Estimating Mail or Web Survey Eligibility for Undeliverable Addresses: A Latent Class Analysis Approach

Paul Biemer[1,2], Joe Murphy[1], Phil Kott[1]
[1]RTI International, 3040 E. Cornwallis Road, Research Triangle Park, NC 27709
[2]University of North Carolina, Chapel Hill, NC 27514

Abstract

Mail surveys, as well as many web surveys, rely on mailings to the sample members inviting them to complete a paper or web questionnaire. Sample members are selected from a frame such as the address-based sampling (ABS) frame derived from the U.S. Postal Service's (USPS) Computerized Delivery Sequence file. A well-known problem with such surveys is determining the eligibility of sample members who mailings are returned as "undeliverable." The undeliverable codes provided by the USPS are often inconsistent across repeated mailings to the same address, yet they typically are treated as accurate in determining case eligibility. This paper describes how sample member eligibility was estimated using a latent class analysis of four indicators of eligibility. In our application, three indicators were based on the USPS codes from 3 mailings sent within a 12-day period to all sampled households and the fourth was the vacancy indicator on the ABS frame. This approach was applied to data from the Residential Energy Consumption Survey National Pilot – a sample survey of 9,650 households – in the calculation of response rates and survey weights.

Key Words: mail surveys, web surveys, occupancy, address-based sampling, latent class analysis

1. Introduction

A challenge with conducting a survey using mail-contact, self-administered modes is how to determine whether a nonresponding housing unit is occupied or vacant since we typically have imperfect indicators of household vacancy status. There are several reasons why it is important to have an accurate indicator vacancy status. First, it is a cost issue. Continuing to mail reminders and followup letters to nonrespondents is a waste of resources if the household is vacant and will never respond. It is also important in computing response rates because vacant units should be removed from the denominator of the response ratio. Finally, it can bias the post-survey adjustments if nonresponse propensity factors include vacant (and ineligible) units rather than treating them as out of scope.

The address-based sample (ABS) frame (the USPS Computerized Delivery Sequence file) includes a readily available indicator of vacancy status for each frame unit, but we know this indicator is not always accurate. For one thing, the frame vacancy status may be out of date by the time we draw the sample and begin mailing to households. This suggests that its accuracy is worse in areas of higher occupancy turnover. Because the target population is usually occupied housing units, it is important that we exclude vacant units from the frame or sample. However, removing occupied units that are mistakenly identified as vacant is also problematic because it can lead to undercoverage bias. Responding households can be classified as occupied accurately but nonresponding households are problematic because they may be either vacant or occupied and not participating.

Another source of information available on vacancy status comes from the USPS-coded outcome of the mailing itself. When survey letters are mailed to vacant units, the USPS will often return the letter indicating it was undeliverable and the reasons why. But we have found these codes to be inconsistent and incomplete, and otherwise imperfect indicators of vacancy status.

In this paper, we consider how to combine the information from these three types of indicators: the ABS frame indicator, undeliverable notices from the post office, and the household response indicator. The next section describes these data in more detail and Section 3 shows how a latent class model can combine these data to provide an estimate of the probability that a unit is occupied or vacant. Section 4 shows how this information can then be used to direct field followup, calculate response rates and correct weighting adjustments.

2. Description of the Data

The Residential Energy Consumption Survey (RECS), sponsored by the Energy Information Administration (EIA), is designed to measure energy characteristics of U.S. households. The data for this paper comes from the 2015 RECS National Pilot (RECS-NP) which preceded the 2015 RECS main survey and was conducted by mail using paper and web modes. The RECS has traditionally been an in-person survey, but, for the RECS-NP, EIA wanted to explore options for conducting data collection by mail-contact, self-administered modes like computer assisted web interviewing (CAWI) and paper and pencil interviewing (PAPI). As a result of the RECS-NP, the CAWI and PAPI modes were subsequently used in part on the latest round of the RECS main survey in 2016.
Given that the RECS, and many other in-person surveys, are transitioning to mail-contact modes, the question of how to deal with the vacancy status question becomes even more important.

Data collection for the RECS-NP was conducted by IMG Crown Consulting and RTI International in late 2015 and early 2016. The RECS-NP focused on an experiment to test four different protocols and two levels of promised incentives using paper and web response with mail contact. Details of this experiment can be found in Biemer, et al. (2016). All protocols for the RECS-NP involved sending out a prenotice postcard, an invitation letter, a reminder letter to all 9,650 addresses in the sample.

For the RECS-NP, the only information that was available on vacancy status for a sample household prior to contact was the frame indicator. Subsequently, the initial three mailings that went to all address provided additional information on vacancy status either through the USPS returned mail notices or from sample addresses that responded to the survey request.

In an earlier analysis, Wiant et al (2016) took an initial look at the consistency of indications of vacancy provided by the USPS for undeliverable mailings. As previously noted, three mailings were sent to households over a 12 day period. Wiant, et al found that about 31 percent of addresses with at least one undeliverable were only undeliverable for one of the three mailings, 29 percent were undeliverable twice and 40 percent were undeliverable for all three mailings. In addition, when multiple undeliverable notices were received, there was considerable inconsistency as to the reason an address was deemed undeliverable. These results suggest that USPS indications of vacancy are quite inconsistent and thus, unreliable.

The ABS frame indicator assigns each address to four categories: not vacant, not seasonal; not vacant, seasonal; vacant, not seasonal; vacant, seasonal. Wiant, et al compared this indicator to two other indicators: one that indicated an address is "vacant" if the USPS declared the address to be vacant for at least one undeliverable notice; otherwise, the indicator was coded as "unknown." The second indicator coded an address as "occupied" if a response was received from the address; otherwise, the address was coded as "unknown." The results of this comparison are shown in Figure 1.

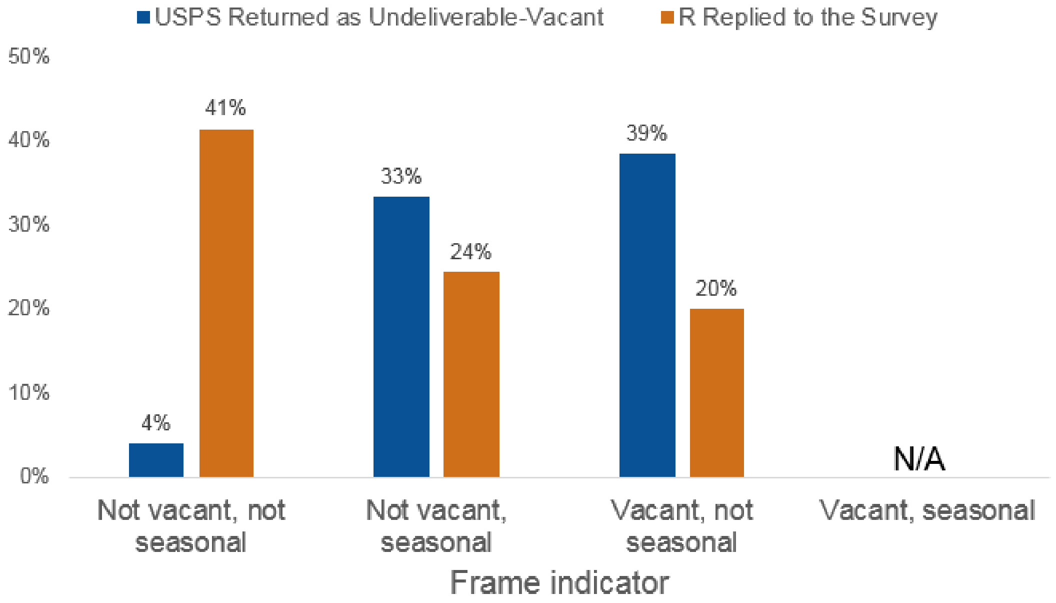

Figure 1: Frame Indicator of Vacancy by USPS Returned Mail and Respondent's Reply (from Wiant et al., 2016)

Note that about one-third of the addresses that were coded by the ABS frame as "not vacant, but occupied on a seasonal basis" were coded as "vacant" by USPS indicator. This suggest high disagreement between the frame and USPS for seasonally occupied households. Further, note that responses were received for 20 percent of addresses that the frame indicated were vacant. As further evidence of inconsistencies in these data, note that among frame vacant, not seasonal addresses, the USPS indicator agreed only 39 percent of the time – a 61 percent disagreement rate!

One explanation for these findings is that the frame information is out of date and the unit may have been vacant (or occupied) the last time the frame was updated, but then subsequently became occupied (or vacant) prior to the survey. Nevertheless, the inaccuracy of the frame indicator is fairly obvious based upon these results.

The question addressed in this paper is how best to use all of these data when they provide incomplete, inaccurate and inconsistent information about the true vacancy status of an address. The goal is to combine these data in some fashion to produce a single output indicator of vacancy having greater accuracy than the individual input indicators. One promising approach for this purpose is latent class analysis, since it can be used to combine fallible indicators statistically to produce a single indicator that combines the input indicators in an optimal way. The next section describes the latent class model that was constructed for the five indicators from RECS-NP and the resulting indicator of vacancy that was produced from it. As a by-product, the model estimation process also produces estimates of the error probabilities for each of the five indicators.

3. Latent Class Model of Vacancy

We analyzed data from the RECS-NP in order to estimate the vacancy rates for the sample and the error rates associated with the post office (PO) undeliverable notices for three early mailings and the frame indicator of vacant. The basic model is as follows.

Let M_k denote the status of a housing unit based upon mailing *k* (= 1, 2, or 3) where $M_k = 1$ if the housing unit is deliverable, $M_k = 2$ if undeliverable (vacant) and $M_k = 3$ if undeliverable (other). Let *F* denote the frame indicator where *F* = 1, if not vacant-non-seasonal, *F* = 2 if vacant, and *F* = 3 if not vacant-seasonal. Let *R* be the response indicator which is 1 if the unit submitted a response and 2 if otherwise. Let *X* denote the true status of a unit as either 1 (not vacant) or 2 (vacant).

Let $M_1M_2M_3FR$ denote the five-way cross-classification table for the indicators. Assume that M_1, M_2, M_3, F, R are all conditionally local independent indicators of *X* given the explanatory grouping variables collectively denoted by the vector, **G**. Thus cell *ijklm* of the $M_1M_2M_3FR \,|\, \mathbf{g}$ table given the value $\mathbf{G} = \mathbf{g}$ has likelihood kernel given by

$$\sum_x \pi_{x\mathbf{g}}^{X|\mathbf{G}} \pi_{i|x\mathbf{g}}^{M_1|X\mathbf{G}} \pi_{j|x\mathbf{g}}^{M_2|X\mathbf{G}} \pi_{k|x\mathbf{g}}^{M_3|X\mathbf{G}} \pi_{l|x\mathbf{g}}^{F|X\mathbf{G}} \pi_{m|x\mathbf{g}}^{R|X\mathbf{G}} \qquad (0.1)$$

where $\pi_{i|xg}^{M_1|X\mathbf{G}} = \Pr(M_1 = i \,|\, X = x, \mathbf{G} = \mathbf{g})$ with similar definitions for M_2, M_3 , $\pi_{l|x\mathbf{g}}^{F|X\mathbf{G}} = \Pr(F = l \,|\, X = x, \mathbf{G} = \mathbf{g})$, $\pi_{m|x\mathbf{g}}^{R|X\mathbf{G}} = \Pr(R = m \,|\, X = x, \mathbf{G} = \mathbf{g})$ for *i* = 1,2,3, *l* = 1,2,3, *m* = 1,2 and *x* = 1,2. Further assume that $\pi_{i|x}^{M_k|X} = \pi_{i|x}^{M|X}$ for all *k* and that $\pi_{1|2}^{R|X} = 0$, i.e., vacant units cannot submit responses. This model can be fit using constrained maximum likelihood estimation.

The vector, **G**, consisted of four grouping variables, viz.: *H* denoting high-rise/other (1, 2), *U* denoting Rural/Urban (1, 2), *S* denoting single/multi (1, 2), and *R* denoting carrier route type where 1 = city route, 2 = highway contract route and 3 = rural route. Both the latent vacancy indicator, *X*, and the indicators (M_k, *k* = 1, 2, 3, *F* and *R*) were allowed to interact (vary) by the grouping four variables. All interactions were tested for significance and only significant interactions were kept in the model. This model fit the data quite well and provided plausible estimates of the vacancy rates and classification error terms. As a measure of model fit, the dissimilarity index (which compares the model expected and observed cell counts) was 0.03 (where 0.05 or less is considered a good fitting model).

Table 1 shows the vacancy rates according to five indicators of vacancy: the three USPS indicators (M_1, M_2 and M_3), the frame indicator (F) and the model indicator which is $\sum \Pr(X_i = 0)$ where X_i is the latent true status for the *i*th unit. The vacancy rate is estimated to be 3%, 2.8%, and 2.4% for each of the mailings, respectively. The sampling frame estimates a vacancy rate of 2.6% while the model estimates 5%. This suggests a downward bias in the estimates of proportion vacant for the observed indicators of vacancy.

Table 1: Vacancy Rates for Each Mailing, the Frame, and Model

	M_1	M_2	M_3	F	Model
Not Vacant	97.0	97.2	97.6	97.3	95.0
Vacant	3.0	2.8	2.4	2.6	5.0

Figure 2 provides the model estimates for the probabilities that an address classified by the USPS is in fact occupied (or vacant) according the model which is regarded as the gold standard or preferred indicator in this analysis. Thus, for addresses where no undeliverable mailing was returned by the USPS, 99% are occupied according to the model. However, among USPS "undeliverable vacant" addresses, 23% are actually occupied. In other words, the probability that an address is occupied given that it is coded as vacant by the USPS is 0.23. Among "other, undeliverables," the true occupancy rate (according to the model) is estimated to be 28 percent.

Now turning to the USPS frame, a similar analysis is conducted. Figure 3 provides the model estimates for the probabilities that an address classified by the frame is in fact occupied (or vacant) according the model which again is regard as the gold standard or preferred indicator in this analysis. Thus, for addresses classified as "occupied, not seasonal," 97% are truly occupied according to the model. Among frame "vacant" addresses, 59% are actually occupied. In other words, the probability that an address is occupied given that it is coded as vacant by the frame is 0.59 which is very high. Among addresses coded as "seasonal, occupied" by the frame, the true occupancy rate is estimated at 74% by the model.

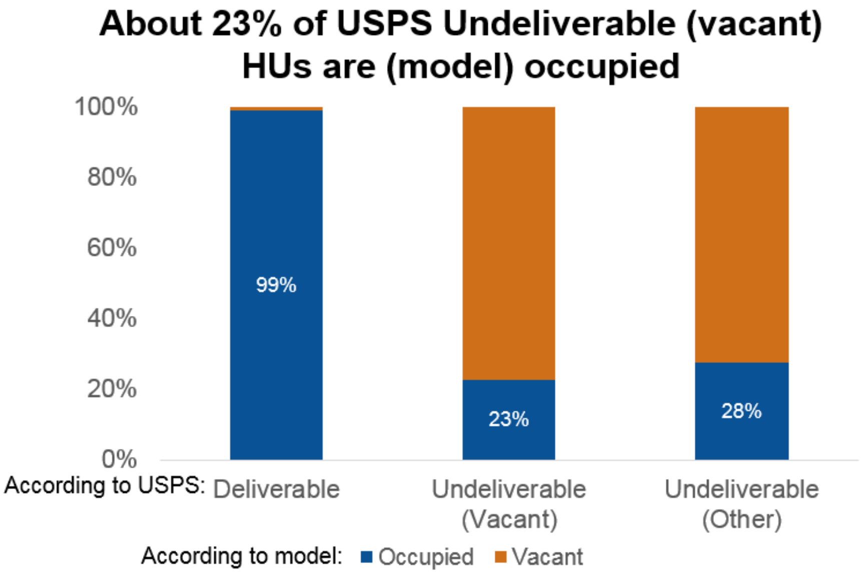
About 23% of USPS Undeliverable (vacant) HUs are (model) occupied

Figure 2. Proportion Occupied or Vacant (According to the Model) for Categories of the USPS Indicator

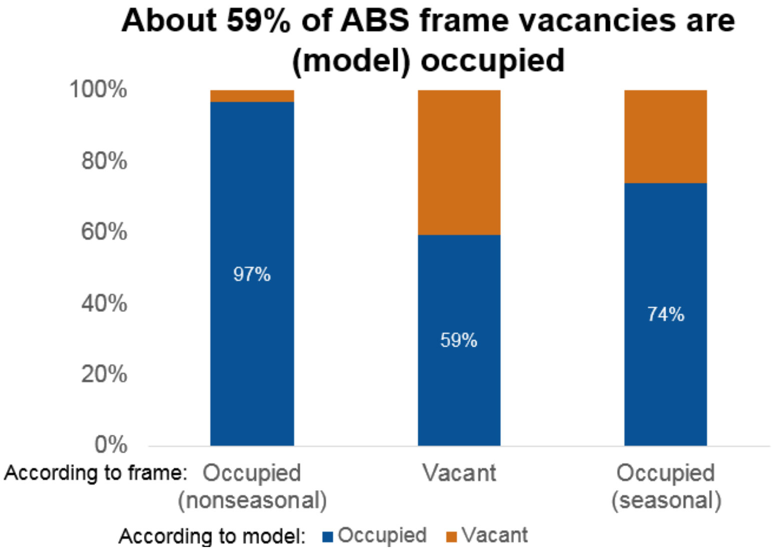

Figure 3. Proportion Occupied or Vacant (According to the Model) for Categories of the Frame Indicator

These results suggest that there is considerable error in the USPS and frame determinations of vacancy status. However, the latent class modeling approach can easily be applied after the first three mailings of any mail-contact survey assuming that the ABS frame indicator is also available. If the frame indicator is not available, the latent class model can still be fit to the three indicators from the first three mailings because the model is still identified. In addition, both the three and four indicator model can be further improved by incorporating the survey response as an additional indicator into the model where the probability the unit is vacant given the unit responds is constrained to be 0. The model could further be expanded by adding additional indicators of vacancy status that might become available for a survey.

4. Other Uses of LCA Indicator

Finally, we mention a few uses of the model and the indicator in practice. The indicator can be used during nonresponse follow-up to suggest cases that may not be worth pursuing further giving their likelihood of being vacant and therefore out of scope. It can also be used in the calculation of response and nonresponse rates where an estimated proportion of eligibility is needed, i.e.

$$\mathrm{RR} = \frac{\sum_{i\in S} R_i}{\sum_{i\in S} \Pr(i \text{ is occupied})} \qquad (0.2)$$

where S is the sample of addresses, and $R_i = 1$ if address i responded, 0 otherwise.

The probability of occupancy can be used as well in nonresponse weight via calibration. The idea here is that full sample totals that are used as calibration targets can be adjusted by the estimated occupancy rate, i.e. find nonresponse adjustment factors $f_i = f(\mathbf{x}_i{}^T\mathbf{g})$ satisfying:

$$\sum\nolimits_{i\in S} d_i R_i f_i \mathbf{x}_i = \sum\nolimits_{i\in S} d_i \Pr(i \text{ is occupied})\mathbf{x}_i \qquad (0.3)$$

where $\mathbf{x}_i$ is a vector of address i's frame characteristics, $\mathbf{g}$ is chosen to satisfy the above equation, and $f(\Box)$ is a function like $f(\mathbf{x}_i{}^T\mathbf{g}) = 1 + \exp(\mathbf{x}_i{}^T\mathbf{g})$, which corresponds to the assumption that unit response is a logistic function of the components of $\mathbf{x}_i$.

For the RECS-NP, the model was used in both the response rate calculation and in the weighting process. This application was only an initial step toward improving the classification of addresses as vacant or occupied that relies on just a single, imperfect indicator. Other indicators of vacancy status could be explored in the latent class context.

The latent class indicator has not be formally evaluated other than by considering the usual model diagnostics which suggest the model fits well. The validity of the model could be further explored by comparing the latent indicator with ground truth collected by interviewers in a mixed mode survey that includes in-person field component. This would permit a more objective evaluation of the validity and utility of the model for estimating vacancy and its effect on survey estimates. However, given the direction the RECS has taken with the recent incorporation of both in-person and mail-contact, self-administered modes, it is going to be important to know early in the data collection whether a unit is likely to be eligible for the survey and to have reliable methods for assessing eligibility when direct evidence is not available from an interviewer making a visit to the unit.

References

Biemer, P., Murphy, Joe, Zimmer, Stephanie, Berry, Chip, Deng, Grace, Lewis, Katie (2016) "The Choice + Protocol for Web/Mail Surveys: Some Empirical Results," Paper presented at 2016 AAPOR in Austin, Texas.

Wiant, K. F., McMichael, J. P., Murphy, J. J., Morton, K. B., & Waggy, M. R. (2016). Consistency and accuracy of undeliverable codes provided by the U.S. Postal Service: implications for frame construction, data collection operational decisions and response rate calculations. Presented at Annual meeting of the American Association for Public Opinion Research, Austin, TX, May 2016.