Using Statistical Models in Place of Clerical Matching in the Census 2020 Post-Enumeration Survey to Produce Estimates of Census Housing Unit Coverage

Prepared for 2019 Joint Statistical Meetings

Michael Beaghen
Mark Jost
Elizabeth Marra
U.S. CENSUS BUREAU

Any views expressed on statistical issues are those of the authors and not those of the U.S. Census Bureau.

Outline

- Research Question
- Background on Post Enumeration Survey
- Methodology
- Limitations
- Results
- Conclusions and Future Research

Research Question

For estimating census housing unit coverage with the 2020 PES, can we replace the clerical matching with logistic regression modeling?

This would save time and money.

Background - The Post-Enumeration Survey

- Census 2020 Post-Enumeration Survey (PES) will evaluate the coverage of the 2020 Census
- The 2010 PES was called the Census Coverage Measurement survey

2010 Post Enumeration Survey Independent Listing

- The 2010 PES selected a sample of census blocks
  - Conducted independent listing of housing units in these sample blocks
  - Resulted in about 170,000 independently listed housing units

Dual System Estimation

- Capture-recapture
- The two systems are
  - the independent listing of PES housing units
  - the correct census enumerations of housing units
    - PES uses clerical review and fieldwork to establish correct census enumerations
- Dual system estimate is an estimate of true population of housing units

Dual System Estimation

	Correctly Enumerated in Census	Not Enumerated in Census
Housing Unit in PES	Match (in Census & PES)	Nonmatch (PES only)
Housing Unit Not in PES	Census only	Missed by both

Dual System Estimation

$$DSE = \frac{correct\ census\ enumerations}{match\ rate}$$

$$match\ rate = \frac{matches}{matches + nonmatches}$$

Match Status for 2010 PES Housing Units by Operational Stage (in Percent)

Status	Computer Matching Only	With Clerical Match
Match	72.9	93.5
Possible Match	11.4	N/A
Nonmatch	15.4	3.7
Duplicate	0.3	0.1
Not a Valid Housing Unit	N/A	2.8

(U.S. only, weighted)

DRB approval number CBDRB-FY19-RAGLIN-B0010

Methodology

- Used 2010 PES data to inform a decision on the 2020 PES
- Simulated eliminating the clerical match
  - Used computer match to determine matches where possible
  - Used logistic regression model for match status of computer nonmatches
  - Concordance and cross-validation as model assessments

Determining Match Status for Independently Listed Housing Units

- Challenge: Computer can establish a match with confidence
  - But it cannot establish a nonmatch with confidence
  - No data available to establish a nonmatch with confidence
- Used 2010 PES data to build predictive model for match status
  - Model the clerical match status of the 26,000+ computer nonmatches (modeling universe)

Limitations

- Would have to use the predictive model determined from the 2010 PES for 2020 PES data
  - Relationships between variables may change between 2010 and 2020
- Matching error – computer or clerk can incorrectly assign a match

Results - A Useful Covariate

- Existence of a person computer match in the housing unit
  - is strongly suggestive that the housing unit is matched
- Odds ratio of 15 for a housing unit match versus a nonmatch
  - given a person computer match versus no person computer match in the HU

Two Competing Models

Model 1 – Fewer Parameters

- Potentially more robust to model misspecification
  - changes in relationships between variables in 2010 and 2020 data
- 9 parameters
- Vacant, occupied renter, occupied owner, single unit, multi-unit, etc.
- No interaction terms

Model 2 – More Parameters

- Potentially better prediction
- Determined by stepwise regression
- 44 parameters
- Three additional covariates
- Includes interaction terms

Percent Concordance

Model	Percent Concordance	Percent Discordant
1	67.0	29.8
2	71.1	28.1

DRB approval number CBDRB-FY19-RAGLIN-B0010

Predicted Match Rates for One of Ten Random Groups Based on the other Nine Computer Nonmatched Housing Units (in Percent)

Region	2010 PES Results	Model 1	Model 2
Northeast	87.5	85.2	84.9
Midwest	81.4	75.4	78.5
South	77.5	73.3	71.4
West	79.0	75.8	75.3

DRB approval number CBDRB-FY19-RAGLIN-B0010

Predicted Match Rate for Florida Based on the Rest of the Nation for Computer Nonmatched Housing Units (in Percent)

2010 PES Results	Model 1	Model 2
68.4	75.6	74.3

DRB approval number CBDRB-FY19-RAGLIN-B0010

Conclusions and Future Research

- Determining match status is problematic
  - Many computer nonmatches for housing units are matches
  - Have to use predictive model based on 2010 PES to predict match status for 2020 PES data
  - Models introduce variance and potential bias
- We cannot recommend with confidence eliminating the housing unit clerical match and replacing it with modeling
- Use the proposed methods with the 2020 PES data

Contact Information

- Michael Beaghen
  - Michael.A.Beaghen@census.gov
- Mark Jost
  - Mary.L.Jost@census.gov
- Elizabeth Marra
  - Elizabeth.Marra@census.gov