RESEARCH REPORT SERIES
(Statistics #2014-08)

The U.S. Census Bureau Mail Return Rate Challenge: Crowdsourcing to Develop a Hard-to-Count Score

Chandra Erdman
Nancy Bates

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 12, 2014

The U.S. Census Bureau Mail Return Rate Challenge: Crowdsourcing to Develop a Hard-to-Count Score

Chandra Erdman* **and Nancy Bates**†

U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233

June 17, 2014

Abstract

In 2012, the U.S. Census Bureau posed a challenge under the America COMPETES Act, an act designed to improve the competitiveness of the United States by investing in innovation through research and development. The Census Bureau contracted Kaggle.com to host and manage a world-wide competition to develop the best statistical model to predict 2010 Census mail return rates. The Census Bureau provided competitors with a block-group-level database consisting of housing, demographic, and socio-economic variables derived from the 2010 Census, five-year American Community Survey estimates, and 2010 Census operational data. The Census Bureau then challenged teams to use these data (and other publicly available data) to construct the models. One goal of the challenge was to leverage winning models as inputs to a new model-based hard-to-count (HTC) score, a metric to stratify and target geographic areas according to propensity to self-respond in sample surveys and censuses.

All contest winners employed data mining and machine learning techniques to predict mail return rates using the database supplied by the Census Bureau along with data from external sources. This made the models relatively hard to interpret (when compared with the Census Bureau's original HTC score) and impossible to directly translate to a new HTC score. Nonetheless, the winning models contained insights toward building a new model-based score using variables from the database. This paper describes the original algorithm-based HTC score, insights gained from the Census Return Rate Challenge, and the model underlying a new, block-group-level low response score.

*Center for Statistical Research and Methodology; Email: chandra.erdman@census.gov
†Research and Methodology Directorate; Email: nancy.a.bates@census.gov

Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the author and not necessarily those of the U.S. Census Bureau.

Contents

1 The U.S. Census Return Rate Challenge

"All you need is data and a question. Our data scientists will provide the answer."

– Kaggle.com

On August 31, 2012 the U.S. Census Bureau launched a nationwide prize competition under the America COMPETES Act. The contest – dubbed the Census Return Rate Challenge, or Census Challenge – encouraged individuals and teams to compete for prize money[1] in predicting 2010 Census mail return rates. Kaggle.com, a company that hosts predictive modeling competitions, was contracted to manage and judge the competition. The objective of the contest was to create a statistical model to accurately predict 2010 Census mail return rates for small geographic areas (namely, census block-groups). Nationwide, 79.3 percent of households that received a 2010 Census mail questionnaire completed it and mailed it back. However, the level of mail return varied greatly by geography. The Census Return Rate Challenge asked participants to model these variations using variables found in the 2012 Census Planning Database.

[1] $25,000 total prize money was awarded under the Challenge: $14,000, $7,500 and $2,500 respectively for 1^{st}, 2^{nd} and 3^{rd} place model winners and $1,000 for the data visualization winner.

The first planning database (PDB) was compiled from 1990 Census data and contained tract-level characteristics of households and people that are associated with response rates (Bruce and Robinson, 2003, 2007). The 1990 PDB was used in the planning, implementation and evaluation of the 2000 Census and was eventually updated with Census 2000 data. The 'Tract-Level Planning Database with Census 2000 Data' has been used in a number of initiatives including creating a "hard-to-count" (HTC) score which identifies areas that are difficult to enumerate (Bruce et al., 2001), segmenting the nation for the 2010 Census advertising campaign (Bates and Mulry, 2011), and developing field interviewer performance strata (Erdman et al., forthcoming).

In 2012, the U.S. Census Bureau developed the first block-group-level planning database, compiled from the 2006-2010 American Community Survey (ACS) and the 2010 Census, as well as several operational variables that describe 2010 Census mail-back behavior. The U.S. Census Bureau provided the 2012 PDB to Census Challenge contestants to build response rate models. A portion of the database was withheld for validation, to confirm models with the lowest weighted mean squared error (MSE). It was also used to update the daily leader-board that tracked each team's standings in the competition. The Census Return Rate Challenge proved very popular – 244 teams and individuals entered the contest.

The practice of soliciting statistical models from the public rather than from internal employees was new to the U.S. Census Bureau. The external competition was initiated by then Census Bureau Director, Robert Groves, and had multiple goals. First, it would encourage creative thinking without the bias of prior approaches to predicting census mail returns. Second, it would attract additional high caliber data scientists, broadening the number and scope of analysts examining the problem. Third, it would position the U.S. Census Bureau as an innovator and potentially improve efficiencies by using crowdsourcing solutions to technical problems. Finally, the Census Bureau hoped to draw upon the winning model as a means to produce an updated block-group-level hard-to-count score, a metric to stratify and target areas according to propensity to self respond to decennial censuses and sample surveys. The intent was to develop a new score that is replicable, easy to interpret and use in the field, and consistent across various levels of geography, in particular, census block-groups and tracts. In this paper, we discuss the winning models and previous method for developing an HTC score. We document the similarities and differences between the two approaches in search of a new, model-based score.

2 The Original HTC Score

Bruce et al. (2001) used the 'Tract-Level Planning Database with Census 2000 Data' to develop a summary indicator for identifying areas that are difficult to enumerate – the hard-to-count score. Based on research regarding barriers to enumeration, Bruce et al. (2001) selected twelve variables for inclusion in the summary score. The variables were guided by ethnographic research designed to identify reasons why people are missed in censuses (de la Puente, 1995). They included both housing variables (e.g., the percentage of vacant houses, percentage of housing units without a phone, and percentage of multi-unit structures) and socio-demographic/economic indicators (e.g., the percentage of people below poverty, percentage of linguistically isolated households, and percentage of renter households). With these variables, the HTC score was created as outlined in the following three-step algorithm:

1. Sort each variable individually from high to low, across tracts. (Each variable correlated with nonresponse and therefore high values indicate the potential for difficulty in enumeration.)

2. Assign integer scores from 0 to 11 to the values of each variable based on the percentiles as follows:

Percentile	50.0	55.0	60.0	65.0	70.0	75.0	80.0	85.0	90.0	95.0	97.5	100
Score	0	1	2	3	4	5	6	7	8	9	10	11

3. Assign each tract an HTC score equal to the sum of the scores over the twelve variables.

Since its creation, the HTC score has been used not only in planning for the 2010 Census but also in managing daily operations of the many national surveys conducted by the U.S. Census Bureau. Prior to Census 2010, HTC scores were updated each decade and appended to the publicly available PDB. In turn, local communities used the scores to identify HTC areas in their jurisdictions and then tailor Census outreach activities to those populations.

Using the same score over time, Bruce et al. (2012) illustrate mail return rates by HTC score deciles, noting useful trends over the 1990, 2000 and 2010 Censuses (see Figure 1). This makes clear the strong correlation between the HTC score and self response.

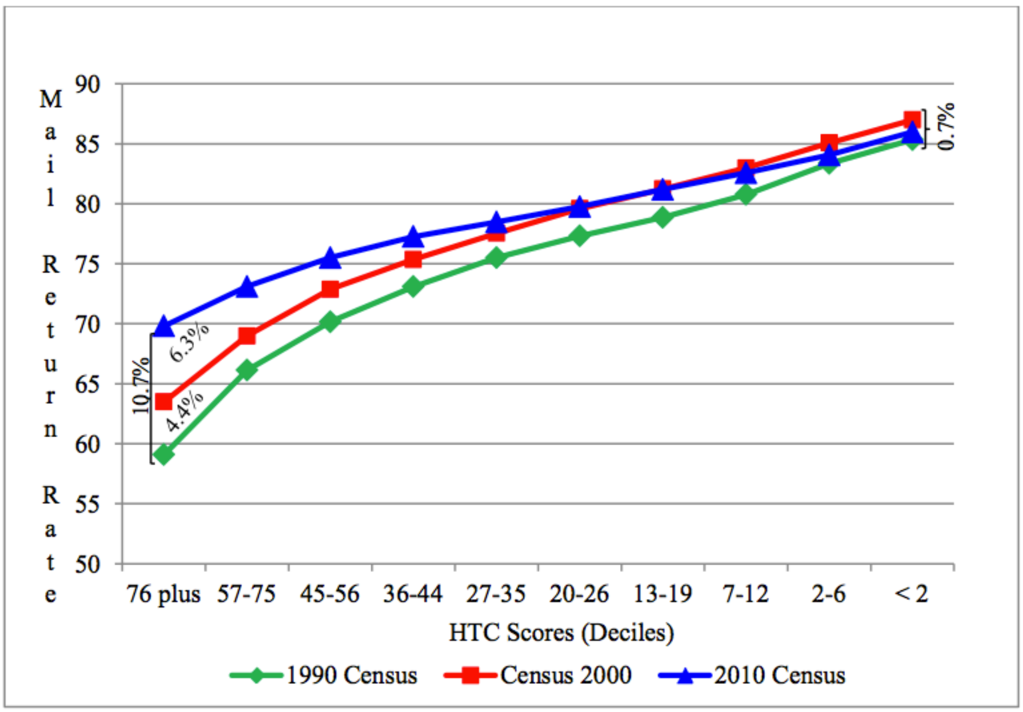

Figure 1: Trends in Census 1990, 2000 and 2010 Census Mail Return Rates by Hard-to-Count Scores. Source: Bruce et al. (2012)

2.1 The Winning Challenge Models

At the conclusion of the Census Return Rate Challenge, Bill Bame, a software developer from Maryland, was awarded the top monetary prize for his predictions that yielded a mean squared error of 2.60.

An examination of the top three Challenge models revealed some commonalities. All top contestants used ensemble methods (gradient boosting or random forests) that fall under the heading of Machine Learning. These are the same methods used in a large number of statistical modeling competitions including the million-dollar Netflix Prize competition (Koren, 2009). These methods generate a multitude of alternative models for prediction or classification of a given data set, fitting model after model in an effort to minimize a loss function (in this case, population-weighted mean squared error). This results in an ensemble of weakly-predictive models that together yield highly-accurate predictions.

Winning contestants also used hundreds of predictors, including variables not supplied by the Census Bureau on the 2012 PDB. The use of external data meant that the winning model could not be directly applied as a predictive model for the new HTC score. However, examination of the winning model predictors in rank order of relative influence proved enlightening and, in fact, confirmed previous research around census mail return behavior. For example, the single most influential predictor in Bame's model was the percentage of renter households in a block-group. Previous research has noted wide variation in census participation between homeowners and renters as far back as the 1990 Census. Word (1997) documents that renters are much less likely to mail back a census questionnaire than homeowners. The percentage of renters in a geographic area is also one of the twelve variables used in the original HTC score. However, other variables ranking high in Bame's model include 2000 Census response rates, nearest neighboring block-group response rates, and margins of error for various ACS estimates. While these are good predictors, they do not provide insight into the reasons for high or low response rates. Nonetheless, the winning model's variables, ranked by relative influence, serve as a starting point for building a new model-based hard-to-count score.

3 Developing an Updated HTC Score

Recall that our goal is to use the results of the Census Return Rate Challenge to create an updated HTC score that is replicable, easy to interpret and apply in the field, and consistent across census block-groups and tracts. Because we want an HTC score that is easy to interpret, we examine individual ordinary least squares (OLS) regression models in lieu of the ensembles of regression trees that won the Challenge. And, because we want model predictors to be actionable (like age and presence of children), we restrict our attention to predictors that fit this criterion.

We begin by examining the relative influence of the predictors in the winning Challenge model. Figure 2 displays this statistic for the top fifty most influential predictors. As noted earlier and displayed in the figure, the most influential variable is the percentage of renters in a block-group. This variable is followed by the percentage of people ages 18 to 24 and the percentage of households headed by unmarried females. The key features of this plot are that relative influence decreases sharply over the first several variables, there are smaller drops over the next twenty variables and, beyond the this point, the relative influence of the predictors is small.

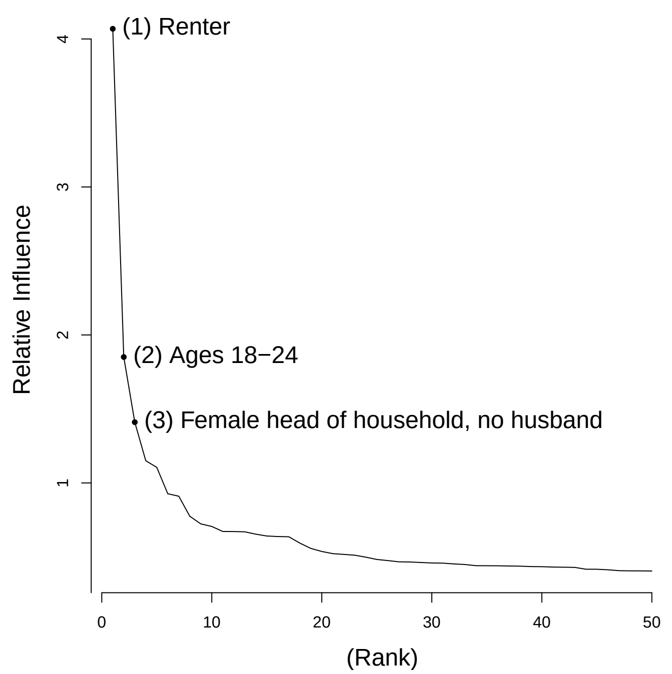

Figure 2: Relative Influence of the Top 50 Most Influential Variables in the Winning Challenge Model

We examined models from the Census Challenge versus a model built upon the original HTC score. Our criteria for determining the "best" model included: (1) a relatively small number of independent variables, (2) high predictive value, measured by adjusted R-squared and population-weighted mean squared error, and (3) because the original HTC score was based on census tracts as opposed to block-groups, accurate predictions at both levels of geography. To estimate how accurately these models will predict in practice, we performed 100 rounds of 2-fold cross-validation.

Table 1 displays the cross-validated R-squared and MSE values of the OLS regression models, averaged over the 100 rounds of testing, with standard errors given in the superscripts. The Bame (2012) model represents the top 25 "actionable" rank order predictors from the winning Census Challenge Model. The Bruce et al. (2001) row represents a model containing only the original twelve HTC score variables. The Bruce et al. (2001) model has fewer predictors (twelve versus twenty-five) and correspondingly lower R-squared and higher MSE values. Both models predicted similarly well between tracts and block-groups. Since the Bame model had higher predictive power, lower error, and a more robust set of variables to delineate HTC areas, we pursued this model to develop a new HTC metric.

Table 1: Comparison of Model Fit Statistics Across Studies and Geographies

	Block-group		**Tract**	
Model	R-squared	MSE	R-squared	MSE
Top 25 Bame (2012)	$56.10^{0.0017}$	$30.48^{0.0013}$	$55.01^{0.0117}$	$23.01^{0.0033}$
Bruce et al. (2001)	$45.47^{0.0007}$	$38.08^{0.0005}$	$45.29^{0.0038}$	$28.34^{0.0013}$

Table 2 displays coefficients of the "Top 25 Bame (2012)" model[2], at the block-group and tract levels. Because we want to predict areas that are hard to enumerate, we use 100 minus return rate – "non-return rate" – as the dependent variable. Most predictors are highly significant at both levels of geography and all are significant at at least one level of geography. Seven of the twelve variables from the original HTC score are found in the model. Other variables include length of residence, presence of young children, and married couple households – variables that describe the "place attachment" construct found in urban sociology literature (Brown et al., 2003). This construct figures into theories explaining behaviors such as level of civic engagement, voting, and even participation in surveys (Guterbock et al., 2006). From Figure 2, we see that (given all other covariates) the presence of renters, vacant units, and persons aged 18-24 in a block group are all positively associated with low response. That is, the *greater* the presence of these characteristics in a block group, the *lower* the self response rates. Alternatively, presence of persons aged 65+, married couples, related children under 6, and persons with a college degree are negatively associated with low response.

While most variables are self-explanatory, a couple require clarification. The variable "different housing unit 1 year ago" is at a person level and is defined as the percentage of people who moved from another residence in the U.S. or Puerto Rico within the last year. The "moved in 2005-2009" variable is at a household level and is defined as the percentage of households where the householder moved into the current unit between 2005 and 2009. The fact that both of these variables measure aspects of mobility but the coefficients have differing signs may be puzzling to some readers. While it is possible that the person-level mobility variable has a positive relationship with mail return rates and household- (householder-) level mobility has a negative relationship with mail return rates, we highlight the fact that no interactions are included in the models and this (in addition to the relationship between mobility and the other covariates) may account for potentially counterintuitive signs.

With this model, we propose a new, Census Challenge-based score – each geography's score is simply the fitted value from the OLS regression. We refer to this new metric as a low response score or LRS. To illustrate this new model-based score, Figure 3 displays deciles of the fitted and actual

[2] In the end, we removed the variable "Ages < 5" as all other age groups are accounted for in the model, and this group is very closely related to the more influential variable, "Households with related child < 6."

non-return rates for block-groups in the District of Columbia. With few exceptions, hard-to-count areas have the highest fitted values (shaded in red and dark orange), and the easiest to enumerate areas have the lowest fitted mail non-return rates (shown in shades of yellow).

Table 2: Low Response Model Summaries at the Block-Group and Tract Levels

	Block-group			**Tract**		
Variable	Coef.	Z-value	Sig.	Coef.	Z-value	Sig.
Intercept	10.29	12.49	***	16.61	10.56	***
Renter occupied units	1.08	50.57	***	0.95	23.51	***
Ages 18-24	0.64	21.53	***	0.47	9.57	***
Female head, no husband	0.58	17.26	***	0.33	5.37	***
Non-Hispanic White	-0.77	-38.76	***	-0.87	-26.17	***
Ages 65+	-1.21	-39.61	***	-1.29	-24.31	***
Related child <6	0.46	15.82	***	0.08	1.38	
Males	0.09	20.43	***	0.04	4.91	***
Married family households	-0.12	-37.43	***	-0.14	-25.46	***
Ages 25-44	-0.06	-1.74		0.11	2.08	*
Vacant units	1.08	52.74	***	0.91	25.11	***
College graduates	-0.32	-17.33	***	-0.53	-12.62	***
Median household income	0.24	4.62	***	0.34	2.88	**
Ages 45-64	-0.08	-2.54	*	-0.16	-2.69	**
Persons per household	3.44	13.19	***	3.30	6.78	***
Moved in 2005-2009	0.09	7.19	***	0.13	4.38	***
Hispanic	0.41	24.45	***	0.52	18.23	***
Single unit structures	-0.52	-53.11	***	-0.56	-27.32	***
Population Density	-0.40	-41.93	***	-0.46	-29.25	***
Below poverty	0.11	9.95	***	0.26	9.56	***
Different housing unit 1 year ago	-0.12	-11.09	***	-0.35	-12.55	***
Ages 5-17	0.17	4.30	***	0.24	3.24	**
Black	-0.04	-2.69	**	0.01	0.24	
Single person households	-0.24	-5.19	***	-0.35	-4.26	***
Not high school graduate	-0.06	-4.84	***	-0.19	-6.75	***
Median house value	0.71	25.56	***	0.78	14.69	***

Sig.: *** $p < .001$; ** $.001 \leq p < .01$; * $.01 \leq p < .05$

Block-group R-squared: 56.1, n = 217,417; Tract R-squared: 55.25, n = 72,763.

Note: All variables are percentages unless otherwise indicated. Most variables are square-root, log or logit transformed (see the Appendix).

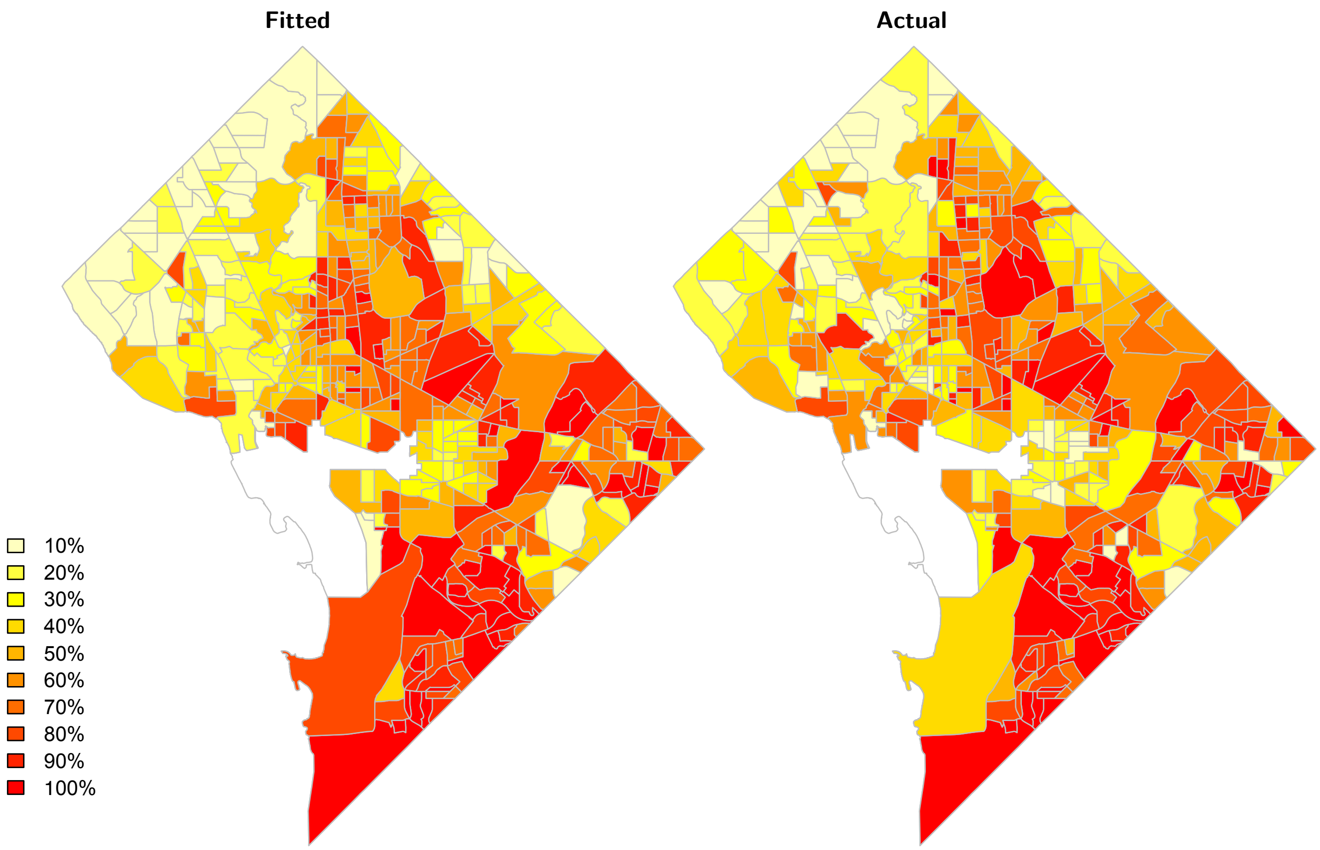

Figure 3: Deciles of Fitted and Actual Non-Return Rates for Block-Groups in the District of Columbia

4 Discussion

In this paper, we apply results from the U.S. Census Bureau's Response Rate Challenge, a crowd-sourcing project to produce the best predictive model of 2010 Census mail return rates. An overall goal of the Challenge was to encourage new approaches to understanding self-response behavior in the Decennial Census. A secondary goal was to use the winning entry as a means for updating the Census Bureau's HTC score. The HTC score is a metric pioneered by the Census Bureau over two decades ago that delineates areas of the country according to ease or difficulty of enumeration.

The winning Challenge model had many predictor variables – 342 in all. After closer examination, we excluded many when constructing a new HTC model because they come from sources external to the 2012 Planning Database, have little meaning, or have low predictive power. Still, our final model containing only twenty-five variables was highly predictive of mail response at the block-group level ($R^2 = 56.10$).

Our model relies heavily on the highest rank order predictors from the winning Census Challenge model which includes a majority of the twelve variables used in the original HTC score. The variables that comprise our score are a robust set that can inform both social marketing and partnership campaigns used in the Decennial Census. Using the District of Columbia as an example, we see how areas with similarly high LRS have very different characteristics. This differentiated outcome is particularly useful when trying to plan a targeted communication campaign, deliver in-language forms, or develop community partnership materials.

Consider the block-groups in three DC neighborhoods displayed in Figure 4. The LRS is 32.5 for Columbia Heights, 37.4 for Trinidad, and 37.8 for Anacostia – all above the 90^{th} percentile. All three neighborhoods have high proportions of renters (82 percent or higher) but are otherwise quite dissimilar. The Columbia Heights block group is comprised of 45 percent Hispanic residents where 33 percent speak a language other than English. Seventy-four percent of the households are multi-unit structures, 52 percent are non-family households, and 50 percent of the householders moved in within the last 5 years. When conducting the census, this block group could benefit from in-language advertising and Spanish language forms. Alternatively, the Anacostia block group is comprised of 98 percent Black residents of whom 46 percent are below poverty and 89 percent are single-unit homes. This block group has a low percentage of non-family households (15 percent) and only 21 percent of householders moved in within the last 5 years. A successful technique for engaging this area might be community partnerships since it is characterized by families who have lived in the area for a longer time. Finally, the Trinidad neighborhood skews younger and mobile with well over one-third (37 percent) aged 18-24 and 59 percent of householders having moved to the address within 5 years. The area is also characterized by poverty and vacant units (one-third are below poverty and just over one-quarter of the houses are vacant). The racial/ethnic composition is varied with 55 percent Black householders and 31 percent White. These characteristics suggest a neighborhood in high transition with in-moving younger households along with the more traditionally hard-to-count.

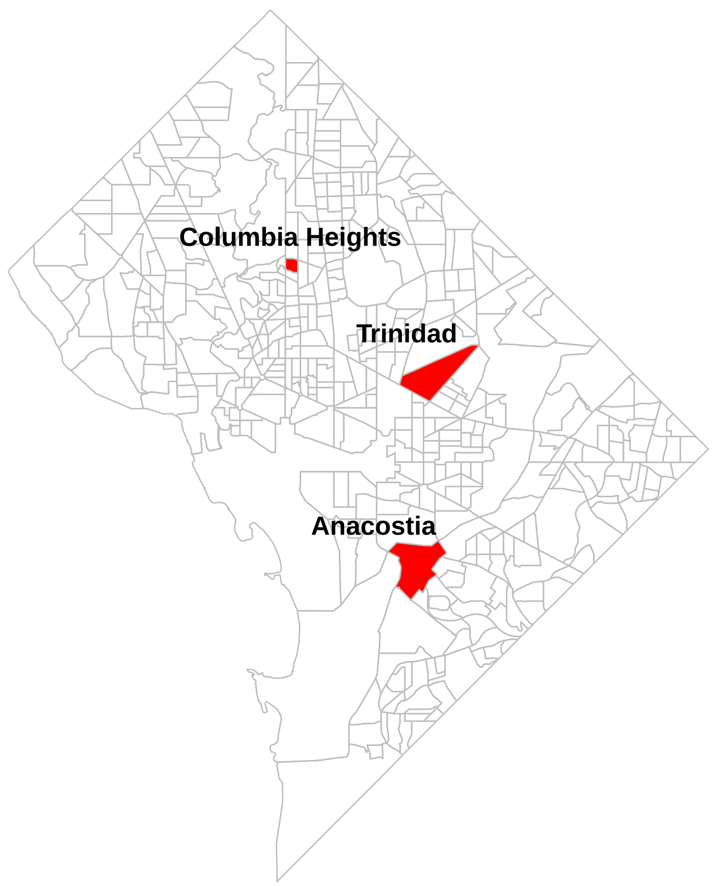

Figure 4: Three HTC Block-Groups in the District of Columbia

Users of the PDB and LRS are encouraged to perform similar exercises at the block-group or tract level for their own neighborhoods and communities. We suggest users not be constrained by the 25 variables that ended up in the final model. Instead, we recommend using the LRS to first pinpoint harder- to-count areas. Then, use the full set of variables found in the PDB to paint a picture of the areas. Finally, use the detailed profiles to develop tailored methods designed to boost cooperation.

While we believe the new LRS will be useful to Census and survey planners, we note an important caveat to the score going forward – namely that the response metric we are predicting is based on a single mode of self-response (mail). For the next Census in 2020, the Internet is expected to be the preferred mode of response with great effort made to promote and encourage its use. The survey literature suggests that, in mixed-mode approaches, different segments of the population adopt the Internet response mode to different degrees (Datta et al., 2002; Nicolaas et al., 2014; Link and Mokdad, 2006).

For example, in the 2009 ACS, a population segment coined by Bates and Mulry (2011) as the "Single Unattached Mobiles" had below average mail return rates (42.5 percent). This segment skews toward single young people living in urban multi-units who rent and move frequently. Beginning in January 2013, the ACS added Internet as means of self-response (in addition to mail). Recent ACS return rates from 2013 that reflect both mail and Internet mode indicate that this segment prefers Internet (63.3 percent of total self response was by Internet). Additionally, when both modes are offered, the overall self-response rate is actually higher for this group than four years prior (44.8 percent, see Baumgardner (2013)). For this segment of the population at least, this represents a relatively unheard of reversal to the downward trend in response rates.

Because our dependent variable does not account for Internet response, our score is not as accurate as it could be and some groups and geographical areas may be characterized as harder to count than they really are. However, in 2013, in addition to adding the Internet as a response option, the ACS also added a question about Internet access at home. When new ACS data become available at low levels of geography, we can revisit our LRS model to adjust both the dependent variable (to include additional modes of self-response) and predictor variables (to include degree of Internet penetration). In the meantime, the new LRS will be appended to the next iterations of the block-group and tract-level PDBs, scheduled for release in 2014, and updated each year thereafter.

References

William P. Bame. Winning Model Documentation. U.S. Census Return Rate Challenge, 2012.

Nancy Bates and Mary H. Mulry. Using a Geographic Segmentation to Understand, Predict, and Plan for Census and Survey Mail Nonresponse. *Journal of Official Statistics*, 27(4):601–918, 2011.

Stephanie Baumgardner. Self-Response Check-in Rate by Mode and Segmentation Group – June 2013. Email communication with the author, U.S. Bureau of the Census. September 17, 2013.

Barbara Brown, Douglas D. Perkins, and Graham Brown. Place Attachment in a Revitalizing Neighborhood: Individual and Block Levels of Analysis. *Journal of Environmental Psychology*, 23(2):259–271, 2003.

Antonio Bruce and J. Gregory Robinson. *The Planning Database: Its Development and Use as an Effective Tool in Census 2000.* Paper presented at the Annual Meetings of the Southern Demographic Association, Arlington, VA, October 24, 2003.

Antonio Bruce and J. Gregory Robinson. *Tract Level Planning Database with Census 2000 Data.* U.S. Government Printing Office, Washington, DC, 2007.

Antonio Bruce, J. Gregory Robinson, and Monique V. Sanders. Hard-to-Count Scores and Broad Demographic Groups Associated with Patterns of Response Rates in Census 2000. In *Proceedings of the Social Statistics Section*. American Statistical Association, 2001.

Antonio Bruce, J. Gregory Robinson, and Jason E. Devine. A Planning Database to Identify Areas That Are Hard-to-Enumerate and Hard-to-Survey in the United States. In *Proceedings of the International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations*. American Statistical Association, 2012.

Pratim Datta, Ken Walsh, and Dek Terrell. The Impact of Demographics on Choice of Survey Modes: Demographics Distinctiveness Between Web-Based and Telephone-Based Survey Respondents. *Communications of the Association for Information Systems*, 9(13), 2002. URL `http://www.websm.org/uploadi/editor/1135694840dattapais.pdf`.

Chandra Erdman, Tamara Adams, and Barbara C. O'Hare. Development of Interviewer Performance Standards for National Surveys. *Field Methods*, forthcoming.

Thomas M. Guterbock, Ryan A. Hubbard, and Laura M. Hoilan. Community attachment as a predictor of survey non-response. Unpublished Paper, 2006.

Yehuda Koren. *The BellKor Solution to the Netflix Grand Prize*, 2009. URL `http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf`.

Michael W. Link and Ali Mokdad. Can Web and Mail Survey Modes Improve Participation in an RDD-based National Health Surveillance? *Journal of Official Statistics*, 22(2):293–312, 2006.

Gerry Nicolaas, Lisa Calderwood, Peter Lynn, and Caroline Roberts. Web Surveys for the General Population: How, why and when? URL `http://www.natcenweb.co.uk/genpopweb/documents/GenPopWeb-final-report.p%df`. National Centre for Research Methods, NatCEN, 2014.

David L. Word. Who responds/who doesn't? Analyzing variation in mail response rates during the 1990 census. *U.S. Census Bureau Population Division Working Paper Series*, 1997.

A Technical Notes

Table 3: List of Transformations.

Variable	**Transformation**
Renter occupied units	Logit*
Ages 18-24	Square Root
Female head, no husband	Square Root
Non-Hispanic White	Logit
Ages 65+	Square Root
Related child <6	Square Root
Ages 25-44	Square Root
Vacant units	Log*
College graduates	Logit
Median household income	Log
Ages 45-64	Square Root
Persons per household	Log
Moved in 2005-2009	Square Root
Hispanic	Logit
Single unit structures	Logit
Population Density	Log
Below poverty	Square Root
Different housing unit 1 year ago	Square Root
Ages 5-17	Square Root
Black	Logit
Single person households	Square Root
Not high school graduate	Square Root
Median house value	Log
Public assistance	Square Root
Unemployed (ages 16+)	Logit
Crowded units	Square Root
Linguistically isolated households	Square Root
No phone service	Square Root

*In order to obtain finite values, a small amount was added to each variable before log or logit transformation. Specifically, the following functions were used:

```
Log <- function(x) {
    return(log(x + min(x[x>0], na.rm=T)))
}
Logit <- function(x) {
    temp <- (1+x)/102
    return(log(temp/(1-temp)))
}
```