Predicting Proxy Status in Nonresponse Followup Workload

Andrew Keller[1]
[1]U.S. Census Bureau, 4600 Silver Hill Rd., Washington, DC 20233

Abstract

For the 2010 Census, the Nonresponse Followup (NRFU) operation verified the status of units that did not respond by mail in areas that received a mailback census questionnaire. The NRFU workload totaled 47 million units at a cost of $1.6 billion. The Census Bureau then classified these units into four categories: occupied, vacant, deleted, and unresolved. Among occupied, vacant, and deleted NRFU units, respectively, 24%, 98%, and 90% of interviews were completed via proxy respondents[2]. In general, survey results often demonstrate reduced data quality when allowing proxy respondents for household interviews. By understanding the unit- and area-level characteristics that lead to proxy respondents, we can better understand where lower data quality might be present in census results. This research implements a modeling approach to predict proxy status. The approach relies on housing unit-level covariates defined by census operations prior to beginning NRFU fieldwork. In addition, the model uses tract-level characteristics known prior to the census as independent variables to predict proxy status for the NRFU universe.

According to the Census Coverage Measurement results, proxy status results in greater erroneous enumerations. Hence, I develop a second model to predict the presence of erroneously enumerated persons within the NRFU universe. The goal is to confirm that the covariates predicting proxy status also predict erroneous enumeration status. By contrasting the covariates that predict proxy status and not erroneous enumeration status, we may find characteristics for units with proxy interviews but not erroneous enumerations.

---

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.

[2] By definition, all vacant and deleted units should be completed by proxy. However, since the NRFU was completed via a paper questionnaire, some of these units were mistakenly marked as completed by a non-proxy.

1. Background

The 2010 Nonresponse Followup Operation occurred from mid-April until the middle of August and was composed of four field operations: Nonresponse Followup (NRFU), Nonresponse Followup Reinterview (NRFU RI), Nonresponse Followup Vacant Delete Check (VDC), and Nonresponse Followup Residual (Residual). The bulk of the 2010 Nonresponse Followup Operation was concentrated on NRFU. NRFU RI was designed as a quality check on NRFU enumerators' results. VDC verified vacant and delete housing units from the NRFU operation as well as a first time enumeration of some units that were not in the original NRFU universe. The Residual operation was intended to (1) obtain population counts for units that were known to be occupied but without population counts and (2) enumerate units that were not in the original NRFU or VDC universes. See Walker et al. (2012) for more information on the 2010 Nonresponse Followup Operation. Because the bulk of the work is concentrated in NRFU, this research is restricted to the NRFU universe.

A NRFU proxy provided information about the NRFU address but the respondent was not a household member. In the case of an occupied unit, NRFU enumerators were permitted to speak with a proxy respondent if a household member was not available following three attempts to contact someone in the household. The 2010 NRFU operation relied upon proxy interviews for about one-quarter of occupied units.

The 2010 Census Coverage Measurement (CCM) program evaluated coverage of the 2010 Census to aid in improving future censuses. The CCM measured the net coverage and components of census coverage of housing units and persons, excluding group quarters and persons residing in group quarters. The CCM sample design was a probability sample of approximately 170,000 housing units. Remote areas of Alaska were out of scope for the CCM.

Keller and Fox (2012) show 2010 components of census coverage that include estimates of correct enumerations and erroneous enumerations for persons enumerated in households responding to NRFU via a proxy. In short, a higher estimate of erroneous enumerations occurred for persons enumerated in units responding via a proxy. In addition, proxy respondents generally know less of the detailed demographic information about household members for which they are trying to provide information. As a result, units responding via a proxy produce a higher whole-person imputation rate than those responding via household member.

2. Introduction

This research uses two logistic models. The goal of the first model is to predict proxy status before going into the field to complete a NRFU interview. The second model predicts units that provide erroneous enumerations before going into the field to complete a NRFU interview. Both models use data from census operations and the American Community Survey (ACS). In addition, the second model incorporates data from the CCM.

3. Data

To begin, I make an informative vs. non-informative proxy distinction if (a) the enumerator completed the interview with a non-household member and (b) at least one person in the unit was data-defined[3]. If (a) and (b) are true, then the respondent is considered an informative proxy. The idea is that, if the Census Bureau knew it would be talking to a non-informative proxy or the unit was vacant, delete, or unresolved, it could consider alternative means to complete the NRFU interview.

Walker et al. (2012) defines the size of the NRFU universe as 47,197,405 units. For this research, I exclude Puerto Rico. This results in 46,436,683 units. I use a multinomial logistic model to predict proxy status. The dependent variable has six categories of proxy status. They are: (1) household member, (2) occupied informative proxy, (3) occupied non-informative proxy, (4) vacant status (through proxy information), (5) delete status (through proxy information), and (6) unresolved status. The frequencies of the six categories of NRFU proxy status are in Table 1.

Table 1: Observed NRFU Proxy Status

Response Type	Household member	Occupied Informative proxy	Occupied Non-informative proxy	Vacant	Delete	Unresolved	Total
Count	21,396,963	5,349,487	1,577,434	13,937,902	3,924,147	250,750	46,436,683
Percentage	46.1%	11.5%	3.4%	30.0%	8.5%	0.5%	100.0%

Independent of the proxy status, it might be helpful if we could identify units prior to a visit by a NRFU enumerator that would more likely result in erroneous enumerations. Like the non-informative proxy, if the Census Bureau knew it would be talking to a respondent that would result in erroneous enumerations, it could consider alternative means to complete the NRFU interview.

The CCM program estimated correct and erroneous enumerations in the census from a sample. Hence, instead of using a universe of 46.4 million units, the dataset for the second model consists of NRFU units in the CCM sample. This results in 37,100 records. Of these, there were 28,377 units without any erroneously enumerated persons and 8,723 units with at least one erroneously enumerated person.

Because I want to predict proxy and erroneous enumeration status before doing a household visit, I use independent variables in the model that were observable before the NRFU operation began. The same variables are in both models.

[3] A census person is data-defined if it has at least two characteristics (name counts as a characteristic).

Table 2: Independent Variables Used in Models

Independent Variable	Description	Possible Values
Multi-Unit Structure	Whether the unit is part of a multi-unit structure	1 – multi-unit
UAA[4] - Vacant	United States Postal Service (USPS) assigned undeliverable-as-addressed (UAA) Vacant indicator after mailing initial questionnaire (IQ)	1 – vacant
UAA - No Such Number	USPS assigned UAA No Such Number indicator after mailing IQ	1 – no such number
UAA - No Such Street	USPS assigned UAA No Such Street indicator after mailing IQ	1 – no such street
UAA - No Mail Receptacle	USPS assigned UAA No Mail Receptacle indicator after mailing IQ	1 – no mail receptacle
UAA - Attempted Not Known	USPS assigned UAA Attempted Not Known indicator after mailing IQ	1 – attempted, not known
Adjoining NRFU Case	Indicates whether NRFU case was next to another NRFU case	1 - NRFU case adjoining
Not Valid Living Quarters	Unit was marked as <u>not</u> valid living quarters prior to census by Master Address File	1 – not valid living quarters
Low Rural Rate (from ACS)	Unit in tract that is not rural (25th percentile)	1 – Yes
Low White Population Rate (from ACS)	Unit in tract with non-Hispanic white population less than 42.89% (25th percentile)	1 – Yes
Replacement Mailing Block	Block was scheduled to receive replacement mailing	1 – Yes

4. Methodology

<u>Predicting NRFU Proxy Status Model</u>

Because there are 46.4 million units, I randomly divide the universe into ten mutually exclusive partitions. For each partition, I fit the model and then apply the coefficients over the remaining 90% of data. For the 90%, I then compare the predicted proxy status against the observed proxy status to assess model accuracy. For example, I have 46,436,683 units. The first 10% partition over which I fit the model has 4,642,860 cases. I fit the model and apply the coefficients to the remaining 41,793,823 cases.

Because I am running a multinomial logistic model, I generate six predicted probabilities for each unit. Each of the 90% cases gets six probabilities that sum to 1. That is, $P(DV = 1) + \ldots + P(DV = 6) = 1$, where DV is the six category dependent variable of proxy status. The cumulative probability is compared to a uniform random number between (0, 1) and then assigned a predicted outcome status as given in Table 1. For example, suppose the predicted and cumulative probabilities are in Table 3.

[4] For more information on UAA assignment, see the Postal Tracking Data Repository Getting Started Guide (Gunnison, 2012).

Table 3: Example Predicted Probability From Proxy Model

Probability	Response Type					
	Household member	Occupied Informative proxy	Occupied Non-informative proxy	Vacant	Delete	Unresolved
Predicted	0.50	0.10	0.05	0.20	0.13	0.02
Cumulative	0.50	0.60	0.65	0.85	0.98	1

Further, suppose the random number is 0.788. Then, the predicted outcome status is vacant.

Predicting NRFU Erroneous Enumeration Model

I use the same technique to estimate the erroneous enumeration model. I first select two partitions. For each partition, I fit the model and then apply the coefficients over the remaining 50% of data. For the 50%, I then compare the predicted erroneous enumeration status against the observed erroneous enumeration status to assess model accuracy. Because I am running a binary logistic model, I generate a predicted probability indicating the probability of the NRFU unit having an erroneous enumeration. The probability is compared to a uniform random number between (0, 1) and then assigned a predicted outcome status.

5. Results

One goal of this project is to get an understanding for the best predictors of proxy status and erroneous enumeration for NRFU units. The idea is that, if we were to know which units would be proxies or give erroneous enumerations, we could make adjustments in the field. To get an understanding for the best predictors, we use the odds ratios from both models.

Predicting NRFU Proxy Status Model

The primary goal of this model is to understand the relationship between the covariates and the prediction of NRFU proxy status. For this model, a household member serves as the reference group. Table 4 displays the odds ratios from the independent variables in Table 2. Nearly every odds ratio differs from 1 due to the large partition size. Because I run the same model over each of the ten partitions, ten odds ratios exist for each independent variable. In the table below, I report the average odds ratio across the ten models.

Table 4: Average Odds Ratios for Predicting NRFU Proxy Status Model

Variable	Occupied Informative proxy	Occupied Non-informative proxy	Vacant	Delete	Unresolved
Multi-Unit Structure	1.87	2.00	1.42	2.12	1.55
UAA - Vacant	1.34	2.00	12.28	8.32	9.08
UAA - No Such Number	0.78	0.76	1.58	7.50	2.69
UAA - No Such Street	0.68	0.81	2.07	2.54	1.61
UAA - No Mail Receptacle	0.66	0.54	1.35	1.47	1.03#
UAA - Attempted Not Known	1.06#	1.30	5.08	5.16	3.75
Adjoining NRFU Case	1.02	1.14	1.44	1.69	1.49
Not Valid Living Quarters	1.07#	1.41	6.59	40.66	25.84
Low Rural Rate	1.13	1.27	0.62	0.49	0.56
Low White Population Rate	0.85	1.07	0.90	1.11	1.33
Replacement Mailing Block	1.08	1.17	1.30	1.32	1.43

means not significant at 95% level

First, it is interesting to look at any independent variable where all proxy categories have an odds ratio greater than 1. For example, look at the odds ratios for multi-unit structure. In this case, the respondent at a multi-unit is always more likely to be someone other than a household member when compared to single units. That is, a respondent in a multi-unit is 1.87 times more likely than a respondent in a single unit to be an informative proxy when compared to a household member. Likewise, a respondent in a multi-unit is 2.00 times more likely than a respondent in a single unit to be a non-informative proxy when compared to a household member.

Similarly, look at the odds ratios for the variable indicating the United States Postal Service (USPS) assigned the unit an undeliverable-as-addressed (UAA) Vacant indicator after the Census Bureau mailed the initial questionnaire. Again, all the odds ratios are greater than 1. In this case, the UAA - Vacant indicator reveals that the respondent for the unit is always more likely to be someone other than a household member when compared to units without the UAA - Vacant indicator.

I measure model quality in two ways. First, I measure model agreement by dividing the number of cases with the same observed and modeled category values by total number of cases with the observed category value. I also measure agreement using the Kappa inter-rater agreement statistic (Agresti, 1990). I use the Kappa statistic where one rater is the observed category determined by the NRFU operation. The other rater is the modeled category. Table 5 introduces the notation used for calculating inter-rater agreement.

Table 5: Inter-rater Agreement Table

Observed Category	Modeled Category		Total
	Yes	No	
Yes	x_{11}	x_{12}	x_{1+}
No	x_{21}	x_{22}	x_{2+}
Total	x_{+1}	x_{+2}	x_{++}

So, using the notation introduced above, model agreement is calculated as $A = \frac{x_{11}}{x_{1+}}$. To determine the inter-rater agreement, the following formula is used:

$$K = \frac{\frac{x_{11} + x_{22}}{x_{++}} - \frac{x_{1+}x_{+1}}{x_{++}x_{++}} - \frac{x_{2+}x_{+2}}{x_{++}x_{++}}}{1 - \frac{x_{1+}x_{+1}}{x_{++}x_{++}} - \frac{x_{2+}x_{+2}}{x_{++}x_{++}}}$$

Kappa takes on any value between -1 and 1. In this context, a '1' value would mean that the observed and modeled results would be exactly the same. A '0' value would mean that the model predicts the observed status by chance. A '-1' value would mean that the observed and modeled results would be always different.

For each of the ten models, I have six inter-relater agreement values, one for each of the six prediction categories. Suppose I predict whether it is a household member. So, using the notation from earlier:
x_{11} : observed household member, modeled household member
x_{12} : observed household member, modeled non-household member
x_{21} : observed non-household member, modeled household member
x_{22} : observed non-household member, modeled non-household member

For the first 10% partition, I fit the model over 4,642,860 units. I then predict the proxy status for the other 41,793,823 units. I do this for each of the six prediction categories. The first eight rows of Table 6 give the values for $x_{11}, x_{12}, x_{21}, x_{22}, x_{1+}, x_{++}, A, K$ for the first 10% partition. The same calculations can be completed for the other nine partitions. Since I run the model over ten partitions, I determine an average for A and K. These are denoted as $\overline{A}$ and $\overline{K}$, respectively and are listed in the bottom two rows.

Table 6: Prediction Results for NRFU Proxy Model

Statistic	Household member	Occupied Informative proxy	Occupied Non-informative proxy	Vacant	Delete	Unresolved
x_{11}	9,994,774	644,766	58,171	4,823,147	561,934	1,729
x_{12}	9,264,911	4,168,990	1,360,947	7,720,067	2,970,537	223,850
x_{21}	9,246,792	4,179,494	1,367,684	7,727,698	2,963,296	224,338
x_{22}	13,287,346	32,800,573	39,007,021	21,522,911	35,298,056	41,343,906
x_{1+}	19,259,685	4,813,756	1,419,118	12,543,214	3,532,471	225,579
x_{++}	41,793,823	41,793,823	41,793,823	41,793,823	41,793,823	41,793,823
A	51.89%	13.39%	4.10%	38.45%	15.91%	0.77%
K	0.1086	0.0209	0.0071	0.1203	0.0817	0.0023
$\overline{A}$	51.94%	13.37%	4.09%	38.44%	15.94%	0.75%
$\overline{K}$	0.1087	0.0209	0.0071	0.1205	0.0818	0.0021

It is interesting to note the difference between $\overline{A}$ and $\overline{K}$. Often, Kappa values between 0.01 and 0.25 indicate minor agreement. However, Viera and Garrett (2005) note that categories with low frequency can have low Kappa values. Because some of these categories have low frequency, the Kappa score is only one consideration in model evaluation.

Using NRFU Proxy Status Model to Inform Field Decisions

We develop the model to inform us about the efficacy of NRFU visits. Conceptually, we would like to visit units that would provide accurate information. Conversely, we would like to minimize the number of units we visit that provide little information as compared to speaking with a knowledgeable respondent. We divide the six categories from Table 1 by these concepts.

In this analysis, we define ‘high information’ units as those where we spoke with a household member or occupied informative proxy respondent. We define ‘low information’ units as those where we spoke with occupied non-informative proxy respondents. We also define vacant, delete, or unresolved units as ‘low information’ units because we could possibly pursue that information through other means without having to visit the unit. Hence, for each unit, I collapse the model predictions into two groups:

P(High Information) = P(Household Member) + P(Informative Proxy)
P(Low Information) = P(Non-Informative Proxy) + P(Vacant) + P(Delete) + P(Unresolved)

From the proxy status model above, each unit has a probability of high and low information. There were 65,046 tracts in NRFU. I average the two probabilities, *P(High Information)* and *P(Low Information)*, separately across all units in the tract. This results in *AVG(P(High Information))* and *AVG(P(Low Information))* at the tract level. Using the model, we identify the tracts providing the least helpful information by those with the highest *AVG(P(Low Information))* values.

Table 7 shows a hypothetical tract. In this tract, we completed 100 NRFU interviews. We observe that 35 interviews came from high information respondents (household members or informative proxy respondents). In total, from Table 1, there were 46.4 million NRFU units. Among these units, I classified 26,746,450 as high information units. For 2010, each NRFU unit was allowed up to six visits. In total, about one hundred million NRFU visits were made. In this tract, the 100 units resulted in 220 visits. Across these 100 NRFU units, the NRFU proxy status model predicts that the average predicted probability of low information across all units in the tract is 0.70.

Table 7: Example Tract

Tract	NRFU Units	High Information Units Observed	Visits Made	*AVG(P(Low Information))*
Example	100	35	220	0.70

As a simulation, I set various cut points for tracts based on *AVG(P(Low Information))*. We then look at the tracts affected, NRFU units affected, high information NRFU units affected, and visits made to the affected NRFU units. In short, we would like to maximize number of NRFU units affected and visits made to those affected units while minimizing the number of high information NRFU units affected.

Table 8 summarizes the effects of setting various cut points based on the predicted probability. For example, for all tracts where *AVG(P(Low Information))* > 0.5, then 15.8% of the tracts and 19.3% of the NRFU workload is affected. This includes 15.3% of the high information NRFU units and 17.7% of the visits.

Table 8: Cut Point Simulation of NRFU Units Affected from NRFU Proxy Status Model

Cut Point – Tracts with Predicted Low Information Probability Above…	Tracts Affected		NRFU Units Affected		High Information NRFU Units Affected		Visits Made to Affected NRFU Units	
	Total	Percent	Total	Percent	Total	Percent	Total	Percent
0.90	5	0.0%	1,370	0.0%	46	0.0%	1,833	0.0%
0.85	14	0.0%	3,463	0.0%	336	0.0%	4,985	0.0%
0.80	31	0.1%	22,143	0.1%	1,093	0.0%	29,967	0.0%
0.75	57	0.1%	47,089	0.1%	3,501	0.0%	79,694	0.1%
0.70	124	0.2%	126,487	0.3%	15,736	0.1%	231,947	0.2%
0.65	342	0.5%	315,863	0.7%	62,632	0.2%	591,293	0.6%
0.60	1,047	1.6%	957,725	2.1%	272,148	1.0%	1,909,977	1.9%
0.55	3,538	5.4%	3,123,568	6.7%	1,192,694	4.5%	6,333,335	6.2%
0.50	10,251	15.8%	8,947,021	19.3%	4,082,664	15.3%	18,226,263	17.7%

<u>Predicting NRFU Erroneous Enumeration Model</u>

Table 9 displays the odds ratios from the independent variables in Table 2. Note that nearly every odds ratio differs from 1. Similar to the proxy status model, I run the same model over the two partitions. In the table below, I report the odds ratio for each model run and the average odds ratio across the two models. For each variable, the 95% confidence intervals on the odds ratio overlap between the two partitions. Hence, I will use the average to describe the odds ratio for each variable.

Table 9: Odds Ratios for Predicting NRFU Erroneous Enumeration Model

	Odds Ratio		
Variable	Partition 1	Partition 2	Average
Multi-Unit Structure	1.46	1.65	1.55
UAA - Vacant	1.70	1.87	1.79
UAA - No Such Number	1.53	1.59	1.56
UAA - No Such Street	0.60	0.55	0.57
UAA - No Mail Receptacle	0.48	0.39	0.43
UAA - Attempted Not Known	2.02	3.30	2.66
Adjoining NRFU Case	1.20	1.09	1.14
Not Valid Living Quarters	4.58	4.72	4.65
Low Rural Rate	1.22	1.05#	1.13
Low White Population Rate	1.20	1.22	1.21
Replacement Mailing Block	1.19	1.22	1.21

means not significant at 95% level

Holding everything fixed, a NRFU unit is 1.55 times more likely to have an erroneous enumeration if it is part of a multi-unit structure. Holding everything fixed, a NRFU unit is 1.79 times more likely to have an erroneous enumeration if the USPS assigned a UAA - Vacant indicator. Holding everything fixed, a NRFU unit is 4.65 times more likely to have an erroneous enumeration if the Master Address File did not classify the NRFU address as a valid living quarters.

Again, I use inter-relater agreement to check model quality. For this model, I have a binary dependent variable, the indication of an erroneous enumeration in the NRFU unit. So, using the notation from above

x_{11} : observed erroneous enumeration in the unit, modeled erroneous enumeration in the unit
x_{12} : observed erroneous enumeration in the unit, modeled no erroneous enumeration in the unit
x_{21} : observed no erroneous enumeration in the unit, modeled erroneous enumeration in the unit
x_{22} : observed no erroneous enumeration in the unit, modeled no erroneous enumeration in the unit

For the first 50% partition, I fit the model over 18,583 units. I then predict the erroneous enumeration status for the other 18,517 units. For the second 50% partition, I fit the model over 18,517 units. I then predict the erroneous enumeration status for the other 18,583 units. Table 10 gives the values similar to Table 6.

Table 10: Prediction Results for NRFU Erroneous Enumeration Model

Statistic	Partition 1	Partition 2
x_{11}	1,144	1,138
x_{12}	3,261	3,180
x_{21}	3,143	3,261
x_{22}	10,969	11,004
x_{1+}	4,405	4,318
x_{++}	18,517	18,583
A	26.17%	26.05%
K	0.0373	0.0347

Again, it is interesting to note the difference between *A* and *K*. For example, although the model predicts that the unit will have a erroneously enumerated household member accurately about 26% of the time, $K \approx 0.036$.

One goal of this research was to contrast the results from the two models. In general, the results support one another. That is, covariates showing higher odds of erroneous enumerations in the unit also show higher odds for conducting a proxy interview. For example, odds ratios greater than 1 for the erroneous enumerations model are consistent with the odds ratios predicting a proxy interview for multi-units and units with a UAA - Vacant indicator. One contrast exists for NRFU units in the low rural rate tracts. From Table 4, a NRFU unit in a low rural rate tract is more likely to be an informative or non-informative proxy when compared to a household member. However, from Table 9, holding everything fixed, the odds are not necessarily higher that the NRFU address contains an erroneous enumeration in the low rural rate tracts.

6. Conclusions and Future Work

This research explores two models dealing with NRFU units – a multinominal logit model that predicts the type of respondent and a binomial logit that predicts the presence of erroneous enumerations in the unit. I develop both models from covariates that are observable prior to conducting the NRFU interview. In general, past CCM results show a correlation between proxy responses and erroneous enumerations. That is, interviewing a proxy respondent results in more erroneous enumerations and imputations. I show the covariates that predict both proxy and erroneous enumeration status. In addition, I note covariates that predict a proxy unit, but not necessarily the evidence of erroneous enumerations. Future covariates can include census block, county, and state-level rates for certain characteristics.

The results of both models can be applied to the decision of whether a visit by a NRFU enumerator is prudent for the unit in question. In this paper, I generate the model predicting proxy status and then run a simulation using those model predictions to show implications on field work. The simulation shows an example of how field work can most effectively be reduced and its effect on the type of interviews received.

References

Agresti, A. (1990), "Categorical Data Analysis," New York: Wiley Interscience.

Gunnsion Consulting Group, Inc. (2012), "Postal Tracking Data Repository Getting Started Guide," Gunnsion Consulting Group, Inc.

Walker, S., Winder, S., Jackson, G., and Heimel, S. (2012), "2010 Census Nonresponse Followup Operations Assessment," DSSD 2010 Decennial Census Memorandum Series #2010-O-C-15.

Keller, A. and Fox, T. (2012), "2010 Census Coverage Measurement Estimation Report: Components of Census Coverage Results for the Household Population in the United States," DSSD 2010 Census Coverage Measurement Memorandum Series #2010-G-04.

Viera, A.J. and Garrett, J.M. (2005), "Understanding the Interobserver Agreement: The Kappa Statistic," Family Medicine, 37(5), 360-363.