# Evaluating the Census Planning Database, MSG, and Paradata as Predictors of Household Propensity to Respond

Xiaoshu Zhu[1], Robert Baskin[1], David Morganstein[1]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850

**Abstract**
As survey response rates drop, researchers are seeking improved methods to predict respondents' propensity to respond and to guide an adaptive design effectively. The Census Bureau created the 2015 Planning Database (PDB) as a research tool that includes selected 2010 Census variables, ACS estimates, and a Low Response Score (LRS) based on response to the 2010 Census at the census block-group and the tract levels. MSG data is a proprietary database of household-level information. Westat selected two surveys that are multistage, area cluster samples of households, designed to produce samples representing the U.S. non-institutionalized population. In both surveys the second level of sampling is area segments comprising Census blocks so the PDB and MSG data are good candidates for modeling segment-level response rates at two stages. We explore whether the PDB or MSG provides useful yield estimates at the household or individual level. The actual results were compared to predictions from the PDB. In our examples, the LRS contained little predictive power. While these results may apply only to these two Westat surveys, these unexpected outcomes may be of interest to other surveys.

**Key Words:** Census Planning Database, MSG, Response Propensity, Logistical Models, Machine Learning Models

## 1. Introduction

This work explored the prediction of household response propensity for in-person surveys, both prior to the start of the survey and during the survey field period, by attempting to fit known survey outcomes from two completed face-to-face household surveys. Prediction of the final response rate is desired before the sample is fielded in order to allocate resources and prepare for field work. Once the survey begins, revised predictions of final response rates can be based on paradata about interview attempts and household responses to those attempts and used in adaptive designs. We used the results from two surveys where the household response was known and attempted to fit data available before the survey began to those outcomes. This was done using cross-validation, fitting the actual outcomes from a sample of records and testing the predictive fit on others. We then repeated the process, revising those models using paradata obtained during the survey field period to see if it would improve the fit.

The sampling frames used to draw the household samples were derived from sampling commercially available address lists based on information from the United States Postal Service's (USPS's) Address Management System (AMS) and in rural areas supplemented by traditional listing. Because 2010 Census information was used to form the segments and segment strata, it was conjectured that the Census information might be useful in predicting nonresponse at the block group or higher level.

For confidentiality reasons the two surveys used in this study are referred to as Survey X and Survey Y. Both of these surveys use a three-stage sample design. They each start with an area survey of Primary Sampling Units, which are counties or groups of counties, followed by a second stage selection of area segments, which are Census blocks or groups of contiguous blocks, with a final stage that is address selection within each segment. Both surveys sampled over 5,000 block groups and 50,000 addresses.

For predicting the pre-field period, we examined the Census Planning Database (PDB) and variables from Marketing Systems Group (MSG) for use in fitting the known outcomes. Since 2000, the Census Bureau has published PDB by assembling measures on housing, demographic characteristics, and socioeconomic status. The PDB also includes census operational data such as mail return rate. In 2014, the Census Bureau conducted crowdfunded research to determine the best models for predicting 2010 Census mail return rates at the block group and tract level. The Census Bureau used this information to construct a Low Response Score (LRS), which was first included in the 2014 PDB. Erdman and Bates (2017) describe research using the PDB to predict the 2010 Census mail return rates. The research involved creating a Hard to Contact score based on the mail return rate. Westat explored the utility of the LRS in predicting nonresponse in other household surveys. In this study, we selected 214 variables from the 2016 PDB, which assembled variables from 2010 Census and 2014 5-year American Community Survey (ACS).

In addition to address listing, MSG appended household characteristics such as household composition, tenure, and income, as well as characteristics of head of household including age, race/ethnicity and education, to the sampled households' address. There are several studies discussing the quality of the appended variables and their use in sampling and response prediction (Brick, Lohr, Edwards, Giambo, Broene et al., 2013; Montaquila, 2014; Roth, Han, & Montaquila, 2013). One issue with the MSG variables is completeness. As a result, we implemented the methods that can properly handle missing values when analyzing the thirteen appended MSG variables.

During the field period, Westat collects paradata on each contact attempt to the individual household. We evaluated the predictive power of the paradata information. The paradata records when and how a contact happens, and how it ends. Westat uses a detailed coding system for interim status of contact attempt, including interim refusal, appointment made, resident not home or not available, etc. For analysis purpose, we grouped the interim status into five general categories and derived variables that indicated the number of all contact attempts, by contact types, and contacts happened at specific time of day, week, and month.

In Sections 2, we examine the value of the PDB and MSG information to predict participation using sources of data available prior to data collection. We first used correlations alone and then modeling with least absolute shrinkage and selection operator (LASSO) and random forest model. We attempt, in Section 2.1, to predict response rates of each survey using variables in the PDB. We built statistical models predicting response at the block group level using variables from the PDB as predictors. The predicted rates from these models were then compared to the actual survey response rates at the block group. Section 2.2 moves to prediction of response by incorporating commercial data from the MSG appended at the household level. In Section 3, we add in survey outcomes obtained after data collection begins. We attempt to predict unobserved household response status based on partially observed response status and by adding in household-level predictors such as paradata, in addition to the PDB and MSG variables. This approach did show promise for future use.

## 2. Prediction Prior To Data Collection

### 2.1 Block Group Level Response Rate

*2.1.1 Correlation with LRS*

We began by looking at the correlation between the block group level LRS and response rates observed in the studies. The LRS has a weak linear correlation with response rates, at both the national and state levels, as shown in Table 1. The scatterplots using randomly sampled 1,000 block groups from each survey are presented in Figure 1, since the scatterplot for the entire country is too dense. Neither of the plots shows a clear association between the two measures.

**Table 1:** Correlation between block group level LRS and survey response rates

| | *Study X* | *Study Y* |
|---|---|---|
| National | 0.147 | 0.146 |
| State 1 | 0.056 | 0.144 |
| State 2 | 0.274 | 0.261 |
| State 3 | 0.234 | 0.220 |
| State 4 | 0.273 | 0.093 |

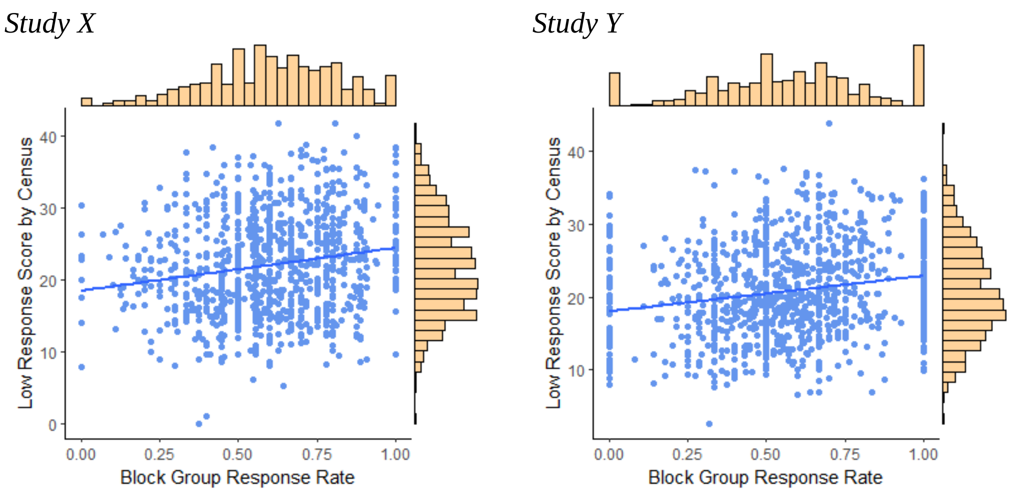
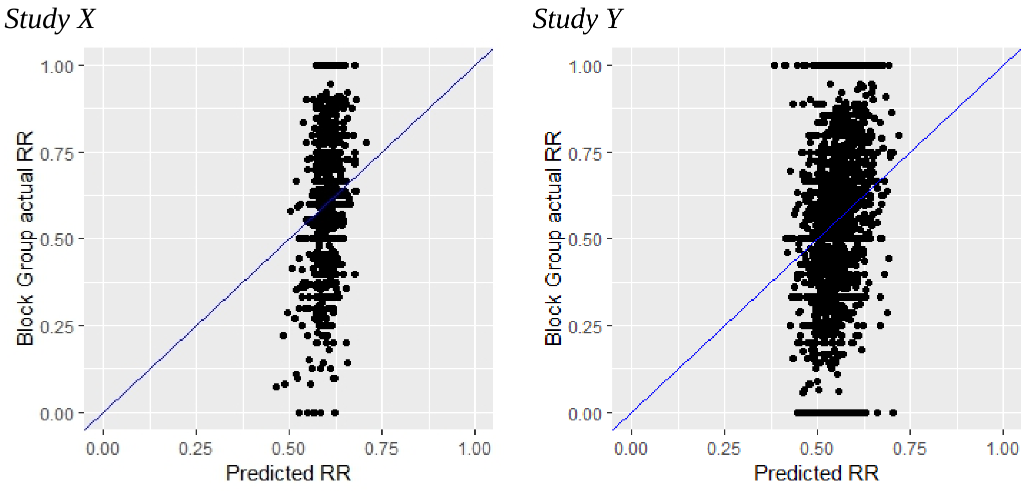


**Figure 1:** Scatterplots between block group level LRS and response rates using 1,000 random sample

*2.1.2 LASSO with PDB variables*

We next ran two LASSO models with the 214 PDB variables using the package “glmnet” in R to test whether the LRS would be selected as an important predictor for response rates. LASSO is a shrinkage and selection method for linear regression. The coefficients of the less contributive variables are forced to be zero, leaving only the most significant ones in the final model. We randomly selected 80% of the sample in Study X and 70% of Study Y to train the model and used the remaining sample were used to evaluate the model fit.

Given that the root mean square error (RMSE) was 0.195 for Study X and 0.255 for Study Y, and the scatterplots between the predicted and observed response rates (shown in Figure 2), both results indicated that the model was not working well for either study.


*Study X*

*Study Y*

**Figure 2:** Scatterplots between predicted and observed response rates in test set using LASSO

Even though the models performed poorly, some variables were more informative than the others. The final model kept 7 and 17 variables for Study X and Study Y, respectively, and the LRS was not in either model (see Table 2). Many selected variables pertained to household economic status. By comparing the two lists of selected variables, we found two common variables (highlighted in bold), pct_pop_25yrs_over_acs_10_14 (Percentage of the ACS population who are ages 25 years and over) and pct_rel_under_6_cen_2010 (Percentage of 2010 Census family-occupied housing units with a related child under 6 years old).

**Table 2:** Variables kept in the final model (not ranked)

| *Study X* | *Study Y* |
|---|---|
| aggr_house_value_acs_10_14[1] | med_hhd_inc_bg_acs_10_14 |
| **pct_pop_25yrs_over_acs_10_14** | med_hhd_inc_tr_acs_10_14 |
| avg_agg_hh_inc_acs_10_14 | aggregate_hh_inc_acs_10_14 |
| avg_tot_prns_in_hhd_cen_2010 | mlt_u2_9_strc_acs_10_14 |
| **pct_rel_under_6_cen_2010** | mlt_u10p_acs_10_14 |
| pct_tot_occp_units_cen_2010 | med_house_value_bg_acs_10_14 |
| pct_vacants_cen_2010 | pct_nh_sor_alone_acs_10_14 |
| | **pct_pop_25yrs_over_acs_10_14** |
| | pct_prs_blw_pov_lev_acs_10_14 |
| | pct_eng_vw_indoeuro_acs_10_14 |
| | pct_pub_asst_inc_acs_10_14 |
| | valid_mailback_count_cen_2010 |
| | pct_urbanized_area_pop_cen_2010 |
| | pct_pop_under_5_cen_2010 |
| | pct_nh_sor_alone_cen_2010 |
| | pct_female_no_hb_cen_2010 |
| | **pct_rel_under_6_cen_2010** |

*Note:* The common variables are highlighted in bold.

### *2.1.3 Random forest with PDB variables*

LASSO assumes a linear relationship between the outcome and predictors, and in the previous model we only evaluated main effects. Random forest, on the other hand, doesn't make assumptions on the nature of the relationship, and it can account for high-order interactions. In addition, random forest is affected less by multicollinearity among the PDB variables than LASSO. We used the same set of data used in LASSO to train and test the random forest model using the "rf" option in the package "caret" in R.

The model performance was similar with the LASSO result in terms of the RMSE (0.194 for Study X and 0.250 for Study Y) and the estimated response rates in the test sets (see Figure 3), although different set of important variables were selected in the random forest. Again, the LRS never entered the top 25 most important list (see Table 3). Among the top 25 most important variables, seven variables appeared on both lists (highlighted in bold). The two lists share variables pertaining to household with or without young-age children, and economic status.

The poor model performance in both LASSO and random forest suggests that the PDB variables do not offer much predictive power on response rate at the block group level for either of the two studies.

[1] The label of PDB variables are available here: https://www.census.gov/research/data/planning_database/2016/docs/2016-Block-Group-PDB-Documentation-V8.pdf

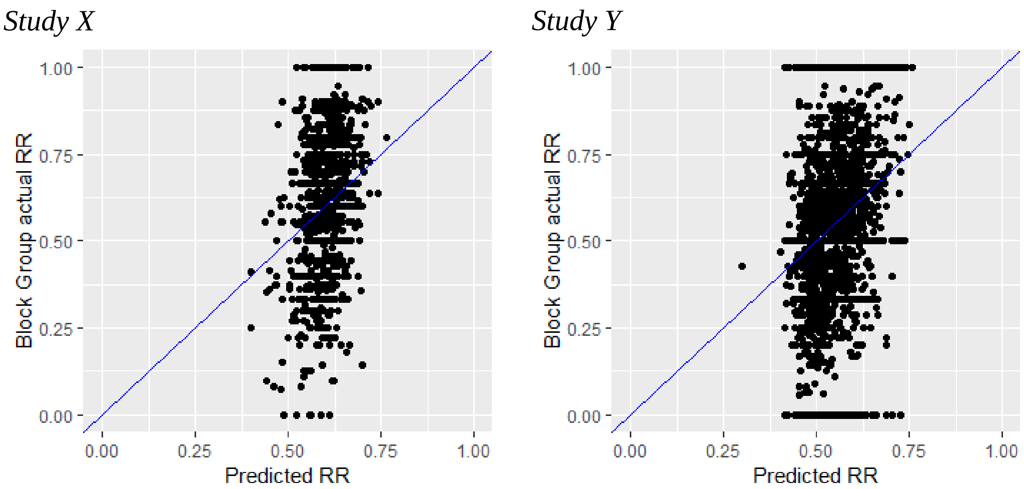


**Figure 3:** Scatterplots between predicted and observed response rates in test set using random forest

**Table 3:** The top 25 most important variables in random forest and importance score

| | *Study X* | | *Study Y* | |
|---|---|---|---|---|
| *Rank* | *Variable* | *Score* | *Variable* | *Score* |
| **1** | mailback_area_count_cen_2010 | 100.0 | tot_population_cen_2010 | 100 |
| **2** | pop_5_17_cen_2010 | 94.9 | college_acs_10_14 | 96.7 |
| **3** | frst_frms_cen_2010 | 94.0 | **rel_child_under_6_cen_2010** | **95.6** |
| **4** | med_hhd_inc_bg_acs_10_14 | 93.2 | one_health_ins_acs_10_14 | 94.7 |
| **5** | **pov_univ_acs_10_14** | **92.0** | nonfamily_hhd_cen_2010 | 94.2 |
| **6** | **aggr_house_value_acs_10_14** | **91.1** | avg_tot_prns_in_hhd_cen_2010 | 93.8 |
| **7** | pct_vacant_units_cen_2010 | 90.6 | **tot_prns_in_hhd_acs_10_14** | **90.2** |
| **8** | mrdcple_fmly_hhd_cen_2010 | 90.1 | pop_1yr_over_acs_10_14 | 89.9 |
| **9** | mrdcple_fmly_hhd_acs_10_14 | 90.0 | pop_5yrs_over_acs_10_14 | 89.2 |
| **10** | tot_prns_in_hhd_cen_2010 | 89.8 | females_cen_2010 | 87.9 |
| **11** | pct_rel_family_hhd_cen_2010 | 88.3 | hhd_ppl_und_18_cen_2010 | 86.9 |
| **12** | males_acs_10_14 | 87.4 | hhd_ppl_und_18_acs_10_14 | 86.0 |
| **13** | **aggregate_hh_inc_acs_10_14** | **86.6** | **pct_nonfamily_hhd_cen_2010** | **85.5** |
| **14** | **tot_prns_in_hhd_acs_10_14** | **86.2** | **aggregate_hh_inc_acs_10_14** | **83.5** |
| **15** | pct_pop_5_17_cen_2010 | 86.1 | valid_mailback_count_cen_2010 | 83.3 |
| **16** | pct_tot_occp_units_acs_10_14 | 85.4 | **pov_univ_acs_10_14** | **83.0** |
| **17** | tot_population_acs_10_14 | 84.2 | **aggr_house_value_acs_10_14** | **82.8** |
| **18** | pct_nh_white_alone_cen_2010 | 84.2 | pct_othr_lang_acs_10_14 | 82.4 |
| **19** | tot_housing_units_acs_10_14 | 83.8 | pop_under_5_cen_2010 | 82.3 |
| **20** | pct_hhd_ppl_und_18_cen_2010 | 83.8 | rel_family_hhd_cen_2010 | 82.3 |
| **21** | **rel_child_under_6_cen_2010** | **83.4** | nh_asian_alone_cen_2010 | 82.3 |
| **22** | **pct_nonfamily_hhd_cen_2010** | **83.1** | tot_housing_units_cen_2010 | 82.0 |
| **23** | nh_white_alone_acs_10_14 | 82.9 | females_acs_10_14 | 81.7 |
| **24** | **tot_occp_units_acs_10_14** | **82.7** | pct_rel_under_6_cen_2010 | 80.4 |
| **25** | census_mail_returns_cen_2010 | 82.6 | **tot_occp_units_acs_10_14** | **80.2** |

*Note:* The common variables are highlighted in bold.

## 2.2 Household Response Status

Because of the poor model performance on response rate at the block group level, we switched the outcome to household response status to see if the performance can be improved.

*2.2.1 Simple logistic regression with LRS*

We first ran two simple logistic regression with the LRS as the single predictor to the household response. Table 4 presents the estimate of the LRS and the pseudo R-squared values for the two studies. The estimates were statistically significant at the 0.05 level, however the R-squared values were all below 0.1. The results echoed the previous analyses that the LRS is not a useful predictor of response rate for either of the two studies.

**Table 4:** Model estimation and pseudo R-square for logistic regression.

| | | *Study X* | *Study Y* |
|---|---|---|---|
| Estimate of LRS | | 0.020 (0.001)* | 0.015 (0.001)* |
| Pseudo R-squared | McFadden | 0.009 | 0.014 |
| | Cox-Snell | 0.041 | 0.061 |
| | Effron | 0.018 | 0.021 |

* p < .05

*2.2.2 Extreme gradient boosting with PDB and MSG*

When moving to extreme gradient boosting model on household response status, we focused on Study Y only. Switching from random forest to extreme gradient boosting model is due to the missing values in the MSG values. Random forest doesn't allow missing values. We ran the two tree boosting models with and without the MSG variables using the "xgbTree" option in the package "caret" in R. We used Receiver Operating Characteristic (ROC) to determine the optimal model using the one SE rule.

The household response status is binary—respondent or non-respondent—and the model outputs the predicted probabilities. Instead of using the default 0.5 to dichotomize the response status, we selected the cutoff value that equalizes the false-negative and false-positive errors. Using this criterion, the accuracy (the percent of correct prediction) was 57.7 with the PDB variables only, barely increasing to 58.0 after adding the MSG variables. The confusion matrix from the two models is presented in Table 5 below. The results indicated that the MSG variable added limited value to the prediction.

**Table 5:** Confusion matrix of models with PDB and MSG variables

| | *Observed Response Status* | | | |
|---|---|---|---|---|
| | *PDB only* | | *PDB and MSG* | |
| | *(Optimal Cutoff: 0.5442)* | | *(Optimal Cutoff: 0.5450)* | |
| *Predicted* | *NR* | *R* | *NR* | *R* |
| NR | 23.0 | 21.2 | 22.9 | 21.0 |
| R | 21.2 | 34.7 | 21.0 | 35.1 |

Although the MSG variables did not improve the model performance much, five MSG variables entered the top 20 list once they were in the model. Among them, income ranked second, tenure status ranked fourth, and age group of head of household ranked fifth. Together with the PDB variables remaining in the top 10, we can see that the measures on social-economic are more informative than the rest of variables in predicting household level response.

**Table 6.** The top 20 most important PDB and MSG variables in Study Y

| | *PDB only* | | *PDB + MSG* | |
|---|---|---|---|---|
| *Rank* | *Variable* | *Score* | *Variable* | *Score* |
| **1** | avg_agg_hh_inc_acs_10_14 | 100.0 | avg_agg_hh_inc_acs_10_14 | 100.0 |
| **2** | college_acs_10_14 | 49.9 | ***income*** | ***71.8*** |
| **3** | med_hhd_inc_tr_acs_10_14 | 42.0 | college_acs_10_14 | 68.0 |
| **4** | pct_nh_asian_alone_cen_2010 | 41.0 | ***ownrentR*** | ***65.8*** |
| **5** | pct_rel_under_6_cen_2010 | 39.3 | ***age_of_hoh*** | ***54.2*** |
| **6** | pct_nh_aian_alone_cen_2010 | 39.0 | med_hhd_inc_tr_acs_10_14 | 48.2 |
| **7** | med_house_value_bg_acs_10_14 | 31.3 | pct_nh_asian_alone_cen_2010 | 43.1 |
| **8** | pct_hispanic_cen_2010 | 30.9 | aggregate_hh_inc_acs_10_14 | 36.8 |
| **9** | urbanized_area_pop_cen_2010 | 28.3 | med_house_value_tr_acs_10_14 | 36.4 |
| **10** | med_house_value_tr_acs_10_14 | 27.3 | pct_mlt_u10p_acs_10_14 | 29.4 |
| **11** | pct_pop_under_5_cen_2010 | 25.3 | pct_rel_under_6_cen_2010 | 29.4 |
| **12** | single_unit_acs_10_14 | 25.2 | pct_single_unit_acs_10_14 | 27.1 |
| **13** | aggregate_hh_inc_acs_10_14 | **22.1** | pct_pop_25yrs_over_acs_10_14 | 26.4 |
| **14** | nh_asian_alone_cen_2010 | 22.1 | pct_urbanized_area_pop_cen_2010 | 26.4 |
| **15** | pct_urbanized_area_pop_cen_2010 | 21.2 | ***numberofchildren*** | ***23.9*** |
| **16** | pct_nh_blk_alone_cen_2010 | 19.2 | ***ethnicityHispanic*** | ***23.4*** |
| **17** | pct_single_unit_acs_10_14 | 19.1 | pct_nh_aian_alone_cen_2010 | 21.9 |
| **18** | pct_college_acs_10_14 | 18.8 | med_house_value_bg_acs_10_14 | 20.9 |
| **19** | pct_pop_18_24_cen_2010 | 18.0 | mail_return_rate_cen_2010 | 19.8 |
| **20** | aggr_house_value_acs_10_14 | 17.6 | pct_nh_white_alone_acs_10_14 | 19.4 |

*Note:* The MSG variables are in *Italic*.

## 3. Prediction During Data Collection

Prior to data collection, the best accuracy of prediction using the PDB and MSG variables was less than 60%. This prediction was better than flipping a coin, but it was barely informative for survey planning and resource allocation. Once in the field, we can add paradata information into the model, which will hopefully improve prediction. Paradata is updated daily, so we were able to extract the information for a specific period. As an example, we used the paradata from the first 15 days and 30 days in the field. We derived 62 variables from paradata for the two periods, including the total number of contacts, the number of contacts occurring during a specific time, and interim status, and added these variables to the tree-boosting model together with the PDB and MSG variables.

Once the paradata variables entered the model, the prediction accuracy improved from less than 60% to 74.6% with only 15 days of paradata information. Adding two more weeks, up to one month, the accuracy reached to almost 80% (see confusion matrix in Table 7). Table 8 lists the top 10 most important variables in the two tree-boosting models. Among them, the top five are the same and play a stronger role in prediction in terms of importance score. The labels of the top five variables are presented below:

- n_refusal1: Have one interim refusal
- p_call_back: Percent of times interim status requesting re-contact
- p_unknown: Percent of times unable to contact
- n_refusal2+: Have two or more interim refusals
- p_not_applicable: Percent of times interim status not applicable

N_refusal1 and n_refusal2+ are two dummy-coded variables for the categorical number of interim refusal. Both variables have high importance in status prediction. This implies that whether or not we had an interim refusal is a strong indicator of the final response status.

**Table 7:** Confusion matrix of models with PDB, MSG and paradata variables

| | *Observed Response Status* | | | |
|---|---|---|---|---|
| | *15 days in the field (Optimal Cutoff: 0.4638)* | | *30 days in the field (Optimal Cutoff: 0.4417)* | |
| *Predicted* | *NR* | *R* | *NR* | *R* |
| NR | 31.5 | 12.9 | 33.7 | 10.0 |
| R | 12.5 | 43.1 | 10.6 | 45.7 |

**Table 8:** The top 10 most important variables in Study Y with paradata in the model

| | *15 days in the field* | | *30 days in the field* | |
|---|---|---|---|---|
| *Rank* | *Variable* | *Score* | *Variable* | *Score* |
| **1** | n_refusal1 | 100.0 | n_refusal1 | 100.0 |
| **2** | p_call_back | 58.2 | p_unknown | 55.4 |
| **3** | p_unknown | 57.3 | p_call_back | 51.9 |
| **4** | n_refusal2+ | 25.5 | n_refusal2+ | 44.7 |
| **5** | p_not_applicable | 15.5 | p_not_applicable | 14.8 |
| **6** | n_call_back | 13.7 | n_unknown | 8.3 |
| **7** | n_unknown | 9.9 | timeinfield | 7.4 |
| **8** | p_appointment | 7.2 | n_call_back | 7.4 |
| **9** | nbrokenappt | 5.0 | p_appointment | 6.8 |
| **10** | n_not_applicable | 4.7 | nbrokenappt | 4.7 |

## 4. Summary

Westat investigated the use of the Census Bureau Planning Database containing numerous block group level statistics as well as a composite Low Response Score. Westat also investigated commercially available data on households and paradata obtained during the field period as potential predictors of nonresponse in two household surveys fielded by Westat using an ABS design.

We found two surprising results from this work. The first is that the PDB and the LRS have little value, prior to data collection, in predicting the final response rates. The correlation of response rates with the LRS was less than 0.15 at the block group at the national level. The simple logistic models were fit to the response rates using the LRS, but the pseudo R-squared was less than 0.1. In fitting the observed response rates using the PDB and the LRS as covariates, the LRS was neither retained in the final models through LASSO, nor did it enter in the top 25 most important covariates in random forest model in either study.

The second surprise came when incorporating paradata in the extreme gradient boosting models for predicting household response status. For survey Y, a complete time series of paradata and eligible responses over time were available. In building models to predict response based on both observed paradata, MSG data and PDB variables, MSG and PDB variables together yielded a model accuracy of less than 60%. Once partial paradata becomes available, the prediction of response rates improved to 75%.

We conclude that for the surveys that Westat investigated the information in PDB and MSG showed little promise in predicting future/unobserved response rates; whereas paradata did provide reasonable estimates and tree-based models are the most effective tool that we have tested. Conjectures as to why the LRS and PDB have little predictive power for the two Westat studies center on the differences in methodology between Census mail-out surveys and in-person surveys. Westat has recently conducted another survey with a mail-out screener, and this new survey could be used to assess whether the methodology difference explains the low predictive power from the PDB variables. On the other hand, the current study didn’t answer the question of whether final response rate is more a function of “effort” on the part of the field other than the characteristics of geographic areas. For future research, we will explore whether it is possible to design a study to estimate “innate response propensity” before “effort” is applied.

## References

Erdman, C., and Bates, N. (2017), “The Low Response Score (LRS): A Metric To Locate, Predict, and Manage Hard-to-survey Populations.” *Public Opinion Quarterly*, 81(1): 144–156.

Brick J. M., Lohr S., Edwards W. S., Giambo P., Broene P., Williams D., and Dipko S. (2013), “National Survey of Crime Victimization Companion Study-Summary of Pilot Results,” prepared for U.S. Bureau of Justice Statistics. Rockville, MD: Westat.

Montaquila J. M. (2014), “Use of Vendor Data in Optimization of Address-Based Sampling Procedures: Discussion,” paper presented at the 2014 Joint Statistical Meetings, Boston, MA.

Roth S. B., Han D., and Montaquila J.M. (2013), “The ABS Frame: Quality and Considerations.” *Survey Practice*, 6(4).