

# 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

Federal Committee on Statistical Methodology  
Washington, DC  
November 5, 2013

**Disclaimer:** The views expressed on statistical issues are those of the authors only.

# The U.S. Census Bureau Return Rate Challenge

*“All you need is data and a question. Our data scientists will provide the answer.”*

*– Kaggle.com*

Our research question: Which statistical model best predicts 2010 Census mail return rates (block-group level)?

Our dataset: 2012 Census Planning Database (PDB)

Product: Updated model-based Hard-to-Count Score

# Census Crowdsourcing Challenge

- 2009 America COMPETES Act
- Contest ran August 31 - November 1, 2012
- 244 teams and individual competitors
- Unanticipated challenges:
  - Non US citizens
  - Use of auxiliary datasets

# Winning model and HTC score

- Software developer from Maryland awarded top monetary prize (MSE=2.60)
- Used random forests and gradient boosting
- Model included 342 variables – many from sources external to Census PDB

*How to apply Challenge results toward new model-based HTC score?*

# HTC-Related Studies

- Bruce et al. (2001); Bruce and Robinson (2003)
  - Original HTC score
- Guterbock et al. (2006)
  - Community attachment theory
- Erdman et al. (2013)
  - Interviewer performance stratification

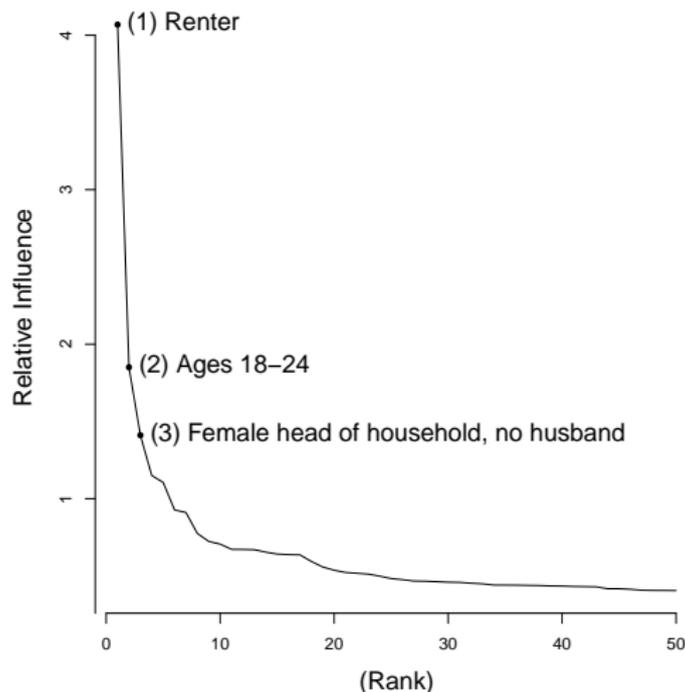
# Model Selection Criteria

- 1 Restrict to PDB predictors
- 2 Small number of predictor variables
- 3 High predictive value (adjusted  $R^2$ )
- 4 Low mean square error
- 5 Model works for both tracts and block-groups

Additional consideration: to include or exclude race/ethnicity composition as predictors?

# Winning Model Predictors

- 90 percent (308/342) of predictors from census
- When ranked by relative influence, 24/25 top predictors from census



# Comparison of Predictors Across Studies

**Table:** Overlap Between Top 25 Predictors in Bame (2012) and Erdman et al. (2013)

Predictor	Bame	Erdman	Bruce	Guterbock
Renter occupied units	✓	✓	✓	✓
Married family households	✓	✓	✓*	✓
Ages 65+	✓	✓	+	✓
Ages 18-24	✓	✓	-	✓
College graduates	✓	✓	-	✓
Moved in 2005-2009	✓	✓	-	✓
Ages < 5	✓	✓	+	✓
Ages 5-17	✓	✓	+	✓
Vacant units	✓	✓	✓	
Single unit structures	✓	✓	✓*	
Males	✓	✓	-	
Non-Hispanic White	✓	✓	+	
Persons per household	✓	✓	-	
Population Density	✓	✓	-	
Below poverty	✓	✓	✓	
Hispanic	✓	✓		
Non-Hispanic Black	✓	✓		

# Comparison of Predictors Across Studies (Cont.)

**Table:** Remaining Top 25 Predictors from Bame (2012)

<b>Predictor</b>	<b>Bame</b>	<b>Erdman</b>	<b>Bruce</b>	<b>Guterbock</b>
Not high school graduate	✓		✓	✓
Different housing unit 1 year ago	✓		✓	✓
Related child < 6	✓		—	✓
Ages 25-44	✓		—	✓
Median household income	✓		—	✓
Ages 45-64	✓		—	✓
Female head, no husband	✓		—	
Single person households	✓		—	

# Comparison of Predictors Across Studies (Cont.)

Table: Remaining Variables from Bruce et al. (2001)

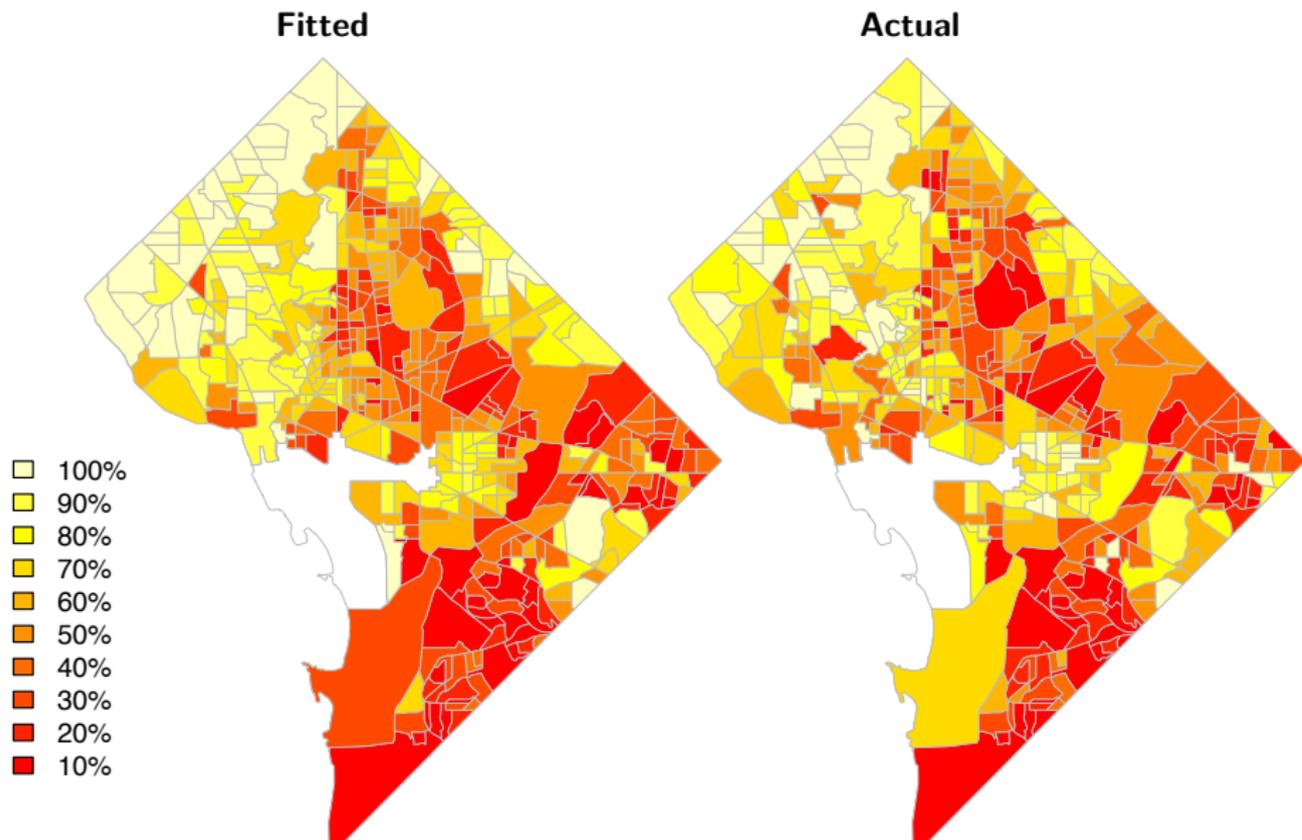
Predictor	Bame	Erdman	Bruce	Guterbock
Public assistance			✓	
Unemployed	—	—	✓	
Crowded units			✓	
Linguistically isolated households			✓	
No phone service		✓	✓	

# Model Fit Statistics

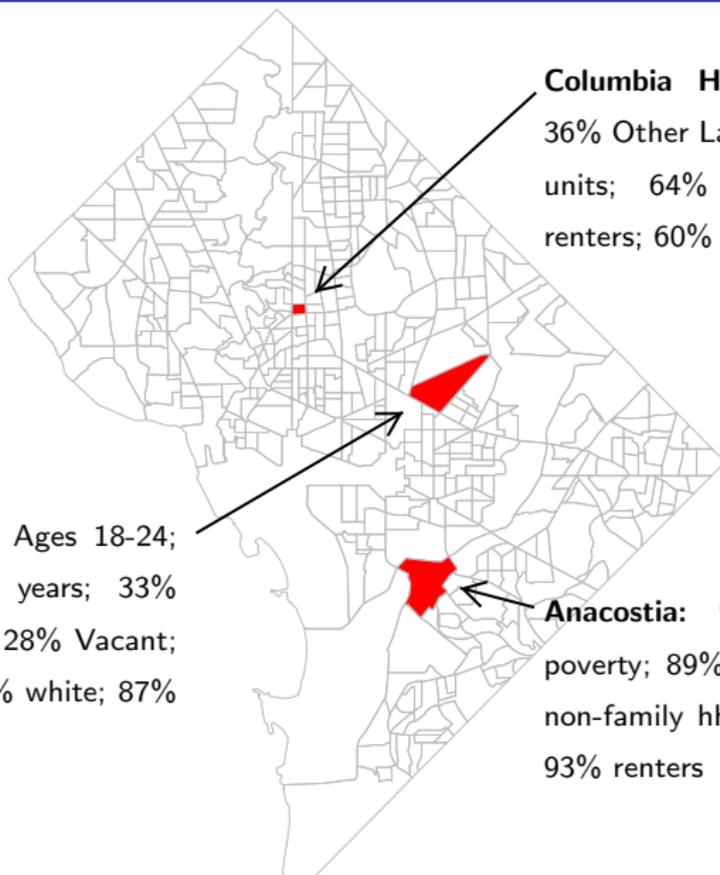
**Table:** Comparison of Model Fit Statistics Across Studies and Geographies

<b>Model</b>	<b>Block-group</b>		<b>Tract</b>	
	R-squared	MSE	R-squared	MSE
Top 25 Bame (2012) including race	56.27	30.36	55.03	23.09
Erdman et al. (2013) including race	56.18	30.37	55.33	22.84
Top 25 Bame (2012) excluding race	55.58	30.85	54.52	23.27
Erdman et al. (2013) excluding race	53.89	32.05	52.06	24.72
Guterbock et al. (2006)	51.17	33.86	49.83	25.75
Bruce et al. (2001)	45.66	37.99	45.78	28.21

# Deciles of Return Rates for Block-Groups in DC



# Three HTC Block-Groups in DC



**Columbia Heights:** 43% Hispanic; 36% Other Language; 92% 10+ multi-units; 64% non-family hhds; 85% renters; 60% moved 5 years

**Trinidad:** 37% Ages 18-24; 59% Moved 5 years; 33% Below poverty; 28% Vacant; 55% Black; 31% white; 87% renters

**Anacostia:** 98% Black; 46% below poverty; 89% single unit homes; 15% non-family hhds; 21% moved 5 years; 93% renters

# Considerations

- Independent variable is mail response; 2020 Census will have an Internet response option
- “Single Unattached Mobiles” (Bates and Mulry, 2011)
  - 64.7 percent of American Community Survey self response by Internet (Baumgardner, 2013)
- In January, 2013, ACS began asking about Internet connectivity

# Summary

- Challenge was successful
- Winning model was complex but predictors in rank order of influence proved useful
- Accurate predictions with relatively few predictors
- Simple HTC score: model fits
- First score at this level of geography
- Useful for planning and targeted advertising

# References



Nancy Bates and Mary H. Mulry.

Using geographic segmentation to understand, predict, and plan for census and survey mail nonresponse.

*Journal of Official Statistics*, 27(4):601–618, 2011.



S. 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.



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.



Chandra Erdman, Tamara Adams, and Barbara C. O'Hare.

Development of interviewer performance standards for national surveys.

Draft Paper, 2013.



Thomas M. Guterbock, Ryan A. Hubbard, and Laura M. Hoilan.

Community attachment as a predictor of survey non-response.

Unpublished Paper, 2006.

[chandra.erdman@census.gov](mailto:chandra.erdman@census.gov)  
[nancy.a.bates@census.gov](mailto:nancy.a.bates@census.gov)