Imputation Models Using Computer Matching Results

Glenn Reisch
U.S Census Bureau, Washington, DC, 20233
(September 30, 2019)[1]

Abstract

The major goal of the 2010 Census Coverage Measurement (CCM) program was to measure coverage for housing units and people in the 2010 Census using dual-system estimation. As part of this process, person interviews were conducted at a sample of housing units collected independently of Census operations. After computer matching of this data to Census results, extensive clerical review and field followup were performed on nonmatches and cases with unresolved statuses.

The 2020 Post-Enumeration Survey (PES) has the same goals and approaches as the 2010 CCM, which includes clerical matching and field followup. However, in this research, we investigated using models instead of clerical review or followup operations. Not including clerical review or followup operations would lead to larger numbers of nonmatches and unresolved statuses. We used models and information from the 2010 CCM to impute the match status and unresolved statuses. We then used the final 2010 CCM results to evaluate coverage estimates based on computer matching output.

Key Words: Post-Enumeration Survey, Imputation, Clerical Matching, Census Coverage

1 Introduction

The purpose of the 2020 Post-Enumeration Survey (PES) is to measure the coverage of the 2020 Census. The coverage measures resulting from the PES allow us to evaluate the quality of census counts as well as help improve future census operations. PES data are matched to census data using a computer matching system. After computer matching, the data undergo extensive clerical matching and field followup. During these operations, matches, potential matches, and nonmatches are reviewed and cases that need additional information are sent back into the field. After collecting additional information for these cases, they are subjected to further clerical review. These operations reduce the number of cases that need imputation and increase the accuracy of the data. However, clerical matching and field followup are time consuming and expensive.

In this research, I investigated estimating census coverage using data directly from computer matching results before clerical review and followup operations. I used data from

[1] This paper is released to inform interested parties of research and to encourage discussion. Any views expressed are those of the author 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-FY19-571

the 2010 Census Coverage Measurement (CCM) program, the post-enumeration survey for the 2010 Census. I calculated estimates from the 2010 CCM using data before clerical review and followup, and compared these estimates with the 2010 CCM results. To do this, I adapted the models used to estimate coverage in 2010 to function with computer matching data and created decision rules and procedures designed to improve the results given the lack of clerical matching and field followup. My research focused on person net coverage for the 50 states and DC, although the 2010 CCM produced other measures of census coverage for people and housing units.

In this paper, I first discuss the 2010 CCM methodology, including dual-system estimation, the 2010 CCM operations, and the imputation and estimation procedures. After that, I present my methodology for two models using data from computer matching. Finally, I compare rates and estimates from my models to those from 2010.

2 Overview of 2010 CCM Methodology

2.1 Dual-system Estimation

Census coverage refers to how completely and accurately a census enumerates the population. The 2010 CCM program used a dual-system estimate (DSE) to measure the coverage of the 2010 Census. Dual-system estimation is used in most post-enumeration surveys, including the 2020 PES, to produce an estimate of the true population size. This type of estimation is based on capture-recapture methodology and has been used by the Census Bureau since 1980 (Mulry and Cantwell, 2010). To implement the DSE, the PES is conducted in an independent area-based sample of housing units. This sample is referred to as the population sample, or P sample. The P sample is compared with a sample of census enumerations in the same areas, referred to as the enumeration sample, or E sample (see 2.2 for further details). The P- and E- sample records can be placed into one of four cells shown in Figure 1.

		In P Sample (PES)		
		Yes	**No**	**Total**
In E Sample (Census)	**Yes**	N_{11}	N_{12}	N_{1+}
	No	N_{21}	N_{22}	N_{2+}
	Total	N_{+1}	N_{+2}	N

Figure 1: Classification of P and E Samples into 2x2 Matrix

Assuming that the P and E samples are independent and each unit has a chance of being in the E sample and the P sample, the "true" population size can be estimated using the DSE as follows:

$$\widehat{N} = N_{1+} \frac{N_{+1}}{N11} \quad (1)$$

Equation 1 is based on a standard Petersen (1896) or Sekar-Deming estimator to measure the size of the true population. Wolter (1986) discusses assumptions and conditions for the DSE. For this DSE, $\widehat{N}$ is an estimator for the unknown population total N. E-sample total (N_{1+}), P-sample total (N_{+1}), and records in both samples (N_{11}) are observed.

In 2010, the DSE for people was calculated using equation 2 with correct enumeration and match probabilities as documented in Viehdorfer (2010):

$$DSE = \sum_{j \in C} \pi_{dd(j)} \times \frac{\pi_{ce(j)}}{\pi_{m(j)}} \times CB_j \quad (2)$$

π_{dd} = predicted data-defined probability
π_{ce} = predicted correct enumeration probability
π_{m} = predicted match probability
CB = correlation bias adjustment factor where $j \in C$ represents people enumerated in the census.

In this equation, the predictions for data-defined, correct enumeration, and match probabilities were obtained through logistic regression modeling. The correct enumeration probability is the probability that a person enumeration is correctly included in the census. Correct enumeration status is determined for each enumeration in the E sample. Then, correct enumeration status is modeled on the E sample using logistic regression. The model coefficients are applied to all census enumerations to compute $\pi_{ce.}$ The match probability is the probability that a record in the P sample matches to a correct census enumeration. The match status is determined for each person record in the P sample. Then, the match status is modeled on the P sample using logistic regression. The model coefficients are applied to all census enumerations to compute $\pi_{m.}$ Sections 2.3 and 2.4 contain more information on match and correct enumeration probabilities. For more information on the data-defined probability and correlation bias, see Viehdorfer (2011). These are out-of-scope for this research.

2.2 **2010 CCM Operations**

To implement the DSE to measure person coverage, the following operations were performed in 2010:

1. Sampling
2. Independent listing

3. Person Interview
4. Computer Matching
5. Clerical Matching and Field Followup
6. Estimation.

The 2010 coverage measurement survey sample was a probability sample of approximately 170,000 housing units in the U.S. (excluding remote Alaska) and approximately 7,500 housing units in Puerto Rico. Two samples are selected to measure census coverage of the household population: the P sample and the E sample. The P sample is a sample of housing units and persons obtained independently from the census for a sample of block clusters. The E sample is a sample of census housing unit and person enumerations in the same block clusters as the P sample. The 2010 CCM used block clusters as their primary sampling unit. Block clusters consisted of one or more geographically contiguous census blocks containing on average 30 housing units. An independent address list was created for each sample block cluster. Within each selected block cluster, a canvassing operation was conducted to construct list of all housing units. This list was made independently of census operations.

For housing units selected during the sampling operation, a person interview (PI) was conducted. The information collected during each PI included name, sex, age, date of birth, race, relationship, and Hispanic origin for each person in the housing unit. The interviewer also collected information about alternate addresses to establish where people lived on Census Day (April 1, 2010). The person data collected during the PI was then matched with census enumerations using a probabilistic computer matching algorithm. After computer matching, clerical matching and field followup operations began. The clerical matching staff reviewed all matches, potential matches, and nonmatches. During their review, they assigned detailed codes that were used to determine which cases were sent to the field in followup and what information needed to be collected during followup to resolve the case. After field followup operations, an additional clerical matching operation used the data collected in field followup to attempt to resolve the remaining cases.

The estimation operation consisted of several processes to produce estimates of coverage. This included imputation procedures to account for missing or unresolved data. In the estimation operation, the DSEs, match probabilities, and correct enumeration probabilities were determined using logistic regression. Match probabilities and correct enumeration probabilities are further described in the next two sections.

2.3 **Match Probability and the P sample**

The match probability is the probability that the P-sample record matches to a correct census enumeration in the correct search area (see Section 2.4 for more information on correct census enumerations). During clerical matching, technicians and analysts assigned an inclusion status, inmover status, and match status for all people in the P sample. Match probabilities were calculated using the codes assigned during clerical matching.

The inclusion status indicated whether a person record should have been included in the P sample. If the person was appropriately included in the P sample, then the person could be used when calculating the match probability.

There was a time gap between the PI and the 2010 Census. This was an issue for matching because people moved between Census Day and the time of the PI. People who moved into a sampled address at the time of PI but lived in a different in-scope housing unit on Census Day are referred to as inmovers.

Each person enumerated during the PI was assigned a residence status code that described where people were living based on their Census Day and interview day addresses. The residence status code was the main factor used to determine the P-sample inclusion status and the inmover status. The residence status codes and their descriptions are listed in Table 1.

Table 1: Residence Status Codes

Residence Status Code	Definition	Assigned during Computer Matching	Assigned Clerically
Inmover	A person who was an interview day resident of the PI housing unit but was a Census Day resident of a different in-scope housing unit.	X	X
Never Resident	A person who should be counted at another in-scope address.		X
Nonmover	A person who was a Census Day resident and interview day resident at the same PI housing unit.	X	X
Outmover	Census Day resident but not an interview day resident	X	X
Out-of-scope	A person who was a Census Day resident of a group quarters or of a housing unit that is outside the nation.	X	X
Review	This code is assigned by PI post processing to cases that need clerical review.	X	
Unclassified	A person who cannot be classified because not enough information was obtained.	X	X

The inmover residence status code indicated if the person was an inmover or nonmover, and specifies if the correct search area for the matching census enumeration surrounds the sample address or a different Census Day address.

When determining match probability, records with a match in the correct search area had a probability of one assigned to them. Nonmatches, duplicates and matches in the incorrect search area were assigned a probability of zero. Possible matches were treated as having an unresolved match status.

More specifically, the following rules were applied to determine match probability based on inmover status:

- For determining match status for an inmover, the correct address was in the Census Day address search area. For inmover cases with an unresolved match status, an inmover match probability was imputed using logistic regression.

- For determining match status for a nonmover, the correct address was in the sample address search area. For nonmover cases with an unresolved match status, a nonmover match probability was imputed using logistic regression.

- Since both the inmover status and the match status could have been unresolved, a final match probability for unresolved cases based on the following conditional probability formula was assigned:

$$P_{mat} = P_{inmov} \times P_{mat|inmov} + (1 - P_{inmov}) \times P_{mat|not\ inmov} \qquad (3)$$

where P_{mat} is the overall match probability,
P_{inmov} is the probability of being an inmover,
$P_{mat|inmov}$ is the probability of matching given the case is an inmover, and
$P_{mat|not\ inmov}$ is the probability of matching given the case is not an inmover.

This final match probability was used to compute the match component of the DSE. See Konicki et al. (2013) for further details.

2.4 **Correct Enumeration Status and the E sample**

The correct enumeration status indicated whether an enumeration should have been included in the census. The records with a resolved enumeration status were either correct or erroneous enumerations. To be a correct enumeration for dual-system estimation, an E-sample person record had to meet four criteria (Hogan, 2003):

1. Appropriateness
2. Uniqueness
3. Completeness
4. Geographic correctness

"Appropriateness" means that the person should be included in the census. This means that they were alive on Census Day and that the records do not refer to fictitious "people," tourists or animals. "Uniqueness" means that there should be one record per person. If two records refer to the same person, one is correct and the other is a duplicate (erroneous enumeration). "Completeness" means that the record must be sufficient to identify a single person. This means that there are at least two characteristics, one of which is a valid name. "Geographic correctness" means that people are included in the census where they should be included. Cases found in the wrong search area considered erroneous enumerations for dual-system estimation.

Technicians and analysts determined the enumeration status of all people in the E sample during the clerical matching activities. If the technicians and analysts could not determine an enumeration status, the enumeration status was imputed during estimation. Then, a logistic regression model to predict enumeration status was fit on the E sample and the coefficients were used to predict the enumeration status for all people enumerated in the census.

3 Methodology

My goal is to estimate census coverage with computer matching data instead of with data that has gone through clerical matching and field followup. In this section, I will present the methodology for two models using computer matching data. To create Model 1, I only made changes to the 2010 CCM methodology to make sure the 2010 CCM imputation and estimation models would run with the computer matching data. The first model was used to establish a baseline to assess the quality of the computer matching data by comparing its match rates, correct enumeration rates, and the DSEs with those of the 2010 results.

To develop Model 2, I analyzed the effect of clerical matching and field followup on the data by looking at frequency tables of residence status and match codes before and after clerical matching and field followup.

3.1 Model Methodology

3.1.1 Model 1

The only changes to 2010 CCM methodology for Model 1 were done to assure that the imputation and estimation models would run with the output from computer matching.

The first step in creating a model with computer matching output is to model the residence status codes for the cases marked "Review" (see Table 1). These were cases marked for clerical review during computer matching. Approximately 19.0% of the cases were marked for review during computer matching. Review cases only exist in the computer matching data and not in the data after clerical matching and field followup. Only the residence status codes after clerical matching and field followup were used in the 2010 CCM imputation models for both the E-sample and P-sample data. Therefore, I assigned the review cases to the after clerical matching and field followup residence status codes. The after clerical matching and field followup residence status codes are those which were assigned clerically (see Table 1). A multinomial logistic regression model was used to impute probabilities with the after clerical matching and field followup residence status codes as the dependent variables. This produced a predicted probability for each residence status code for each person record. For the cases marked for review, a new record was created for each predicted residence status code. The predicted probabilities were used as weights for these new records. For cases not marked for review, no changes were made.

I ran the 2010 CCM imputation and estimation programs with these residence status codes and the computer match codes from the E- and P-sample data. The results of Model 1 are in section 4.1.

3.1.2 Model 2

In Model 2, I created decision rules and recodes to help model the match rates and correct enumeration rates. I used frequency tables of computer matching codes crossed with codes from after clerical matching and field followup to analyze the differences between the data. These cross tabulations allowed me to see which codes were affected the most by clerical

matching and field followup. For the codes where the frequencies differed greatly, I adjusted the computer matching data to more closely resemble the data after clerical matching and field followup. I assumed this would lead to a more accurate model. For most codes, the data did not change much, and I did not make any adjustments. I adjusted the inmover residence status code using a decision rule. I also adjusted the P-sample match codes for nonmatches and possible matches by using recodes. These adjustments are described in detail below.

3.1.2.1 Residence Status Codes

Just as in Model 1, the first step in creating Model 2 was to model the residence status codes. However, this time I analyzed how residence status codes assigned during computer matching were impacted by clerical matching and field followup. Table 2 shows frequencies by residence status codes from computer matching crossed with residence status codes from after clerical matching and field followup.

Table 2: Residence Status Codes (Unweighted)

Computer Matching	After Clerical Matching & Field Followup				
	Inmover (Row %)	Nonmover (Row %)	Unresolved (Row %)	Other (Row %)	Total (Col %)
Inmover	3,100	200	90	100	3,490
	88.8	5.7	2.6	2.9	0.9
Nonmover	1,700	309,000	2,000	2,700	315,400
	0.5	98.0	0.6	0.9	77.8
Review	23,000	24,000	7,900	22,000	76,900
	29.9	31.2	10.3	28.6	19.0
Unresolved	70	1,700	600	450	2,820
	2.5	60.3	21.3	16.0	0.7
Other	80	100	70	6,300	6,550
	1.2	1.5	1.1	96.2	1.6
Total	27,950	335,000	10,660	31,550	405,160
	6.9	82.7	2.6	7.8	100.0

From Table 2, it is clear that inmovers are not well identified in the computer matching data. Inmovers from computer matching data represent only 0.9% of the observations. However, in the data after clerical matching and field followup, they make up 6.9% of the cases. This difference is important because of the role inmovers play in determining the match probability (see Section 2.3). Therefore, I looked for a way to adjust the number of inmovers in computer matching data to be closer to the numbers from after clerical matching and field followup.

By looking at the specification for residence status coding (Linse, 2009) and the residence status coding input data, it was able to be determined that many of the cases identified initially as inmovers were later marked for clerical review. Therefore these cases appeared not as inmovers, but as review cases in computer matching. Over 92.1% of the cases originally marked as inmovers remained inmovers after clerical matching and field followup. Based on that, I created a decision rule to treat cases originally coded as inmovers to remain inmovers even if they were marked for review. This led to 25,500 inmovers instead of 3,490 inmovers in the computer matching data. After that, I fit the multinomial logistic regression model (as described in Section 3.1.1) to impute probabilities for the remaining review cases. This decision rule was the only change to modeling the residence status codes.

3.1.2.2 P-sample Match Status Codes

Table 3 contains frequencies of the P-sample match codes from computer matching crossed with match codes from after clerical matching and field followup.

Table 3: P-sample Match Codes (Unweighted)

Computer Matching	After Clerical Matching & Field Followup				
	Match (Row %)	Nonmatch (Row %)	Possible Match (Row %)	Other (Row %)	Total (Col %)
Match	326,000	300	150	650	327,100
	99.7	0.1	0.0	0.2	83.2
Nonmatch	21,000	34,500	200	2,400	58,100
	36.1	59.4	0.3	4.1	14.8
Possible Match	5,400	250	80	150	5,880
	91.8	4.3	1.4	2.6	1.5
Other	600	200	0	1,100	1,900
	31.6	10.5	0.0	57.9	0.5
Total	353,000	35,250	430	4,300	392,980
	89.8	9.0	0.1	1.1	100.0

One of the goals of clerical matching and field followup is to match non-matched cases. As shown in Table 3, the percent of matches is higher in the data after clerical matching and field followup (89.8%) than in computer matching data (83.2%). In addition, the percent of non-matches is lower after clerical matching and field followup (9.0%) than after computer matching (14.8%).

Matches identified by the computer matching system tend to remain matches after clerical review. As shown in Table 3, 99.7% of the 327,100 matches found during computer matching remained matches after clerical matching and field followup.

For nonmatches, this is not the case. Of the 58,100 nonmatches from computer matching, 21,000 (36.1%) became matches. Of these 21,000 cases, approximately 19,000 (90.5%) were matched during clerical matching without the need for additional information from field followup. Thus, the computer matching system rejected many cases that were later matched based on the same information. By looking at the matching tips from the 2010 CCM Before Followup Clerical Matching specification (Whitford, 2010), it seems that many of these clerical matches are easy to identify by a clerk although they may be hard to automate. By treating nonmatches as possible matches, I imputed a positive match probability for certain nonmatches while taking into account the lack of confidence in nonmatches from the computer matching system.

As 91.8% of possible matches from computer matching became matches after clerical review, I treated them as matches.

3.1.2.3 Adjustments Model 2

Based on the analysis in 3.1.2.1 and 3.1.2.2, I made the following changes for Model 2:

- Used an inmover decision rule for residence status codes.
- Treated P-sample nonmatches as possible matches and imputed a match probability.
- Treated P-sample possible matches as matches.

4 Results

4.1 Results Model 1

As explained in Section 3.1.1, Model 1 uses computer matching data with no adjustments except for the review cases.

In Table 4, the correct enumeration and match rates using Model 1 for the nation and by age/sex are compared to the 2010 CCM results. Table 4 also shows the percent differences between Model 1 and the 2010 correct enumeration and match rates.

Table 4: Model 1 Correct Enumeration and Match Rates (Weighted)						
Category	Correct Enumeration Rate			Match Rate		
	Model 1	2010	Difference	Model 1	2010	Difference
National	93.3%	91.8%	1.5%	81.8%	91.1%	-9.3%
0 to 4	92.3%	90.4%	1.9%	77.1%	88.3%	-11.2%
5 to 9	93.4%	91.8%	1.6%	81.7%	90.5%	-8.7%
10 to 17	93.9%	92.2%	1.7%	83.0%	91.5%	-8.4%
18 to 29 male	90.0%	87.1%	3.0%	67.9%	84.8%	-16.9%
18 to 29 female	91.1%	88.3%	2.9%	68.1%	86.4%	-18.3%
30 to 49 male	93.1%	92.1%	1.0%	82.3%	90.5%	-8.2%
30 to 49 female	94.0%	93.4%	0.6%	84.4%	92.0%	-7.6%
50+ male	94.0%	92.8%	1.2%	87.7%	93.7%	-6.0%
50+ female	94.5%	93.5%	1.0%	88.4%	94.4%	-6.0%

The correct enumeration rate for the nation is 93.3% using Model 1, which is only 1.5% higher than the actual 2010 value. The correct enumeration rates from Model 1 are higher than the 2010 estimates for all age/sex groupings. The differences for the different age/sex groupings ranged from 0.6% to 3.0%.

The national match rate for Model 1 is 81.8%. That is substantially lower than the 2010 match rate of 91.1%. The national match rate in Model 1 is lower than the 2010 match rate primarily because the percent of matches is higher in the data after clerical matching and field followup and the percent of non-matches is lower after clerical matching and field followup (see 3.1.2.2). The match rates by age/sex from Model 1 are lower than the 2010 estimates for all age/sex groupings. As can been seen in Table 4, there was much more variability across the age/sex groups for the match rates than the correct enumeration rates as the differences ranged from -18.3% to -6.0% . 18-19 males and females had the biggest discrepancies for both rates, while 30-49 male and female and 50+ male and female tended to have lower differences for both rates.

As the correct enumeration rate is in the numerator and the match rate is in the denominator of the DSE, a higher correct enumeration rate or a lower match rate will lead to a higher DSE. Model 1 has both a high correct enumeration rate and a low match rate. Table 5 contains national and state level data for census counts, 2010 DSEs, the Model 1 DSEs and the differences between Model 1 and the 2010 DSEs in thousands. The states displayed in the table are those that performed the best and the worst based on the percentage difference calculated as follows:

$$\% \text{ Difference} = (\text{Model 1 DSE} - \text{2010 DSE}) / \text{2010 DSE}.$$

Table 5: DSE Highlights					
Category	2010 Census Counts (×1000)	2010 DSE (×1000)	Model 1 DSE (×1000)	Difference of Model 1 and 2010 DSE (×1000)	% Difference of Model 1 and 2010 DSE
National	300,703	300,667	343,900	43,233	14.4%
New Jersey	8,605	8,574	9,462	888	10.4%
Connecticut	3,456	3,440	3,813	373	10.8%
Alaska	629	624	739	115	18.5%
DC	562	575	693	118	20.6%

The national Model 1 DSE is greater than the 2010 DSE by approximately 43.2 million. This would imply that the 2010 Census did not count approximately 43.2 million people. This overestimation of the population is primarily due to the Model 1 national match rate being 9.3% lower than the 2010 national match rate. The percent difference between the Model 1 and the 2010 DSEs for the states ranged from 10.4% to 20.6%.

4.2 **Results Model 2**

Based on my analysis of the frequency tables and the results of Model 1, I made adjustments to the P-sample match codes for non-matches and possible matches and the residence status code for inmovers. The main changes to Model 1 were designed to increase the match rate to more closely align with 2010 estimates. The correct enumeration and match rates for Model 2 and 2010 for the nation and by age/sex are in Table 6. Table 6 also shows the differences between the Model 2 and the 2010 correct enumeration and match rates.

Table 6: Model 2 Correct Enumeration and Match Rates (Weighted)						
Category	Correct Enumeration Rate			Match Rate		
	Model 2	2010	Difference	Model 2	2010	Difference
National	92.5%	91.8%	0.7%	93.5%	91.1%	2.4%
0 to 4	91.4%	90.4%	1.0%	92.6%	88.3%	4.3%
5 to 9	92.8%	91.8%	0.9%	93.1%	90.5%	2.7%
10 to 17	93.4%	92.2%	1.2%	93.1%	91.5%	1.7%
18 to 29 male	88.2%	87.1%	1.1%	86.1%	84.8%	1.3%
18 to 29 female	89.4%	88.3%	1.1%	86.8%	86.4%	0.4%
30 to 49 male	92.5%	92.1%	0.4%	95.1%	90.5%	4.5%
30 to 49 female	93.4%	93.4%	0.1%	96.1%	92.0%	4.0%
50+ male	93.5%	92.8%	0.7%	95.5%	93.7%	1.8%
50+ female	94.1%	93.5%	0.6%	95.8%	94.4%	1.5%

The national match rate for Model 2 is greater than the 2010 match rate by 2.4%. The percent differences for the different age/sex groupings ranged from 0.4% to 4.5%. The changes made to Model 2 have reduced the differences between the Model 1 and 2010 match rates. In fact, the Model 2 match rates are higher than the 2010 rates.

The changes implemented have also lowered the correct enumeration rate from Model 2 compared with Model 1. This is due to identifying more inmovers. A match to an inmover is an erroneous enumeration because inmovers lived at a different address than they were matched to on Census Day. The national correct enumeration rate for Model 2 is 0.7% higher than the actual 2010 correct enumeration rate. For each age/sex category, the correct enumeration rates from Model 2 still exceed the 2010 correct enumeration rates. The differences for the different age/sex groupings ranged from 0.1% to 1.2%.

The DSE highlights for Model 2 are shown in Table 7. The national Model 2 DSE is less than the 2010 DSE by approximately 5.9 million. The percent differences between the Model 2 DSEs and the actual 2010 DSEs ranged from -5.3% to 0.3%.

Table 7: Model 2 DSE Highlights

Category	2010 Census Counts (×1000)	2010 DSE (×1000)	Model 2 DSE (×1000)	Difference of Model 2 and 2010 DSE (×1000)	% Difference of Model 2 and 2010 DSE
National	300,703	300,667	294,800	-5,867	-2.0%
Kansas	2,774	2,756	2,763	7	0.3%
North Dakota	648	648	650	2	0.3%
Texas	24,564	24,804	23,800	-1,004	-4.1%
DC	562	575	544	-30	-5.3%

The results of Model 2 are an improvement over the results of Model 1. However, the Model 2 DSE may appear to be more accurate than it really is. This is because the effect of having a high correct enumeration rate is mitigated by the effect of having a high match rate.

5 Conclusions and Future Research

I looked at estimating census coverage using computer matching data instead of after clerical matching and field followup data. Model 1 used no additional information to measure coverage, while Model 2 used a decision rule and recodes based on analyzing 2010 data before and after clerical matching and field followup.

I compared match rates, correct enumeration rates, and DSEs from two models to those from 2010. The Model 1 DSE exceeded the 2010 DSE by 43.2 million. By using an inmover decision rule and treating nonmatches as unresolved, I was able to produce a DSE within 5.9 million of the 2010 CCM DSE. However, the correct enumeration rate and match rate were both higher than the 2010 rates.

The DSEs calculated with computer matching data are not accurate enough to be used in production. The Model 2 DSE was approximately 5.9 million less than the 2010 DSE. That is roughly the population of Missouri in 2010, the 18th largest state by population that year.

Improving the computer matching algorithm, identifying additional decision rules and tailoring the imputation models to better work with computer matching data are potential areas of future research.

Acknowledgements

Thank you to Krista Heim, Scott Konicki, Mark Jost and Julianne Zamora for their comments and input on this document.

References

Hogan, H, (2003), "The Accuracy and Coverage Evaluation: Theory and Design", Survey Methodology, Vol. 29 No. 2, pp 129-138, Statistics Canada, Catalogue No. 12-001.

Konicki, S., et al. (2013), "2010 Census Coverage Measurement Estimation Methods", DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-J-03.

Linse, K. (2009), "2010 Census Coverage Measurement Person Interview Post Processing Attachment 1", DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-D7-06.

Mulry, M. and Cantwell, P. (2010). "Overview of the 2010 Census Coverage Measurement Program and Its Evaluations." CHANCE, 23:3, 46-51, DOI: 10.1080/09332480.2010.10739823.

Petersen, C.G.J. (1896), The Yearly Immigration of Young Plaice into the Limfjord from the German Sea. Report of the Danish Biological Station. 6, 1-48.

Viehdorfer, C. (2011), "The Design of the Coverage Measurement for the 2010 Census – REVISION #1", DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-B-07-R1.

Whitford, D. (2008), "Overview of 2010 Census Coverage Measurement Program", DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #A-19.

Whitford, D. (2010) "Specification for the 2010 Census Coverage Measurement Person Before Followup Clerical Matching", DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-D9-16. Attachment E-1.

Wolter, Kirk M. (1986), "Some Coverage Error Models for Census Data". Journal of the American Statistical Association, Vol. 81, No. 394, pp. 338- 346.