Using Statistical Models in Place of Clerical Matching in the Census 2020 Post-Enumeration Survey to Produce Estimates of Census Housing Unit Coverage[1]

Michael Beaghen[1], Mark L. Jost[1], Elizabeth Marra[1]
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
The Census Bureau will conduct a Post-Enumeration Survey (PES) to assess the 2020 Census' coverage of population and housing units (HU). Census HU coverage errors include omissions, duplication, HUs enumerated in the wrong place, and HUs that should not have been enumerated. The 2020 PES will be a probability sample of about 170,000 HUs nationwide. PES methodology requires an Independent Listing of HUs and people in HUs in a sample of geographies. This independence is necessary to satisfy the requirements for the dual-system estimator. The PES listings are matched to the census listings of HUs and people. Accurate matching includes automated, probability-based matching, followed by clerical matching, and by field interviewing to resolve differences in the listings. Since the clerical matching operations are time consuming and expensive, successful models to replace them could save time and money in future Census coverage measurement programs.

Key words: Census coverage error, dual-system estimation, record linkage

1 Introduction

In this research we explored the feasibility of using logistic regression models to replace the HU clerical matching operations in the 2020 PES. We simulated a PES without clerical matching using the results of the 2010 PES. This report only documents research into eliminating the HU clerical match and field followup (we consider the followup a part of the clerical match). Reisch (2019) conducted analogous research that explored the elimination of the PES person clerical match and field followup. The results presented in this report are a summary of key results which are more completely documented in Beaghen et al. (2019).

2 Overview of PES Methodology

In this section we provide an overview of the PES design, operations, and estimation. For more details on the 2010 PES, see U.S. Census Bureau (2008). The 2020 PES will be very similar in design to the 2010 PES.

[1] Any views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. CBDRB-FY19-582

Starting with the 1950 Census, the U.S. Census Bureau has conducted post-enumeration surveys to evaluate the census coverage of the population of people and HUs. However, it was not until the 1980 Census that the Census Bureau started using dual-system estimation to measure the true population size. Similarly, the Census Bureau will conduct a PES and use dual-system estimation to assess the coverage of the 2020 Census and to aid in the design of future censuses. The Census 2010 PES was called the Census Coverage Measurement and the Census 2000 PES was called the Accuracy and Coverage Evaluation.

2.1 **PES Sample Design**

The 2020 PES will be a probability sample of about 170,000 HUs nationwide. Remote areas of Alaska, group quarters facilities, and people residing in group quarters facilities are out of scope for the PES. The PES will also have a sample of about 7,500 HUs in Puerto Rico.

2.2 **PES Operations**

The PES first conducts the Initial Housing Unit (IHU) operation. The IHU begins with the Independent Listing, a field operation to create an address list of all HUs in the geographic areas selected for the 2020 PES sample. This listing will be conducted in the winter of 2020, and will be completely independent from the 2020 Census operations.

The IHU operations have two goals. First, is to determine which of the independently listed HUs match to census HU enumerations. Second, to classify census enumerations as correct or erroneous. The list of independently listed HUs used to determine the match rate to census enumerations is called the P sample, and the list of census enumerations in the PES sample geography is called the E sample.

The matching begins with automated, probability-based computer matching of the P-sample HUs to the census enumerations. The computer match identifies matches and possible matches; the remainder are nonmatches. Expert clerical matching staff reviews both matches and nonmatches, and assigns detailed codes that are used to determine which cases are sent to the field in followup and what information needs to be collected during the followup to resolve the case. These matching and follow up operations allow clerks to determine the HU status of the listings, that is, whether the address existed as a livable HU at the time of the Census. For the E-sample, these statuses are referred to as correct or erroneous enumerations.

After the completion of the Independent Listing and IHU clerical matching operations, the PES Person Interview (PI) operation lists people in the valid P-sample and E-sample HU listings. The PI also collects information that allows for the determination of the HU status. A computer match attempts to match P-sample people to census person enumerations. This computer match is followed by clerical matching and a field followup to resolve differences.

Lastly, there is an additional round of HU clerical matching and field followup, the Final Housing Unit operation, to process late census changes to its final inventory of HU enumerations.

2.3 **Census Coverage Definitions**

Census coverage refers to how completely and accurately the census enumerates the population. Coverage errors include omissions and erroneous enumerations. Erroneous enumerations include duplicated enumerations of HUs and those that should not have been enumerated anywhere on Census Day (April 1 of the Census year), such as a coffee shop. Furthermore, to be correctly enumerated, HUs should be counted in a geographic area defined by the sample basic collection unit and a ring of surrounding basic collection units (the basic collection unit is the smallest geographic area for 2020 Census operations and roughly equivalent to a census block).

The PES estimate for the population is referred to as the dual-system estimate, or DSE. The net coverage error is defined as follows.

Net Coverage Error = Census Count - DSE

2.4 **PES Estimation Methodology**

The 2020 PES will use estimation methods that are very similar to the 2010 PES methods. For details on the 2010 PES estimation see Olson and Viehdorfer (2013).

Dual-system estimation requires two independent systems of measurement. In the PES, these are the P sample and the census correct enumerations (which are determined by the matching and followup). To estimate a DSE, the correct enumeration and match probabilities are derived from the sample cases and their statuses as determined in the HU operations. These probabilities are applied to each individual HU in the census as described below.

The DSE for a domain C is the sum of the HUs in that domain (e.g., total, owner/renter/vacant, age/sex groupings, etc.), weighted by the correct enumeration rate and the inverse of the match rate. The formula for the DSE for domain C is shown below.

Where

$\pi_{ce,j}$ is the probability the census enumeration is a correct enumeration
$\pi_{m,j}$ is the probability the census enumeration was matched
j is a census enumeration

$$DSE_C = \sum_{j \in C} \frac{\pi_{ce,j}}{\pi_{m,j}}$$

3 2010 PES Initial Housing Unit Matching Results

To illustrate what the clerical matching and followup accomplish, we summarize the 2010 PES IHU matching results in Table 1. Table 1 shows the HU match status assignment before and after clerical matching in the 2010 PES IHU match. The computer matching was able to identify the bulk of the HU matches. After computer matching, 72.9 percent of the HUs found in the Independent Listing operation matched to HUs in the preliminary 2010 Census inventory of HU enumerations. (The results in Table 1 are weighted.) Clerical matching analysts and technicians reviewed the results of the computer match in the clerical matching operation and found new matches. A field followup collected additional information during the clerical match phase to help clerical technicians determine the match status. As a result of the IHU clerical matching operations, the share of matches

increased from 72.9 percent to 93.5 percent. At the same time, the share of nonmatches decreased from 15.4 percent to 3.7 percent (Contreras et al., 2012).

Table 1: 2010 PES Initial Housing Unit Match Status Assignment

Match Status	Computer Matching	Clerical Matching
Match	72.9	93.5
Possible Match	11.4	N/A
Nonmatch	15.4	3.7
Duplicate	0.27	0.08
Not a Housing Unit	N/A	2.8

4 Research Methodology

The goal of the research was to produce sound DSEs by estimation domain, as the PES produces estimates of census coverage for the nation and each state by certain domains of interest. Thus our criteria for success was how well we could produce correct enumeration rates and match rates at the domain level; it was not how well we could predict statuses for the individual HUs. For our assessment we compared the predicted match rates to those of the 2010 PES at the level of domains.

In our research we used the 2010 PES data for the U.S. excluding Puerto Rico.

During the clerical matching, analysts and technicians assigned three statuses. Since these statuses were used to produce HU estimates, we would need to develop methods to assign the three statuses, if we were to skip the clerical match operations. The three statuses are:

- E-sample enumeration status: correctly enumerated or erroneously enumerated.
- P-sample Census Day HU status: valid or invalid HU on Census Day (it is the P-sample analogue to E-sample enumeration status).
- P-sample match status: either the P-sample HU refers to a census HU enumeration; or no match is found to a census HU enumeration.

We used available data wherever possible to directly assign these statuses. Such data included the following.

- The PI
- The HU computer match results

The steps for assigning statuses were as follows.

- We used the PI to directly assign HU status to the great majority of P-sample HUs and to assign the enumeration status to the great majority of E-sample HUs. We discuss the use of the PI in more detail in Section 6.
- We used the computer match to obtain the match status for many of the valid P-sample HUs.
- Where we could not obtain a status from the computer match or PI, we used logistic regression models to assign predicted probabilities of the status. In our logistic regression models we used covariates obtained from the PES processing, which we discuss in Section 8.

We used the 2010 PES delete-a-group jackknife replication method (Imel et al., 2013) to estimate the standard errors of the estimates and predicted values.

5 Limitations

In this report we give the greatest attention to the assignment of P-sample match status. We more briefly discuss the assignment of HU status to the P-sample and correct enumeration status to the E-sample. We made this choice of emphasis because modeling the match status was the biggest challenge to make this project work. Determining the HU status of the P-sample HUs and the enumeration status of the E-sample HUs was not problematic, as we used the PI information to assign the great majority of statuses (described in Sections 6 and 7). In contrast, the biggest limitations of the DSE produced without a clerical match resulted from the challenges in assigning match status.

The reason assigning the match status for computer nonmatches is challenging is that we would have to use a predictive model fitted on 2010 PES data to predict match status for the 2020 PES. But the 2020 PES could differ in important ways from the 2010 PES data in the relationships between the predictive variables and match status. To use 2010 PES data to predict 2020 PES match status would require an assumption that we cannot test at this time.

6 Assigning the P-Sample Housing Unit Status

In this section we describe how we used the results of the PI to assign HU status to P-sample HUs. In addition to collecting person information, the PI collects information that allows for the determination of the Census Day status of the HU. After the collection of the PI data, the PES automatically assigns codes indicating whether the HU was valid at the time of the Census. All P-sample HUs go to PI, except those determined in the IHU match and followup not to be valid HUs. Thus P-sample HUs with an uncertain Census Day status also go to PI. Such cases include a fair number of HUs under construction or future construction sites at the time of IHU operations. The PI is of particular value because it would take place even if the clerical match did not.

In Table 2 we see we can use the PI to reliably assign the HU status for the great majority of P-sample HUs. The HU status determined by PI is highly consistent with the PES HU status determined by PES HU operations, i.e., the clerical review and field followup (these include both the initial and final HU operations). This consistency is not surprising; if the PES HU operations and the PI results were not consistent, we would have to doubt the overall soundness of PES methods.

We make the following observations of Table 2.

- Of the 163,000[2] HUs which PI determined to be valid HUs, 162,000 had a valid HU status as determined by the PES HU operations; this includes both initial and final housing unit operations.
- Of the 4,200 HUs which PI determined to be invalid HUs, the PES HU operations determined 3,700 to be invalid.
- There were more than 1,000 inconsistently coded HUs: about 600 valid by PI but invalid by PES, and about 450 invalid by PI but valid by PES. These 600 and the 450 HUs mostly balance each other out, suggesting that the differences between

[2] Note that the housing unit counts presented in Sections 6 and 7 have been rounded.

the PI and PES results were largely due to random interview or respondent error. It was beyond the scope of this research to investigate these inconsistencies.

Table 2 HU Status from PI Versus HU Status from Housing Unit Operations 2010 PES[3]

HU Status from Person Interview	HU Status from HU Operations		
	Valid	Invalid	Total
Valid Census Day Status	162,000	600	163,000
Invalid Census Day Status	450	3,700	4,200
Total	163,000	4,300	167,000

Not shown in Table 2 are the 4,200 P-sample HUs for which the PI could not determine the HU status because of incomplete interviews. We built logistic regression models based on the 167,000 HUs with 2010 PI data to assign the HU validity status for these 4,200 cases (see Beaghen et al., 2019).

7 Assigning Correct Enumeration Status

The 2010 PES assigned correct enumeration status to the E-sample HU enumerations based on the computer match and the clerical match with field followup operations. For this research the PI provided the HU status for the large majority of E-sample HUs. As with the P-sample HUs, for the E-sample HUs, the PI collects information to determine the HU status at the time of Census Day, that is, their enumeration status. The PI could successfully assign enumeration status to about 160,000 E-sample HUs.

For the 7,900 E-sample HUs that we could not assign based on the PI (not shown in any table), we predicted enumeration status with a logistic regression models built on the 160,000 whose status could be assigned with the PI. For details on these logistic regression models, see Beaghen et al. (2019).

8 Assigning Match Status

In contrast to determining HU status or correct enumeration status, which we largely obtained from the PI, assigning the match status is more challenging. While we have confidence in those matches made by the computer match, without clerical match and followup, we did not have PES data to allow us to confidently build a model to predict the match status of those P-sample nonmatches from the computer match operation.

8.1 Assessing the Automated Computer Assignment of Match Status

To assess the soundness of the match status assigned by the computer match, we compared the computer match statuses with the final, 2010 PES match statuses after HU clerical matching and followup. We examined HUs with the following three match statuses from the computer matching: match, possible match, and nonmatch. The results are below. Note that these estimates are unweighted and exclude P-sample HUs determined not to be valid HUs or with unresolved HU status.

- 99.9% of computer matches stayed matches in clerical matching.
- 98.5% of computer possible matches were confirmed matched in clerical matching.
- 81.0% of computer nonmatches were matched in clerical matching.

[3] Note that the sums of columns and rows do not equal the totals because of rounding.

We conclude that we can confidently use the computer match codes to assign most matches. Of the HUs which the computer identified as matches, 99.9% remained matched after the clerical match. The assignment of match status to HUs that the computer match identified as possible matches was also not problematic, as even simple models yield predictions sufficiently sound for our purposes (see Beaghen et al., 2019).

8.2 Covariates in the Model for Predicting Match Rate

We identified several variables that were predictive of match status. Foremost, the existence in a P-sample HU of a household member who matches to a census person enumeration in the Person Computer Match (Person Link Indicator) is strong evidence that the HU also matches; see Beaghen et al. (2019). Also with strong predictive power was the HU status determined by the Independent Listing; addresses that the Independent Listing indicated were not valid HUs at the time of the listing, such as under construction or unfit for habitation, were less likely to be matched. In addition to these two covariates, there are variables that have a history of predictive value for match status or correct enumeration status in PES estimation, such as type of HU structure, owner/renter/vacant, or city style/non city style address (see Olson and Viehdorfer, 2013). We describe these covariates and others in Section 8.4.

8.3 Methodology for Assigning Match Status for Computer Nonmatches

We fitted two logistic regression models to the 2010 PES data to predict match status. We wanted a highly parameterized model that would have strong predictive value, even at the risk of over parameterization. For comparison, we wanted a model with main effects we knew from experience would be predictive of match status.

The model universe was the 27,000 nonmatches from computer matching. For the response variable we used the match status from the 2010 PES HU clerical match and followup. The success in the logistic regression was a match and a failure was a nonmatch. We applied the parameter estimates to predict a probability between 0 and 1 for match status for these HUs.

To build the first, more parameterized model, we ran a stepwise variable selection with alpha = 0.1. To build the second model with fewer parameters, we were guided by experience with PES data in addition to hypothesis tests. To assess the fit of the logistic regression models, we used the percent concordance and cross-validation methods.

8.4 Results of Modeling for Match Status

The covariates that the stepwise selection found significant are listed below.

- Person Link Indicator: whether a P-sample HU had a person who matched to a census person enumeration
- Independent Listing Status of HU: the HU was or was not a valid HU at the time of Independent Listing
- Address Characteristic Type: city style/non-city style
- Independent Listing Type of Structure: single-unit structure, multi-unit structure, mobile home/trailer park, etc.
- Region: one of four census Regions in the U.S.

- Metropolitan Statistical Area (MSA) Size by Type of Enumeration Area: large MSA mail out/mail back, medium MSA mail out/mail back, small MSA mail out/mail back, other
- Bilingual Block: whether the collection block received the bilingual questionnaire
- Replacement Mailing Status: none, targeted, blanketed interaction of bilingual block and replacement mailing status
- Interactions of the Person Link Indicator with: Metropolitan Statistical Area, Region, Replacement Mailing Status, Bilingual Block, IL Status of the HU, IL Type of Structure, and Address Characteristic Type; the interactions of Bilingual Block and Replacement Mailing Status

Model 1 included all of the terms included by the stepwise regression, which we listed above. Model 2 included a subset of the terms included in Model 1: Person Link Indicator, Type of Structure, Occupied/Vacant, IL Status of the HU, and Region. Further, Model 2 included no interaction terms. Model 1 had 44 degrees of freedom, while Model 2 had nine. For more details on the models see Beaghen et al. (2019).

8.5 **Model Fit for Match Status**

Table 3 shows the model fits in terms of the percent concordance. As one would expect, Model 1, with more parameters in the model, had a higher concordance. These concordance rates are similar to rates obtained in other predictive models used to impute enumeration status or match rates.

Table 3 Model Percent Concordance

Model	Percent Concordance
1	71.1
2	67.0

We implemented the cross-validation to assess the model fit as follows.

1) Each P-sample HU was randomly assigned to one of ten groups.
2) For each of the ten groups, we predicted the probability of match for each HU, with a predicted value calculated from the model fitted to the data in the remaining nine groups.
3) The measure of fit was the sum of the squared differences between the predicted values for each left out group and the observed values for each group.

Table 4 shows the results of the cross-validation. We see that the differences between Models 1 and 2 in the cross-validation measures of fit are small compared to the overall measures of fit for each model, suggesting the two models have about equal predictive power.

Table 4 Cross-Validation Results for Predicted Match Rates for Two Models[4]

Squared Differences by Race and Hispanic Origin of the Householder			
Race Group	Model 1	Model 2	Difference
Overall	419	424	-5
Vacant Unit	106	106	0
AIR on Reservation	31	29	3
AIR off Reservation	5	5	0
Hispanic	33	33	0
Non-Hispanic Black	39	39	0
NHPI	3	3	0
Asian	21	20	1
White or Other	186	185	0

8.6 Comparisons of Predicted Match Rates for Computer Nonmatches by Estimation Domain

We compared the match rates for the computer nonmatches predicted by the models and the official 2010 PES IHU match rates for several domains of interest. Table 5 shows the predicted match rates for the domain Race and Hispanic Origin of the Householder. We chose this domain because it was not included in the predictive model. The 2010 PES match rates and predicted match rates are similar but not the same in these domains because of model misspecification and sample variation. Note that the standard errors of the estimated match rates are in parenthesis below the estimate itself.

Table 5 Predicted Match Rates in Percent for Computer Nonmatches by Race and Hispanic Origin of the Householder

Race and Hispanic Origin of the Householder	2010 PES	Model 1	Model 2
American Indian Living on Reservation	82.0 (5.2)	76.6 (4.5)	71.7 (5.1)
American Indian Living off Reservation	77.1 (3.7)	74.5 (2.6)	73.8 (2.3)
Hispanic Origin	82.7 (2.0)	81.5 (1.6)	80.6 (1.6)
Non-Hispanic Black	77.8 (3.6)	77.4 (2.5)	79.5 (1.9)
Native Hawaiian/Pacific Islander	62.4 (12.4)	69.0 (7.0)	70.0 (6.0)
Asian	65.0 (17.0)	71.9 (7.8)	75.3 (4.9)
White or Other	77.4 (1.4)	77.3 (1.4)	76.8 (1.6)

8.7 Predicting the Match Rate for Florida

We predicted the Florida match rate for P-sample HUs not matched in the computer match, based on the model fitted on data from the rest of the nation minus Florida (i.e., the other

[4] Note that the values in Table 4 are rounded.

49 states and DC). Since Florida may be systematically different from the rest of the nation, there is greater potential for model misspecification bias. Indeed, in Table 6 we see the predicted match rates for the category Single Family are suggestive of model misspecification.

We also notice in Table 6 that the standard errors of prediction may be lower than the standard errors of the estimates based on the observed data. This is an artifact of prediction, because the predicted values are dispersed about a mean while the observed data are either 1 or 0.

Table 6 Predicted Match Rates in Percent for Computer Nonmatches for Florida by Type of Structure

Type of Structure	2010 PES	Model 1	Model 2
Overall	68.4 (8.7)	74.3 (4.5)	75.6 (2.6)
1- Single Family	43.6 (17.0)	64.5 (5.8)	67.8 (3.3)
2- Multiunit	79.0 (13.0)	76.6 (5.4)	76.4 (4.1)
3- All Other	80.5 (5.8)	83.6 (3.6)	85.4 (2.7)

8.8 **Discussion of the Predicted Match Rates**

A model with more parameter terms potentially has more predictive power, whereas a model with fewer terms may be more robust to model misspecification. Model misspecification is a particular concern because the 2010 PES data may differ in important ways from 2020 PES data. Model 1 had a higher concordance rate, but its advantage in predictive power was not supported by the cross-validation analysis. For these reasons the simpler Model 2 may be preferable to Model 1.

The predicted match rates for Race and Hispanic Origin of the Householder are arguably close to the observed 2010 PES match rate, despite potential model misspecification and sample variability. In contrast, the predicted match rates for Florida fitted on U.S. data excepting Florida suggest model misspecification. This sensitivity of the predicted match rates to model misspecification is a caution for using 2010 PES data to predict 2020 PES match rates.

9 Conclusions

In this research we demonstrate that with the information collected in the Independent Listing and the PI, with the HU and person computer matching results, along with statistical modeling, we can soundly assign enumeration status to E-sample HUs, and the analogous HU status to P-sample HUs. However, we do not as of yet have methods for assigning match status in which we are equally confident. For most P-sample HUs we can assign a match status with high confidence based on computer-identified matches. We developed predictive models based on the 2010 PES data for the P-sample HUs without computer matches. But they assumed the 2020 PES data have similar relationships between variables as do the 2010 PES data. If the mechanisms generating P-sample to census nonmatches differ between 2010 and 2020, then the estimates of match rates based on this model could be seriously biased.

If the Census Bureau were unable to implement clerical matching and field followup in the 2020 PES, we would have a methodology prepared to produce DSEs of Census 2020 HU coverage. But the proposed methodology for assigning match status would remain unproven and present risks to the quality of the DSEs.

10 Future Research

Assuming the Census Bureau implements the HU clerical match in the 2020 PES, we can test our proposed methodology and compare its results to the actual 2020 PES results.

There may be partial alternatives to eliminating all of the HU and person clerical matching and followup operations. For example, it may be feasible to eliminate some of the clerical matches and field followup operations, but not all of them. An important future line of research could be to eliminate the clerical match but incorporate the field followup. Another possibility is that it may prove possible to eliminate HU clerical matching and followup, though not the analogous person operations.

Ultimately, improving the computer matching would reduce the risks from having to model match rates for nonmatching HUs. A sufficiently strong computer match algorithm could pave the way for a completely automated match.

References

Beaghen, M., Jost, M., and Marra, E. (2019). "Using Statistical Models in Place of Clerical Matching in the Census 2020 Post-Enumeration Survey to Produce Estimates of Census Housing Unit Coverage." DSSD 2020 Post-Enumeration Survey Memorandum Series #2020-E-36.

Contreras, G., Cronkite, D., Rosenberger, L., Wakim, A., and Argarin, A. (2012). "Assessment for the 2010 Census Coverage Measurement Initial Housing Unit Independent Listing, Matching, and Followup Operations – Reissue." DSSD 2010 Census Coverage Measurement Series #2010 – I – 14 – R1.

Imel, L., Mule, V., and Seiss, M. (2013). "2010 Census Coverage Measurement Estimation Methods: Measures of Variation." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-J-03.

Olson, D., and Viehdorfer, C. (2013). "2010 Census Coverage Measurement Estimation Methods: Net Coverage Estimation." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-J-04.

Reisch, G., (2019). "Imputation Models Using Automated Probability Matching Results." Proceedings of the 2019 Joint Statistical Meetings. American Statistical Association.

U.S. Census Bureau (2008). "The Design of the Coverage Measurement Program for the 2010 Census." DSSD 2010 Census Coverage Measurement Memorandum Series #2010-B-07. https://www.census.gov/coverage_measurement/pdfs/2010-B-07.pdf