Predictive Modeling Using an Enhanced Address-Based Sampling Frame

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

Abstract
Address-Based Sampling (ABS) frames are well suited for linking to auxiliary data, either by geocoding frame addresses or by direct address matching. Survey researchers commonly append publicly available census data to ABS frames. Additionally, the frame can be linked to proprietary data sources compiled and sold by large data brokers such as Experian, Acxiom, Epsilon, and CoreLogic. When these auxiliary data are appended to the ABS frame and combined with reported values and response paradata from a previous survey, researchers can develop address-level models that predict demographic characteristics and respondent behaviors, helpful in sample designs for surveys of targeted subpopulations. These models can be used for stratification, decisions about data collection protocols, and weight adjustments, but this paper focuses on stratification and sampling for eligible subpopulations. This paper reviews the general principles of predictive modeling with ABS and gives examples from our experience.

Key Words: Address-Based Sampling, ABS, frame, auxiliary data, stratification, predictive modeling

1. Stratification for Sampling Rare Populations

Often when the target population is a relatively rare subpopulation or domain (e.g., Alaska Native/American Indian, narrow age range, persons with a specific medical condition), no list or frame of the specific subpopulation is available. Screening for the subpopulation often begins with a general population sample, such as a sample of residential addresses as in address-based sampling (ABS) designs, the context of this discussion. Screening general population samples from address frames can be very inefficient and expensive for subpopulations, even if the subpopulation is not that rare. Stratification with disproportionate sampling is a common way to reduce costs and improve efficiency when trying to sample subpopulations. (See, for example, Lohr (1999) and Kish (1965).) For example, if the frame can be stratified into high- and low-density strata with respect to the target subpopulation, one can sample the high-density stratum at a higher rate. This approach is common practice, whether the goal is estimation for the subpopulation alone or for the subpopulation along with the total population. Kalton (1986, 2009) discussed the approach of stratification with disproportionate sampling and other methods for sampling rare populations. Many of the ideas and principles in this paper came from the Kalton papers.

If the rare subpopulation is clustered, one can sample geographic clusters (primary sampling units or PSUs) with probability proportional to the rare subpopulation's measure of size (density) for the cluster. Then sample addresses within the sampled clusters can be

sampled. This is a two stage or cluster design (Lohr 1999). Or, one can put the high-density clusters/geographies into one stratum and the low-density clusters/geographies into another and select addresses directly from the strata. Either way, stratification and oversampling the high-density stratum assumes that the densities of the subpopulation in the strata and clusters are available. When strata or clusters are defined geographically (e.g., counties, census tracts, census block groups), density information often is available from the decennial census or the American Community Survey (ACS) for many demographics. To define the strata, one selects a threshold for the measure of size and assigns the geographies to the strata accordingly. Samples selected from the high-density stratum will have a higher percentage of the rare subpopulation, so selecting from that stratum at a higher rate than from the low-density stratum is more efficient than an unstratified design for reaching target numbers of the rare subpopulation. Multiple strata of varying density levels or subpopulations are possible.

Census and ACS demographics are commonly used for measures of size of subpopulations for strata and geographical clusters. The measures of size then inform the sampling rates within strata and expected counts of the desired subpopulation in the sample. ABS lends itself well to these designs because addresses can be geocoded (assigned a latitude and longitude) and classified into geographical clusters.

If the subpopulation is not geographically clustered, but is more or less uniformly distributed (e.g., females aged 15-44), stratifying geographically will not help with efficiency. Even if productive stratification is possible, the survey might still require more screening of addresses than desired or affordable.

1.1 Address-Level Stratification with ABS Frames

What if a survey designer could stratify individual addresses rather than clusters or geographical areas? If the address-level stratification is any good, the survey might save considerably on screening costs. However, stratifying individual addresses generally requires more frame information than the usual U.S. Postal Service files provide. Consumer marketing data, on the other hand, are abundant. Marketing variables are often at the person level, and they often take the form of flag indicator variables. Marketing data are generally expensive, often incomplete, or just plain wrong. Still, marketing data have proven to be useful for stratification and oversampling for rare populations such as specific age groups.

Consumer marketing data are used heavily for direct-mail marketing. Advertisers want to reach as many potential customers as possible. Sometimes marketers' customers are in specific geographical areas or in specific demographic groups. Marketers want address lists that cover their potential customers as completely as possible. False positives may be less of a concern to advertisers than they are to survey researchers. Vendors of marketing data accumulate information from multiple sources, including publicly available administrative records and purchasing behavior. Marketing data can be wrong, of course. For example, the purchase of a toy does not guarantee that the purchaser lives in a household with children. Nor does the absence of a toy purchase indicate that a household does not have children. In that sense, yes/no flags may more accurately be described as probably/don't know. Furthermore, purchasers can move so that the marketing data are out of date. Consequently, these databases may have many missing values and inaccuracies (Harter 2016). But the databases are useful for their intended purpose—direct mail marketing. Survey researchers who try to use the same marketing data for survey research must be realistic about the shortcomings of the data and the potential impact of these shortcomings on their sample design.

A relatively simple way to use marketing data is to roll up the person-level data to the address level and use the person-level flags as address-level flags. For example, if the marketing data indicate that a female aged 15-44 lives at the address, then the address automatically goes into the high-density stratum for that subpopulation. Because the marketing data for individual addresses may be wrong, not all of the high-density addresses have eligible members of the subpopulation, and not all of the low-density addresses are ineligible. The assumption is that the flags are better (ideally substantially better) at assigning addresses to density strata than random assignments.

1.2 Predictive Modeling for Stratification

West et al. (2015) tried using commercial auxiliary data (marketing data) with models to predict survey eligibility at the household level for the National Survey of Family Growth (NSFG), with some success. Sponsored by the National Center for Health Statistics, the NSFG collects statistics on family growth, formation, and dissolution including factors associated with childbearing, fertility, medical services for family planning, sexual and drug-use behaviors related to HIV and STD risk, and child adoption. The target population for the 2011-2019 NSFG was men and women aged 15-44 in the United States. For this research, the authors used sampled housing units from June 2012-March 2013 NSFG, their eligibility status, frame variables (including census and ACS demographics), and paradata. Covariates in the model came from the address frame and from marketing data, either from Aristotle or from Marketing System Group's three unidentified marketing data sources. Although the marketing data were often missing or incorrect, the models with frame variables and marketing data fit the eligibility outcomes much better than the models with frame variables alone. In the paper, the authors did not actually predict eligibility on a separate sample; however, the models were applied in subsequent cycles of NSFG, making use of the predictive nature of the models. This application is an example of predictive modeling for subpopulations.

In March 2022, McPhee (2022) gave a webinar sponsored by the American Association of Public Opinions Research (AAPOR) that was an overview of predictive modeling. The presentation covered many applications, but this discussion focuses on the specific application of static predictive modeling for sampling rare subpopulations with ABS. First, the modeler needs good training data, an ABS frame with additional covariates and known eligibility outcomes. The results from a prior cycle of a survey may suffice. A model is estimated where the dependent variable is subpopulation eligibility. The model is applied to a current frame to predict eligibility, or the probability of eligibility. The new frame is stratified based on the eligibility predictions, and a disproportionate sample is selected from the stratum of households with high probability of eligibility. Predictive modeling for stratification is a variation on the original stratification/disproportionate sampling methodology. The goal of the model is prediction, not the fit of the model on the training data or the model parameters. Even mediocre models with imperfect auxiliary data (e.g., marketing data and ACS characteristics at the area level) can be useful for stratifying the frame.

2. Predictive Modeling with ABS at RTI

RTI International maintains its own copy of an ABS frame and has enhanced the frame with marketing data. RTI leases a copy of the United States Postal Service's Computerized Delivery Sequence File (USPS 2016). Geographers assign geocodes to the addresses in the frame. In turn, the geocodes are spatially linked to census geographical areas. RTI appends

auxiliary data from the decennial census, the ACS, and other federal sources by matching on the geographical areas.

RTI also leases consumer marketing data. RTI aggregates the person-level data to address level and merges the resulting indicator variables onto the ABS frame. The enhanced frame has supported extensive ABS research (https://abs.rti.org/).

In the sections that follow, we summarize RTI's efforts in recent years to sample subpopulations efficiently with the enhanced frame, including predictive modeling**.**

2.1 Tobacco Surveys

The Food and Drug Administration (FDA) and state agencies often sponsor studies to evaluate anti-tobacco campaigns aimed at particular subsets of tobacco users. These studies are particularly suitable for use of marketing data to help find eligible subpopulation members.

2.1.1 New York Adult Tobacco Survey

The New York Adult Tobacco Survey is a quarterly 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 (CBGs) in the state. The CBGs were then stratified by ranges of predicted smoking rates, and addresses in the strata of CBGs with higher smoking rates were sampled disproportionately (Harter 2016). This early version of predictive modeling was used for geographical stratification, not address-level stratification.

2.1.2 Evaluation of Public Education Campaign on Teen Tobacco

The first cycle of the longitudinal Evaluation of Public Education Campaign on Teen Tobacco (ExPECTT) survey, sponsored by FDA's Center for Tobacco Products, targeted youths aged 11-16. The survey was designed to evaluate FDA's general market youth tobacco prevention campaign. The ABS frame was stratified directly on a composite age group flag derived from marketing flags. This was the relatively simple form of stratification, but not predictive modeling. Using the marketing flag directly led to a few observations (Ridenhour et al. 2014):

- The flag matched only about 5% of addresses.
- Flagged addresses were three times more likely than unflagged addresses to have an eligible youth.
- Sixty-one percent of eligible households were flagged as eligible.

In other words, the flag was useful for stratification and reducing costs in finding eligible youths, but it was far from a perfect predictor. The national area probability sample of about 45,000 addresses for this in-person survey was selected from the ABS stratified frame. After data collection, samplers used the ExPECTT screener data to build predictive models for the next study.

2.1.3 Rural Smokeless Tobacco Education Campaign

The Rural Smokeless Tobacco Education Campaign (RuSTEC) was another FDA study to evaluate a public education campaign to prevent and reduce smokeless tobacco use among rural male youths (ages 11-16 in the baseline year) in 30 geographical areas. RuSTEC had a stratified frame of addresses based on the predictive models developed from the extensive ExPECTT screener data. Ridenhour and McMichael (2017) created multiple propensity

strata from the models, and they summarized the expected and observed eligibility rates among screened households, as shown in Table 1.

Table 1: Eligibility Rates for the Rural Smokeless Tobacco Education Campaign

Stratum	*Expected Eligibility Rate*	*Observed Eligibility Rate*
1	2.9%	1.2%
2	4.4%	2.5%
3	9.3%	7.5%
4	13.3%	16.1%
5	25.6%	35.8%
6	30.0%	42.9%

In turn, RuSTEC screener data (24,000 screened households) were used to build predictive models for two other studies. One was the second round of the ExPECTT study, which again stratified and screened for youths aged 11-16, but this time the stratification was based on the predictive models that used RuSTEC for training data.

2.1.4 Point of Sale Intervention for Tobacco Evaluation

The other study to benefit from the RuSTEC models was Point of Sale Intervention for Tobacco Evaluation (POSITEv), which stratified and screened for households with smokers aged 25-55 (McMichael & Wiant 2019). FDA's POSITEv study was an evaluation of a public education campaign using advertising in and around tobacco retail outlets to educate consumers about the dangers of tobacco products. Nationally, only 11% of U.S. households were eligible, but the budget required that more than 17% of sampled households be eligible. RuSTEC screener data were used to predict the eligibility of each household in the 30 counties selected for the study. The ABS frame for these counties was divided into 10 equally sized strata based on the predicted probability of an address having a smoker aged 25-55. The predicted and observed eligibility rates for the 10 strata are shown in Figure 1 and Table 2.

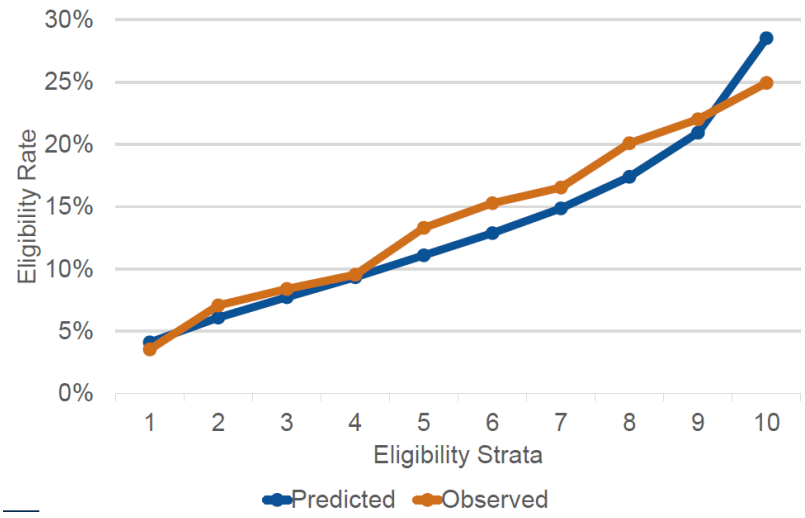

Figure 1: Predicted and Observed Household Eligibility Rates by Density Strata for POSITIVe Study

Table 2: Eligibility Rates for the POSITEv Study

Eligibility Strata	Sample	Eligibility Rate Predicted	Observed	Diff
Low - 1	4,822	4.1%	3.5%	-0.6%
2	5,661	6.1%	7.1%	1.0%
3	5,972	7.7%	8.4%	0.7%
4	6,882	9.3%	9.6%	0.2%
5	7,344	11.1%	13.3%	2.2%
6	8,504	12.9%	15.3%	2.4%
7	10,460	14.9%	16.5%	1.7%
8	13,367	17.4%	20.1%	2.7%
9	17,261	20.9%	22.0%	1.1%
High - 10	24,268	28.5%	24.9%	-3.6%
Overall	**104,541**	**17.2%**	**16.9%**	**0.3%**

Although the predicted eligibility rates for the strata were not perfect, they were reasonably accurate, enabling stratification and oversampling to improve efficiency. Equal allocation of the sample to the strata would have yielded a 13.3% eligibility rate, better than random sampling, but not as efficient as desired. By oversampling the higher density strata, the authors were able to achieve a 17.2% eligibility rate among the sampled addresses. The authors recognized that an optimal allocation for this population could have resulted in even greater improvements in efficiency.

2.2 Boating Surveys

Boat ownership is a rarer characteristic than tobacco use and many age groups. Finding suitable training data for predictive modeling of this subdomain was particularly challenging. Nevertheless, some gains in efficiency were achieved.

2.2.1 National Recreational Boating Safety Survey

The National Recreational Boating Safety Survey was a recreational boating survey for the U.S. Coast Guard (Ridenhour et al. 2021). The survey involved two frames, an incomplete registry of boat owners and an ABS frame without boat ownership indicated. The ABS frame provided complete coverage, but it would have yielded boat-owning households very inefficiently without predictive modeling. In addition to the two frames, a geospatial database was available that contained the number of boats for geographical areas. However, the leased geospatial database was not permitted to be linked to the frames. The geospatial database was used to develop a model to predict the presence of boats for census block groups. This model was combined with the registry data to develop a second model for the presence of boats at the address level. The second model was applied to the ABS frame. The two-model approach definitely helped find boat-owning households in the ABS frame more efficiently (43.2% eligibility rate; see Table 3), but the performance was not as strong as the statisticians expected.

Table 3: Data Collection Rates* for the National Recreational Boating Safety Survey

Rate	*Registry Frame*	*Stratified ABS Frame*	*Overall*
Screener	33.6	15.1	22.2
Eligibility	91.9	43.2	71.4
Yield	30.9	6.5	15.9

*Based on 9 of 12 completed cohorts
Source: Ridenhour et al. (2021)

2.2.1 Recreational Boat Fishing Survey

The Recreational Boat Fishing Survey (RBFS) is a survey of people who fish by boat, a very rare subpopulation. The RBFS is a subset of the Fishing Effort Survey (FES) conducted for the National Marine Fisheries Service of the National Oceanic and Atmospheric Administration. The FES monitors recreational saltwater fishing activity by residents of Atlantic and Gulf Coast states. The ABS frame was supplemented with a state database of licensed saltwater anglers. The results of predictive modeling for this study did not meet expectations and are not as informative as the other study results presented.

2.3 Current Surveys

Without going into detail because the studies are still in progress, RTI samplers are using predictive modeling on two more studies. The first application is to help find multigenerational households. The second is to find age groups for the NSFG, the same survey that West et al. (2015) first wrote about, coming full circle.

3. Discussion

Oversampling high-density strata for a target population is not without risks. If one samples from the high-density stratum and not from the low-density stratum, coverage error and the potential for coverage bias result. If one samples from the low-density stratum, but at a lower rate, the members of the subpopulation obtained from the low-density stratum will have much larger sampling weights, increasing the design effect and reducing the effective sample size. In other words, the stratification approach may get the target number of respondents from the subpopulation at lower cost, but the resulting sample of the subpopulation does not provide as much statistical value to the estimates as if the sample had been selected with equal probability (and greater cost). Levine (2016) illustrated that it is possible to overdo the disproportionate sampling from the high-density stratum, leading to higher variances than proportionate sampling. Kalton (1986, 2009) gave formulas for the optimal sample rate in the high-density stratum relative to the low-density stratum. Even then, he showed that the reduction in variance from disproportionate sampling is small unless the density of the subpopulation in the high-density stratum is high *and* the high-density stratum contains a high proportion of the subpopulation.

Stratifying and oversampling for a subpopulation involves other tradeoffs, too. Sampling for the subpopulation is less efficient for estimates of the total population because of disproportionate sampling and unequal weighting effects. If the study requires both total population and subpopulation estimates, ideally the loss of precision to the total population estimates will be small relative to the gains for the subpopulation. Often, the stratification with some oversampling is still an effective design.

The level of success with stratification and oversampling of subpopulations will vary by the target population, the outcomes being measured, and the availability and quality of auxiliary data. Ideally, the auxiliary data should be reasonably complete and accurate, and the match rate to the frame should be high. It is important that the subpopulation densities in the strata be known, or at least approximated well. These conditions can be tricky with marketing data for direct stratification, as we saw with the ExPECTT 1 example. Predictive modeling requires good training data, such as screener data from a large sample. Repeated cross-sectional samples may be ideal for predictive modeling, but it is possible to use models from one study to predict eligibility for another study. Even then, as McMichael and Wiant (2019) pointed out, what counts is not the eligibility of the sample so much as

the eligibility of the *responding* sample. Thus, having accurate response rate estimates for the strata is important for an optimal allocation.

Now, the frame and marketing data do not need to be in the sampler's possession to take advantage of these methods. One can obtain samples from a vendor such as Marketing Systems Group that has both an address frame and separate marketing databases. Vendors may not be permitted to match the marketing data to the entire frame, but they can match the marketing data to samples of addresses. In that case, one can purchase a very large first phase sample and have the vendor match the sample to the marketing data. The marketing flags can be used to stratify directly, as in the ExPECTT 1 example. Alternatively, if suitable models are available, the models can be used to predict eligibility of the phase 1 sample addresses. Then the phase 1 sample can be stratified as the frame for selecting the phase 2 sample. Generally, this is much less costly than extensive address screening without stratification for the target subpopulation.

Acknowledgements

The authors thank RTI International for supporting ABS research generally and this paper specifically.

References

Harter, R. (2016). The Quality of Auxiliary Variables in an Enhanced Address-Based Sampling Frame. Invited presentation in *JSM Proceedings, Government Statistics Section*, pp. 74-89, Alexandria: American Statistical Association.

Kalton, G. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society, Series A*, Vol. 149, part 1, pp. 65-82.

Kalton, G. (2009) Methods for Oversampling Rare Subpopulations in Social Surveys. *Survey Methodology*, 35, 125-141.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons.

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

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury Press.

McMichael, J. & Wiant, K. (2019). *Improvements in Sample Design With Address-level Prediction Models*. Presented at AAPOR, Toronto.

McPhee, C. (2022). *Applications of Predictive Modeling to Survey Design & Operation in Address-based Samples*. Webinar presented March 17, 2022. American Association of Public Opinion Research.

Ridenhour, J. L. & McMichael, J. P. (2017, May). *Propensity Stratification With Auxiliary Data for Address-Based Sampling Frames*. Presented at the 2017 American Association for Public Opinion Research conference, New Orleans, LA.

Ridenhour, J., McMichael, J., Harter, R., & Dever, J. (2014). *ABS and Demographic Flags: Examining the Implications for Using Auxiliary Frame Information*. Presented at the Joint Statistical Meetings, Boston, August 7, 2014.

Ridenhour, J., McMichael, J., Krotki, K., & Speizer, H. (2021). Using big data to improve sample efficiency. In *Big Data Meets Survey Science* (C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov & L. E. Lyberg, eds.). https://doi.org/10.1002/9781118976357.ch17

U.S. Postal Service, 2016. *CDS User Guide*. Retrieved October 6, 2022. Available at Computerized Delivery Sequence (CDS) User Guide | PostalPro (usps.com)

West, B.T., Wagner, J., Hubbard, F., and Gu, Haoyu (2015). The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, 3, 240–264.