The Quality of Auxiliary Variables in an Enhanced Address-Based Sampling Frame

Rachel Harter
RTI International, P.O. Box 12194, Research Triangle Park, NC 27709-2194

Abstract
Address-Based Sampling frames contain auxiliary variables from the U.S. Postal Service with characteristics of addresses, which can be used to restrict the frame. ABS frames can be enhanced with many other auxiliary variables. Geocodes assign addresses to specific geographic areas. Then Census Bureau demographic variables at area levels can be appended for stratification and disproportionate sampling. Household and person-level marketing variables can reduce screening costs or influence planned contact attempts. Auxiliary variables may be useful for weighting, imputation, or estimation.

Two quality factors affect the usefulness of an auxiliary variable. First is its completeness. Marketing variables, for example, are not available for all addresses. The second factor is accuracy. Area variables are not accurate for all households in an area, and household or person variables may be incorrect. In this paper, completeness of many variables is evaluated using an Enhanced ABS Frame. Accuracy is more difficult; here we compare area aggregations of auxiliary variables to other reliable sources. The completeness and accuracy inform the appropriate uses of an auxiliary variable.

Key Words: ABS, auxiliary variables, sample design, completeness, match rate, accuracy

1. Introduction

Address-based sampling (ABS) loosely refers to survey methodologies for samples selected from address frames. In the U.S., ABS frames are usually based, in part, on the U.S. Postal Service (USPS) sources. A vendor of marketing address lists may have a license to have its lists corrected and updated to be consistent with USPS delivery files for more efficient mailing. Although the address files were not originally intended as sampling frames for housing units, they nevertheless form the most comprehensive commercially available sources for surveys of the residential U.S. population. A useful overview of ABS was recently published by the American Association of Public Opinion Research (2016).

One characteristic of ABS is the availability of auxiliary variables that can be matched to the sample or frame of addresses. Section 2 lists various types of auxiliary variables that can be used. Section 3 describes many of the potential uses of auxiliary variables in ABS methods, and provides examples of such uses. Section 4 describes the quality aspects of auxiliary variables that contribute to their usefulness in ABS methods, and reviews those aspects for a number of examples. Section 5 introduces a new website that allows users to interactively explore characteristics of one ABS frame and some of its auxiliary variables. Section 6 concludes with a few summary remarks and suggestions for additional investigations.

2. Types of Auxiliary Variables in the ABS Context

Auxiliary variables can be obtained from many sources. This section identifies some of the most common types of auxiliary variables used in ABS studies. Some of the variables are unique to ABS surveys, but many are available for other types of surveys, as well.

2.1 USPS Variables

Some auxiliary variables come from the USPS as part of their mail delivery system. These variables relate specifically to each address or mailing point in their system. Vendors include these variables in their address lists to assist their customers in customizing their mass mailings according to their needs. Some of the USPS variables commonly used in ABS are listed in Table 1.

Table 1: Examples of Auxiliary Variables from U.S. Postal Service Records

Variable	*Description of Indicator*
Address type	Nature of the address (e.g., city-style, PO box)
Vacancy flag	Vacant at least 90 days
Seasonal delivery flag	Mail delivery for part of each year (e.g., vacation home, dormitory)
Drop indicator	A common mail receptacle for multiple units
Drop count	Number of units sharing the same drop point
OWGM indicator	Mail is delivered to a PO box only and not also the street address

2.2 Geocodes

For most addresses, a latitude and longitude can be assign so that the address can be located geographically on a map. The coordinates enable addresses to be assigned to specific census geographies including blocks, block groups (CBGs), and tracts. The ability to "geocode" addresses in this way is useful for assigning addresses to targeted areas or to geographically-defined strata.

Not all addresses can be geocoded so precisely. PO boxes, for example, can be geocoded to a ZIP code or post office, but generally not more precisely than that. Such addresses are considered "unlocatable" because they cannot be placed on a map or in a block.

Many address vendors provide geocodes as a value-added service. Alternatively, users of address lists can send addresses to a geocoding vendor or use a software product to generate the geocodes.

2.3 Aggregate Data for Geographical Areas

Once addresses have been geocoded, data about address' specific geographies can be assigned. American Community Survey (ACS) estimates are available for states, counties, and sometimes tracts or block groups. Federal statistical agencies provide a wide array of variables for states and counties. State and private agencies may also provide county-level data. These variables at geographic levels are estimates of aggregations, and all addresses within the same geography will have the same aggregate values. Examples of aggregate level variables are shown in Table 2.

Table 2: Examples of Aggregate Data for Geographical Areas

Variable
Percentage of population that is African-American
Percentage of households with children
Average household income
Average household size (number of persons)
Percentage of homes that are rented
Percentage of householders with college education

2.4 Person or Household Variables

Marketing data vendors (direct marketers) are in the business of helping businesses reach their target consumers. For this reason, the data vendors assemble and maintain vast amounts of information about

persons and households. The marketing variables may be basic demographics at the person level such as age, gender, or race/ethnicity. The data might include personal interests related to magazine subscriptions or group memberships, or financial information such as credit scores. The data might also include household-level information on the home, such as size and occupancy status. A common variable of interest is a telephone number associated with an address. The potential number of variables is vast, and varies considerably by vendor. For illustration, and to contrast with aggregate variables, a few examples of possible person and household variables are shown in Table 3.

Table 3: Examples of Person or Household Variables

Variable
Telephone number
Presence of an African-American person in the home
Presence of a child in the home
Household income
Household size (number of persons)
Householder's surname
Householder's education level

2.5 Modeled Predictions

With variables of interest available for a subset of addresses, such as survey responses from a prior cycle, and with auxiliary variables available for the frame, analysts can model the relationships and generate predictions of the variables of interest for all members of the frame. In a sense, modeled predictions can be considered mass imputations of variables. Table 4 lists some examples of variables that might possibly be modeled.

Table 4: Examples of Modeled Variables

Variable
Probability of an African-American person in the home
Probability of a child in the home
Propensity to be eligible for the study

2.6 Paradata

The term "paradata" refers to data about the data. In surveys, the paradata often summarize data collection experience. Examples of paradata include the length of time to complete a screening interview at an address, the number of interviewer visits before finding someone at home, or interviewer observations about the sampled addresses. West (2016) provides excellent discussion of the use of paradata in surveys. This paper focuses on the other types of auxiliary variables.

3. Uses of Auxiliary Variables

Auxiliary variables have many uses in designing and conducting surveys. Some of the uses for ABS studies are comparable to other types of surveys. This section describes some of the possible uses, with examples.

3.1 Subsetting the frame

The 2015 Residential Energy Consumption Survey (RECS) has an area probability design; that is, it has a multi-stage cluster design where the first two sampling stages are based on geographic areas (http://www.eia.gov/consumption/residential/). At the third sampling stage, addresses are selected from lists in the selected geographies. For efficiency, the national frame of addresses was subset in two ways.

First, only addresses that geocoded into the selected geographies were retained in the sampling frame for the selection of addresses. Second, unlocatable addresses (PO boxes and other address that could not be geocoded) were removed from the frame because field interviewers would not be able to find them. Thus, the 2015 RECS used geocodes and address types to subset the frame.

3.2 Stratification, disproportionate sampling for subdomains

The New York Adult Tobacco Survey is a quarterly in-person survey of adult tobacco users in the state of New York. Using person-level data from the prior survey, models were estimated to predict population smoking rates for all census block groups in the state. The CBGs were then stratified by ranges of predicted smoking rates, and addresses in the strata with higher smoking rates were sampled disproportionately. The auxiliary data in this design included geocodes, past survey screener data, ACS 5-year data at the CBG level, and modeled predictions.

3.3 Modeling response propensities for improving data collection

Both the National Longitudinal Study of Adolescent to Adult Health and the National Survey of Child and Adolescent Well-being modeled response propensities for selected sample units and used those propensities to direct the incentive levels offered to the households.

3.4 Weight adjustments

The 2015 RECS had a companion study to test alternative methods of data collection. The RECS National Pilot Study invited households at sample addresses to participate by mail or web. With contact attempts by mail and many cases with nor response, the pilot study had many cases of unknown eligibility—households that might be vacant or second homes. Rather than apply a CASRO-type eligibility adjustment to the weights, the National Pilot Study modeled the eligibility of completed cases using auxiliary variables and predicted the eligibility of nonrespondents with unknown eligibility status. The predictions were used to generate calibration totals that estimated the total eligible population. These control totals were used to calibrate the weights of respondent cases. The auxiliary data included various frame variables for the models, known eligibility status for screener respondents, and modeled predictions of eligibility.

3.5 Variables difficult to collect in a survey

Although no specific example is provided, it is conceivable that questionnaire items that are difficult or sensitive to collect might be replaced by auxiliary variables such as administrative data for modeled predictions. Using auxiliary variables in this way would be a special case of mass imputation.

3.6 Imputation

Auxiliary variables may be used in various ways to impute missing survey response variables. Auxiliary variables might be used in models for hot deck donor pools, distance calculations for nearest neighbors, selection of donors, or model values. Imputation for the RECS National Pilot Study included the variables in Table 5 in one way or another.

Table 5: Some Variables Used in RECS National Pilot Study Imputation

Variable
Census 2010 Urban Type Code
Climate Zone (collapsed)
CBG median income
CBG proportion of owned housing units
CBG proportion of housing units with 2 or fewer bedrooms
RECS geography
Type of housing unit

3.7 Model-based estimation

The Small Area Estimation procedures for the National Survey of Drug Use and Health (NSDUH) use Hierarchical Bayes models with auxiliary variables at various geographic levels as covariates (http://www.samhsa.gov/data/sites/default/files/NSDUHsaeMethodology2014/NSDUHsaeMethodology2014.pdf). Although not strictly an ABS study, NSDUH nevertheless illustrates the use of auxiliary variables in estimation. Area-level auxiliary variables for NSDUH have been obtained from the Federal Bureau of Investigation, the Centers for Disease Control and Prevention, the Bureau of Labor Statistics. Nielsen Claritas, the U.S. Census Bureau, the Bureau of Economic Analysis, and the Substance Abuse and Mental Health Services Administration.

4. Aspects of Quality and Utility

4.1 Evaluating Auxiliary Variables

Eltinge, et al. (2015) presented a framework for combining survey and auxiliary data. The designer must be mindful of the potential improvements from auxiliary data, but also the potential costs and potential risks resulting from errors and biases in the auxiliary data. As noted previously, relevant auxiliary data have many potential uses. The trick is to find the uses that have the biggest positive net impact on the design.

A common use of auxiliary data is to stratify the frame for sampling when responses are sought for the subpopulation with a rare characteristic. In the case of two strata, a high-density and low-density stratum for the characteristic of interest, the high-density stratum is often oversampled to reduce costs; however, disproportionate sampling introduces design effects that increase variances. Assuming that the stratifying auxiliary variable is complete and accurate, Kalton 2009, Kalton and Anderson, 1986, and Waksberg 1973 have shown optimal allocations to the strata that minimize the variance subject to a cost function, or that minimize the cost subject to a specified precision. This approach can be tailored for different cost functions.

In the two-stratum situation, Kalton (2009) indicated that the usefulness of an auxiliary variable can be evaluated by the prevalence of the rare population in each stratum, the proportion of the rare population in the two strata, and the cost ratios for the two strata. In the ABS context, some types of auxiliary variables are often incomplete and inaccurate. McMichael et al. (2014) stated that the usefulness of an auxiliary variable in stratifying for a rare population is dependent on the prevalence of the rare population overall, the match rate (coverage) of the auxiliary variable to the units on the frame, the accuracy rate of the auxiliary variable, and the cost ratios.

Some authors have developed graphical techniques to help the designer use the optimization formulas. For example, in Figure 1 Tao (2016) illustrated the cost savings expected from using an auxiliary variable to define the two strata for different coverage rates and cost ratios. (This example happens to be for a telephone survey, but the same approach could be used for ABS.)

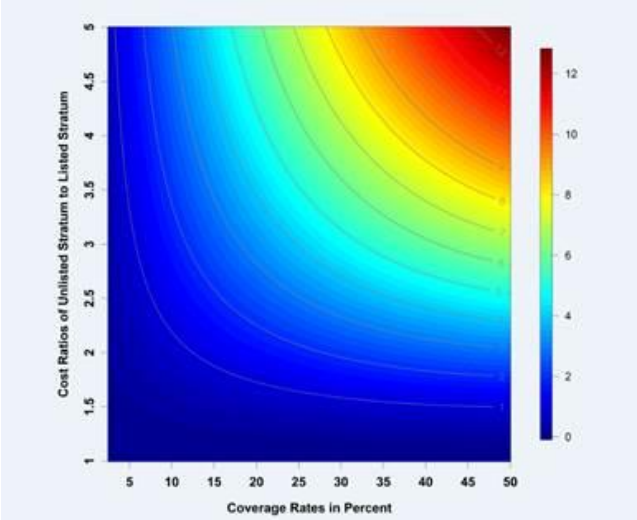

Figure 1: Cost Savings from Optimal Allocation to High- and Low-Density Strata

Levine (2016) assumed a more complex scenario in which estimates are desired for the total population and the rare subpopulation. Because all occupied households are eligible in this scenario, the cost ratios are irrelevant. With disproportionate sampling, minimizing the variances is equivalent to maximizing the effective sample size (ESS) for both the total population and the rare subpopulation. For the New York Adult Tobacco Survey, where African-Americans are 14% of the total New York population, Figure 2 shows the impact on the effective sample sizes for various percentages of African-Americans in the sample. The effective sample size for total population estimates is maximized when African-Americans are sampled proportionately. Oversampling the high-density stratum helps the African-American estimates with minor impact on total population estimates. But only to a point. Oversampling the high density stratum too much makes both estimates worse. For this particular example, Figure 2 shows that the African-American estimates are optimized when approximately 25% of the sample is African-American; at this proportion the effective sample size for African-American estimates is increased about 30%, while the effective sample size for the total population is reduced by about 10%.

Figure 2 helps to optimize sample proportions, but not necessarily the sample allocation for the high- and low-density strata, depending on what is assumed for match rates and accuracy rates. Optimization formulas can be derived for other scenarios, depending on the assumptions about match rates and accuracy rates. These examples are shown to illustrate tools that some designers have developed to evaluate the use of auxiliary variables.

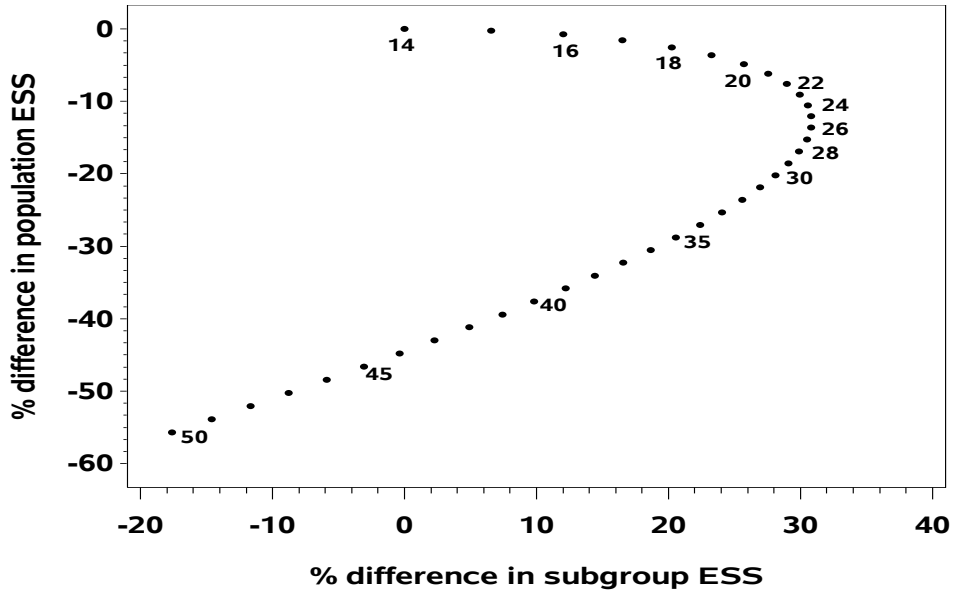

Figure 2: Change in Effective Sample Sizes for Estimates for the Total and African-American Populations in New York State for Various Proportions of African-Americans in the Sample

4.2 Examples of Match Rates and Accuracy Rates

The population prevalence of a characteristic of interest is a given, but the survey designer can choose which auxiliary variable to use based on the other factors: match rate, accuracy rate, and cost ratio, where such information is available. The remainder of this paper focuses on match rates and accuracy rates.

Match rates can be determined prior to data collection when the auxiliary variables are obtained. There are two basic ways of assessing accuracy. Prior to a study, often the best that can be done is to summarize the auxiliary data to an aggregate level and compare to aggregate geographical figures from authoritative sources. This method can be misleading, however, because a rate that appears to be on target may be the net result of both incorrect matches and nonmatches. The best approach is the direct one—compare to survey responses. This approach does not shed light on a variable prior to a study, however. Therefore, it is instructive to review accuracy rates in the literature.

This section summarizes match rates and accuracy rates of a variety of potential auxiliary variables discussed in the literature or experienced by colleagues. These are examples only; individual results may vary by vendor, by geography, by variable definition, and over time. Even the examples below gloss over these differences.

4.2.1 USPS variables

USPS variables are generally considered to be 100% complete for records included in ABS frames. There are exceptions for survey purposes, however. First, drop point addresses correspond to mail receptacles that serve multiple housing units (drop units). An example of a drop point is a main office for a senior community with no records for individual units and no unit identifiers. The use of drop points is more common in a few older cities such as New York, Chicago, and Boston. Nationally, the percentage of drop point addresses in an ABS frame is 0.7%, but the percentage in New York City is 17.7% (Amaya 2016).

Second, when nonlocatable addresses are removed from the frame, such as for in-person surveys, the frame has coverage error. Nationally, 10.9% of residential addresses are considered unlocatable, but the rate varies considerably by state. The unlocatable rate in Mississippi is 15.3% (McMichael 2016).

Perhaps the USPS variable that has been evaluated most for accuracy is the vacancy flag, which indicates whether the home associated with an address has been vacant for at least 90 days. By definition, the vacancy flag will not identify all currently vacant homes, and actual vacancy status can change frequently. The current accuracy of the flag is relevant if one considers using it to identify homes that are out-of-scope for a survey to reduce mailing costs.

The most precise way of evaluating the accuracy of the vacancy flag is through personal visits. It is difficult to confirm by mail that a vacancy flag is accurate, although sometimes post office returned mail is used as confirmation of vacancy status. When mail is not returned, it is not clear whether the home is vacant. It is far easier to count the instances in which the evidence contradicts the flag. Table 5 summarizes instances in which addresses flagged as vacant actually responded to a survey, or addresses not flagged as vacant were later deemed to be vacant. These observed inaccuracy rates provide lower bounds on actual inaccuracy rates.

Table 6: Inaccuracy of USPS Vacancy Flag for Current Vacancy Status

Source	*Flagged as Vacant*	*Flagged but Occupied*	*Not Flagged but Vacant*
Wiant et al. 2016	3%	37%	4%
Kali et al. 2014	< 3%	40%	
Amaya et al. 2014	6.5%	9%	8%

4.2.2 Geocodes

In general, city-style addresses can be geocoded, while non-city-style addresses cannot be geocoded except to a ZIP code or post office level. Therefore, the match rate for geocodes is the rate of city-style addresses. Nationally, 89.1% of addresses are city-style (McMichael 2016).

Accuracy of geocodes depends on the geographic level to which an address is assigned. Geocodes to county level should be very accurate. As the size of the geographies decreases, the geocoding error increases. The level of accuracy depends on the underlying database and on the algorithm for geocoding. Side-of-street errors and misplacements along a block segment sometimes occur, as any user of a GPS device can attest.

4.2.3 Aggregate data for geographical areas

The match rate for aggregate geographical data depends on the level of geography. For the county level and above, aggregate variable match rates are generally 100% because geocoding at that level is accurate, and even unlocatable addresses are associated with ZIP codes, and ZIP codes do not cross county lines. Below county level, the match rate for aggregate geographical variables is consistent with the geocoding rate.

Government data at aggregate geographical levels are considered to be quite accurate for the geographies they represent, subject to sampling and other survey errors. On the other hand, a variable on the percentage of children in an area population, for example, cannot be used to determine the presence of a child at any particular address. The aggregate level accuracy makes aggregate variables very useful for

geographic stratification, say, but their built-in coarseness makes them unsuitable for purposes requiring accurate values at the address level.

4.2.4 Person or household variables

The match rates for marketing variables at the person or household level are highly dependent on several factors. First, the variables come from a variety of sources, with varying degrees of completeness. Second, similar variables may have definitional differences that affect match rates. For example, a flag for homes known to have a child will have a different match rate than a flag for homes suspected of having a child. Third, variables from different sources must be linked to the address file, and the methodology used to link the files record-by-record can affect the completeness of the auxiliary variables. Some vendors offer multiple related variables, varying only by the certainty of the linkages. Fourth, person-level data are often rolled up to form a household level variable, and the roll-up methodology can affect the completeness or match rate of the household level variable. Finally, many marketing variables take the form of an indicator variable or flag for the presence or absence of a certain condition, such as presence or absence of a child in the home. The problem is that knowledge about the variable is limited, and addresses are flagged for the "yes" condition, but "no" is often mixed with "don't know." Thus for flag variables, a match rate is often defined as the proportion with a "yes" value.

Telephone numbers are sometimes appended to ABS samples to provide additional contact options. Listed landline telephone numbers have been the mainstay for the purpose because of the ready availability of addresses for matching. Recently some vendors have begun offering cell telephone numbers, as well. We refer to telephone numbers matched to addresses as phone appends.

Harter et al. (2016) obtained flag variables for the availability of a landline or cell phone append for a large sample of addresses; the results are summarized in Table 7. The flags were based on higher-certainty matches from the vendor's internal sources; rates could have been higher if the vendor had accessed additional sources. The landline match rate is similar to the 47% landline rate obtained by Yancey and Nair (2016).

Table 7: Best Telephone Append Rates (Match Rates) By Phone Type in Total U.S.

Phone Type	*Append Rate*
Landline	43%
Landline Only	27%
Cell	32%
Cell Only	16%
Landline and Cell	16%
None Available	42%

Amaya, Skalland, and Wooten (2010) investigated high-certainty matches as well as "any" match, where any available phone number was a potential match. As expected, "any" match resulted in a substantially higher match rate of phone appends, as shown in Table 8. Note the very high match rate for multi-unit buildings.

Table 8: "Any" Telephone Append Rates (Match Rates) by Address Type

Address Type	*N*	*Match Rate*
All Address Types	69,123	73.6%
Single-Unit Building	51,616	69.0%
In Multi-Unit Building	16.565	92.8%
P.O. Box	723	9.6%
Rural Route	219	29.9%

Sometimes a higher match rate corresponds to a lower accuracy rate. The multi-unit buildings with the higher match rate in Table 7 have a much lower accuracy rate in Table 9. Also, the accuracy of the cell phone appends for the California Health Interview Survey is much improved over the 2012 McMichael and Roe study.

Table 9: Match and Accuracy Rates for Telephone Appends

Source	*Phone Type*	*Initial Match Rate*	*WRN Rate*	*Accuracy Rate*	*Effective Match Rate*
McMichael and Roe (2012)	Overall	71%	75%	80%	45%
	Landline	54%	81%	96%	42%
	Cell	17%	65%	29%	3%
	Address Type	*Initial Match Rate*	*WRN Rate*	*Accuracy Rate*	*Effective Match Rate*
Amaya, Skalland, and Wooten (2010)	All	74%	79%	92%	54%
	Single-unit building	69%	82%	96%	54%
	Multi-unit building	93%	72%	75%	51%
	P.O. box	10%	92%	92%	8%
	Rural route	30%	89%	1%	27%
	Phone type	*Addresses with Appends*	*Numbers Reached*	*Accurate Addresses*	*Accuracy Rate*
California Health Interview Survey: Building Healthy Communities	Landline	7,883	1,804	1,556	86%
	Cell, never LL	2,337	514	305	59%
	Cell ported from LL	2,892	623	444	71%

Yancey and Nair tested a related auxiliary variable, the vendor's prescreening of telephone numbers for working residential number status. About 99% of the telephone numbers were screened and classified as working or nonworking, and the classification was accurate 93% of the time.

Table 10 lists the match rates and accuracy rates for a variety of demographic variables. The presence of a person in a particular age range (child or youth) seems to be a commonly desired auxiliary variable.

Table 10: Match Rate and Accuracy Rate for Person and Household Variables

Source	*Variable*	*Match Rate*	*Accuracy Rate*
DiSogra, Dennis, Fahimi (2010)	(various)	73-95%	
	Home ownership		93%
	White		84%
	Black/African-American		66%
	Hispanic		73%
	HH income < $25k		44%
	HH income > $75k		52%
McMichael et al. (2014)	surname	77%	
	Hispanic		78%
	child 3-17		60%
	Hispanic child 3-17		60%
ZuWallack et al. (2016)	own/rent flag VT	75%	
	VT - own		93%
	VT - rent		77%
	own/rent flag CA	35%	
	CA - own		95%
	CA - rent		58%
Community-based surveys	"any" child flag	31%	
	flagged		44%
	not flagged		79%
	DOB-based child flag	10%	
	flagged		69%
	not flagged		76%

The experience of Ridenhour et al. (2014) is particularly instructive. The 5% of addresses flagged as having a youth aged 11-16 is substantially lower than the 15% expected from occupied housing units according to the American Community Survey or American Housing Survey. This flag variable is an example of "don't know" being combined with "no." The flag was accurate for 68% of the sample, with 61% of eligible households being flagged and 71% of ineligible households not being flagged. Furthermore, the authors noted the following aspects of the flag variable that relate to potential biases:

- Flagged addresses are more likely to be occupied HUs
- Flagged addresses are less likely to respond to the screener
- Eligible youth in flagged addresses are less likely to use tobacco

Clearly match rates and accuracy rates are not the only ways in which auxiliary variables can affect survey quality.

4.2.5 Modeled predictions

The match rate of modeled prediction variables can be 100% under certain conditions:

- The variable of interest (dependent variable) is available for a sample of addresses
- Model covariates are available for the entire frame
- A reasonable model can be found for predicting the dependent variable

If those conditions are met, the accuracy of the predictions is dependent on the accuracy of the variables in the model and the adequacy of the model to predict true values.

For the NYATS, stratification by modeled propensities to smoke was compared with observed smoking rates from the survey for both RDD and ABS samples. Table 11 shows that the smoking rates estimated from the survey were considerably lower than the predicted smoking rates for both samples and all strata. The observed smoking rates did increase with the predicted rates, however. Interestingly, as the predicted smoking rate increased, the ABS response rate decreased. It is possible that propensity to respond is confounded with the propensity to smoke.

Table 11: Predicted and Actual Smoking Rates for RDD and ABS Samples

Predicted Smoking Rate Stratum	*Listed Landline Sample*		*ABS Sample*	
	Response Rate	*Smoking Rate*	*Response Rate*	*Smoking Rate*
0-15%	16%	5%	43%	7%
15-20%	13%	11%	40%	9%
20-25%	14%	14%	41%	15%
25-30%	16%	13%	36%	18%
30% or more	16%	21%	35%	20%

Hubbard et al. (2014) purchased auxiliary variables from two vendors and modeled eligibility and response propensities. They found that variables in common between the two vendors improved the predictions of eligibility, but none of the auxiliary variables consistently improved predictions of response.

5. An Interactive Tool for Exploring an ABS Frame and Auxiliary Variables

RTI International has a national ABS frame based on the USPS' Computerized Delivery Sequence File and No-Stat File obtained through CIS. RTI has supplemented the USPS variables with geocodes, ACS data, and Acxiom marketing data. Most survey organizations are not so fortunate to have their own frame, so RTI has created a website called ABS @ RTI (http://abs.rti.org/) for other survey designers to explore the characteristics of the RTI frame, including summary statistics of frame variables at the national, state, and (sometimes) county levels. The site includes white papers on ABS topics and an interactive ABS Atlas for drilling into the summary statistics. One use of the site is to determine the likely match rates for some auxiliary variables. This new website will be expanded and updated over time.

Figure 3 is a screen shot of an interactive map in ABS Atlas where the user can filter on a number of USPS variables and see summary counts by county.

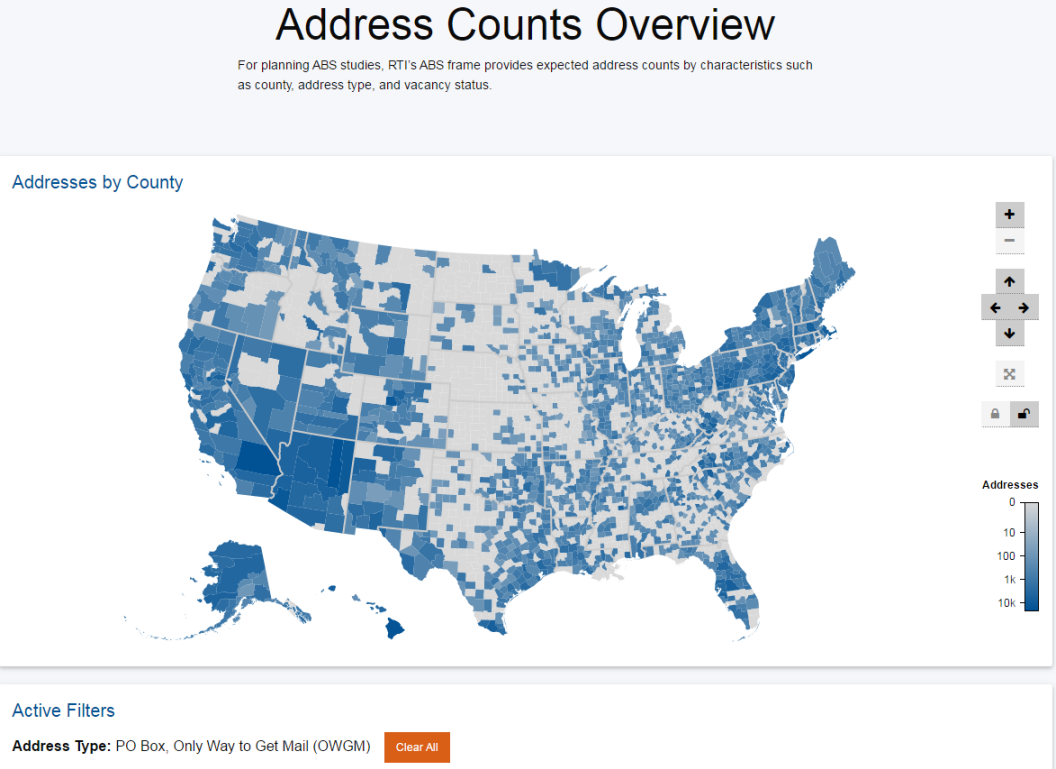

Figure 3: Example of a filtered map showing the counts of OWGM PO Box addresses by county

Summary statistics can be presented in tables or bar charts. Figure 4 illustrates some bar charts of address types and other USPS variables, and Figure 5 illustrates the table format at the website.

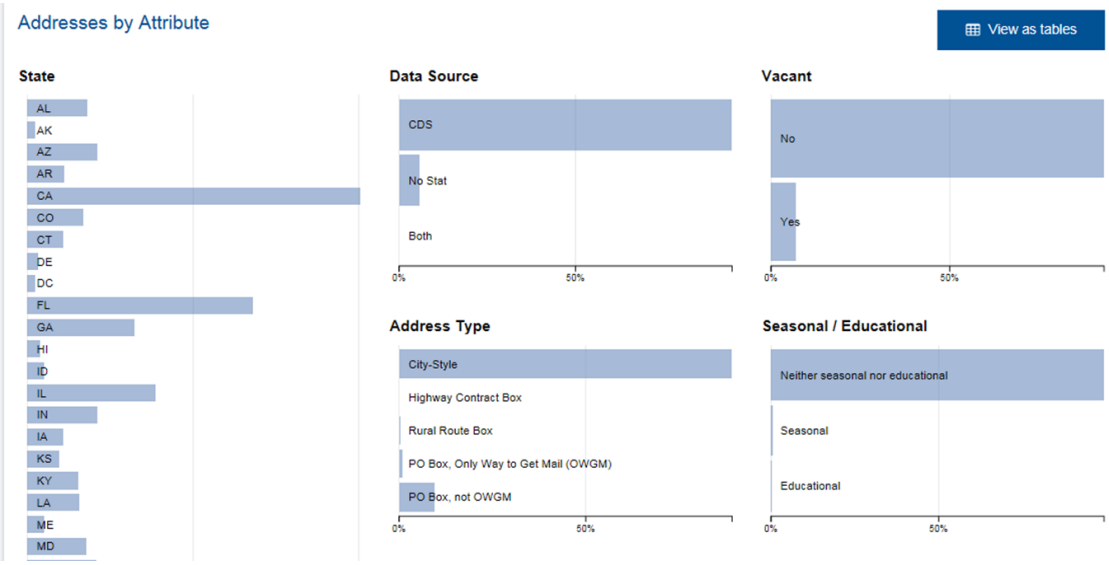

Figure 4: Example of bar chart summary in ABS Atlas

Address Type

Value	Addresses	Percent
City-Style	2,080,021	94.9%
Highway Contract Box	0	0.0%
Rural Route Box	3	0.0%
PO Box, Only Way to Get Mail (OWGM)	2,039	0.1%
PO Box, not OWGM	109,155	5.0%

Drop Point

Value	Addresses	Percent
No	2,107,445	96.2%
Yes	83,764	3.8%
Commercial Mail Receiving Agency	9	0.0%

Figure 5: Example of a table summary in ABS Atlas

Using Figure 5 as an example, a designer planning a survey for Cook County would see that 5.1% of the addresses are not city-style and could not be geocoded below ZIP code or post office level. The match rate for locatable geocodes and for ACS data at tract and CBG levels would be 94.9%.

6. Review and Next Steps

This paper provides a high level overview of the use of auxiliary variables in ABS methodologies. Many examples were described briefly to illustrate the versatility and risks associated with auxiliary variables in the ABS context. The main points of the paper can be summarized as follows:

- ABS frames can have many and varied auxiliary variables.
- Auxiliary variables have many uses in survey design and estimation.
- Usefulness of auxiliary variables depends primarily on match rates, accuracy rates, prevalence of attribute, and cost.
- Match rates and accuracy rates vary widely – *caveat emptor*.
- Characteristics of auxiliary variables can affect survey quality in unexpected ways.
- Auxiliary variables do not have to be complete and entirely accurate to be useful, depending on the application.
- Understand the limitations of variables considered for use.

- Check whether assumptions are satisfied.

In the era of expanding availability of data, the range of potential auxiliary variables will offer exciting potential. On the other hand, the completeness and accuracy of auxiliary variables affects their usefulness. Even so, the temptation to take advantage of the wide array of possibilities is likely to lead to the development of more methods involving data measured with error.

For example, Valliant et al. (2014) demonstrated effective use of imperfect vendor data in a sample design using a linear programming technique. West and Little (2013) developed nonresponse adjustment methods using auxiliary variables measured with error. Kott (2016) presented a method of nonresponse weight adjustments where the auxiliary variable is measured with error and corrected values are available for survey respondents. Additional research in the use of imperfect auxiliary variables as both logical and necessary.

Acknowledgements

The author expresses deep appreciation to the following RTI International colleagues who assisted with examples or contributed substantially to the ABS @ RTI website: Ashley Amaya, Paul Biemer, Derick Brown, Burton Levine, Joseph McMichael, and Joey Morris. The author also thanks session organizer John Eltinge for encouraging this work.

References

Amaya, A. (2016). "Drop Points". RTI International white paper. Available at http://abs.rti.org/ .

Amaya, A., LeClare, F., Fiorio, L., and English, N. (2014). "Improving the Utility of the DSF Address-based Frame through Ancillary Information," *Field Methods* 26 (1), pp. 70-86.

Amaya, A., Skalland, B., and Wooten, K. (2010). "What's In a Match?" *Survey Practice* 3 (6).

American Association for Public Opinion Research (2016). *Address-based Sampling*. Report prepared for AAPOR Council by the Task Force on Address-based Sampling.

DiSogra, C., Dennis, J.M., and Fahimi, M. (2010). "On the Quality of Ancillary Data Available for Address-Based Sampling" *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 4174-4183.

Eltinge, J., Fricker, S., Yang, D., Harter, R., and Ridehnhour, J. (2015). "Practical Approaches to Design and Inference Through the Integration of Complex Survey Data and Non-Survey Information Sources." Presented at ENAR 2015, Miami, FL.

Harter, R., McMichael, J., Brown, F., Amaya, A., Buskirk T., and Malarek, D. (2016). "Telephone Appends", RTI International white paper. Available at http://abs.rti.org/.

Hubbard, F., West, B., Wagner, J., and Gu, H. (2014). "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth" Presented at AAPOR 2014, Anaheim, CA.

Kali, J., Sigman, R., Wren, W., and Jones, M. (2014). "Experiences with the Use of Addressed Based Sampling in In-Person National Household Surveys." *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 3050-3059.

Kott, P. (2016). “Calibration Weighting for Nonresponse with Proxy Frame Information,” Presented at Joint Statistical Meetings, Chicago.

Levine, B. (2016). “The Overpromise of Oversampling,” Poster presented at AAPOR, Austin, TX.

McMichael, J. (2016). “Address Counts“. RTI International data visualizations. Available at http://abs.rti.org/ .

McMichael, J., Harter, R., Shook-Sa, B., Ridenhour, J., and Morris, J. (2014). “Evaluation of Combining Consumer Marketing Data Used for Address-Based Sampling,” Presented at the Joint Statistical Meetings, Boston, MA.

Ridenhour, J., McMichael, J., Harter, R., and Dever, J. (2014). “ABS and Demographic Flags: Examining the Implications for Using Auxiliary Frame Information,” Presented at the Joint Statistical Meetings, Boston, MA.

Valliant, R., Hubbard, F., Lee, S., and Chang, C. (2014). “Efficient Use of Commercial Lists in U.S. Household Sampling.” *Journal of Survey Statistics and Methodology*, 2, pp. 182-209.

West, B.T. (2016). “On the Quality and Utility of Alternative Auxiliary Data Sources Used in the National Survey of Family Growth,” Presented at the Joint Statistical Meetings, Chicago, IL.

West, B.T., and Little, R.J.A. (2013). “Non-response adjustment of survey estimates based on auxiliary variables subject to error,” *Journal of the Royal Statistical Society, Series C* (Applications) 62(2), 213-231.

Wiant, K., McMichael, J., Murphy, J., Morton, K., and Waggy, M. (2016). “Consistency and Accuracy of Undeliverable Codes Provided by the US Postal Service: Implications for Frame Construction, Data Collection Operational Decisions, and Response Rate Calculations,” Presented at AAPOR 2016, Austin, TX.

Yancey, T., and Nair, V. (2016). “Should We Always Use the Telephone Numbers Matched to an ABS Sample?” Presented at AAPOR 2016, Austin, TX.

ZuWallack, R., Brown, J., Brassell, T., Williams, R., Dion, R., and Cooper, V. (2016). “Using Auxiliary Data to Increase Efficiency of Sampling Rental Units in Metropolitan Regions,” Presented at AAPOR 2016, Austin, TX.