, and sometimes, competing vendor data sources. Knowing these append rates may help determine the order of appending or even more importantly may assist in determining that one vendor source is more feasible for a given study relative to another. Another implication gleaned from this study involves attempting to use independent vendor sources to "fill out" a demographic record for a given sample of addresses. For example, if a researcher desires to know gender, age and religion for a given sample of addresses. If vendor 1 can supply age and gender and vendor 2 can supply religion, why not run the sample of addresses through two append processes? Well, if both vendors use the same PRP, then the append process is likely to yield consistent information for the PRP and household. However, as this study indicated for many person-level variables, discordance across vendors for a given address can sometimes be attributed to differences in PRP for a given address. And so, in the case of our example, it is possible that the gender and age data appended from vendor 1 may in fact refer to the woman of the house whereas the religion appended from vendor 2 may be that of her husband.

The results of this study also have implications for researchers who are cross validating the information contained in vendor appended data with survey data collected in the field. Overall, researchers and practitioners should keep in mind that data from vendor data sources are appended via the address and should generally be thought of as properties of the household unit. In some household studies that have reported vendor flag accuracy, the surveyed adult in the household was randomly selected and information about the sex, education, age and other similar variables was collected and ultimately compared to the appended information. But for a given household, the PRP, upon which the appended information is based, may not be the adult that was selected using the sampling design. Or, in the case that the selected adult is the PRP, information on time varying covariates like education, marital status and age may differ between appended information and reported information because of lags in administrative data updates or because of recent changes in these variables that are more current than the most current administrative update could provide. Perhaps one way to improve the evaluation in these cases would be to create "buffers" around ages (plus or minus one year) and determine consistencies between a buffered appended age and reported information for a roster of adults in the household. Where there are differences in appended versus reported data on variables like education (e.g. appended data indicates B.S. where the reported information indicates M.S.) perhaps an additional question inquiring about educational attainment of other adults in the household, or recent changes in educational attainment for the survey respondent could provide more comprehensive data about the household to use in evaluating the accuracy of the appended data.

As target populations become more specific and as population members become harder to reach or harder to find, survey researchers need better tools to sample and survey members. While certainly not the silver bullet, household and person-level can be appended to ABS samples and is available for a health majority of addresses within residential ABS frames. Variability in append rates across a wider spectrum of demographic variables does exist across vendors and researchers should inquire about these append rates for variables they deem most important for their survey design or sampling plan.

Acknowledgements

We would like to thank Dr. Stanislav Kolenikov for organizing an informative session about the use of vendor related information in the design and analysis process and to Dr. Charles Disogra for chairing it! We would also like to thank Dr. Jill Montaquila for her excellent and insightful discussion of the papers in our session.

References

Bethlehem, J., Cobben, F. and Schouten, B. (2011). Handbook of Nonresponse in Household Surveys. Wiley, Hoboken.

Biemer, P.P. and A. Peytchev (2013). Using Geocoded Census Data for Nonresponse Bias Correction: An Assessment. *Journal of Survey Statistics and Methodology*, 1(1), 24-44.

Burks, A.T. and Buskirk, T. D. (2012). Can Response Propensities Grow on Trees? Exploring Response Propensity Models Based on Random Forests Using Ancillary Data Appended to an ABS Sampling Frame. Paper presented at the 2012 Midwest Association of Public Opinion Research, Chicago, IL. http://www.mapor.org/confdocs/progarchives/mapor_2012.pdf (accessed on March 10, 2014).

Buskirk, T.D., West, B. T. and Burks, A-T (2013) "Respondents: Who Art Thou? Comparing Internal, Temporal, and External Validity of Survey Response Propensity Models Based on Random Forests and Logistic Regression Models," Presented at the 2013 Joint Statistical Meetings of the American Statistical Association, Montreal, Canada.

Dohrmann, S., Buskirk, T.D., Hyon, A. and Montaquila, J. (2014). Address Based Sampling Frames for Beginners. In *JSM Proceedings*, Survey Research Methods Section. Alexandria, VA: American Statistical Association.

Iannacchione, V., K. Morton, J. McMichael, B. Shook-Sa, J. Ridenhour, S. Stolzenberg, D. Bergeron, J. Chromy, and A. Hughes. (2010). "The best of both worlds: a sampling frame based on address-based sampling and field enumeration." In JSM Proceedings, Survey Research Methods Section, Alexandria, VA: American Statistical Association.

Link, M.W., and Burks, A.T. (2013). Leveraging Auxillary Data, Differential Incentives, and Survey Mode to Target Hard-to-Reach Groups in an Address-Based Sample Design. *Public Opinion Quarterly*, 77(3), 696-713.

Strobl, C., Boulesteix, A-L., Zeileis, A. and Hothorn, T. (2007). Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. BMC Bioinformatics, 8(25).