Recursive Partitioning for Racial Classification Cells

Aaron Gilary
Center for Statistical Research and Methodology,
U.S. Bureau of the Census, Washington DC, 20233
aaron.j.gilary@census.gov

Abstract[1]

From the time the Census Bureau introduced an option to identify with multiple races on its survey forms, researchers within the Census Bureau have sought the best way to aggregate the possible responses into categories while preserving the information from an increasingly multiracial country. Classifying racial data into categories helps provide information to Census stakeholders so they can measure the Census Bureau's performance in identifying and correctly enumerating each population. As planning intensified for the 2010 Census Coverage Measurement study, research staff analyzed the Matching and Correct Enumeration rates of multiracial populations, in order to model the data. The paper details the techniques used to build models for Census Coverage data, by applying stepwise regression to the concept of CART modeling to partition the data into cells, and adding information criteria as a method of cross-validation. The paper also discusses: the specific issues inherent in modeling Dual-System Estimation data for this topic, and how they were addressed; the patterns of racial identification that were discovered; and the recommendations that were ultimately proposed.

Key Words: Recursive Partitioning, Stepwise Regression, Multiracial, Modeling

1. Introduction: Multiracial Modeling for Census Coverage

After many decades of collecting respondent race as a single characteristic, the Census Bureau provided each respondent the option to identify with more than one race on the 2000 Census questionnaire. In order to link racial identification from the 2000 Census to the earlier classifications where each respondent could only identify with one race, a system of rules was developed to collapse multiple race responses into domains which approximated the traditional, single race categories (Farber 2001). The Census Coverage Measurement (CCM) Estimation Team reviewed these classification rules for 2010. The Census currently uses post-stratification to form cells for racial classification, but debated whether logistic regression should be used instead to formulate these cells and whether the current arrangement was the optimal one for a population increasingly reporting more than one race.

The Census asks about race using six categories: American Indian, Black, Native Hawaiian/Pacific Islander, Asian, White, and Some Other Race. Because each respondent can identify with one or more of these categories, as well as with Hispanic origin, the Census offers 126 different options for race/origin classification. Should the Census Bureau use these options to expand its classification structure for 2010 to provide more information about populations identifying with more than one race?

[1] This report is released to inform interested parties of research and to encourage discussion. Any views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

To help answer this question, I researched an alternative method for forming race/origin subgroups, using the 2000 data. Recursive partitioning methods were applied to the CCM data sets to find the best set of race domains, based upon a balance between a minimal number of classification errors, and homogeneity in the rates.

1.1 Dual-System Estimation, and the Partitioning Methods

The Census Coverage Measurement sample survey is a follow-up to the Census itself, conducted for an independent verification of the Bureau's performance capturing the nation's population. It is an area sample that contains two sources of data – the Post Enumeration Sample ("P-Sample") selected randomly from across the country as the verification sample, and the Enumeration Sample ("E-Sample") that comes from the corresponding Census records (Bell and Cohen 2009).

Census Coverage Measurement uses Dual-System Estimation as a means of evaluating the Census data. The standard dual-system estimator employed by Census Coverage holds that the estimate of the total population ($\hat{\mathbf{N}}_{++}$) can be derived from the Census and P-Sample totals as in the equation:

$$\hat{\mathbf{N}}_{++} = ((N_{+1} * N_{1+})/N_{11}),$$

Where

$\hat{\mathbf{N}}_{++}$ is the estimate of the total number of people;
N_{+1} is the number of people counted in the E-Sample;
N_{1+} is the number of people counted in the P-Sample;
N_{11} is the number of people counted in both the E-Sample and the P-Sample.

The counts in the formula are adjusted to eliminate duplicate, fictitious, or otherwise erroneous enumerations, and therefore reflect the number of *actual* people counted in each survey (Wright and Hogan 1999). This estimator is used for subgroups of the population (poststrata) as well as for the overall total.

The Dual-System Estimate assumes that the P-Sample and E-Sample are independent and that the ratios of survey totals in the above equation are synthetic at an aggregate level, meaning that higher level ratios can be applied to lower levels with no error other than classification error (Hogan 2003). This model estimates total population using the Post-stratification Assumption, which holds that producing estimates across any variable used for poststratification and then aggregating them will provide an estimate of the total (Wolter 1986).

The *Match Rate* (i.e., the proportion of persons in the P-Sample with a valid Census record) was used as the dependent variable for the P-Sample modeling, and the *Correct Enumeration (CE) Rate* (i.e., the proportion of Census records that are correct) was the dependent variable for the E-Sample modeling in this research study. The Match and CE rates are important because they measure the Census Bureau's success at population estimation, both overall and with respect to variables such as race, age, sex, and household tenure status (i.e., owners versus renters). The Census Bureau currently uses these covariates to separate the data into post-strata that are homogenous according to match and CE rates. There were 416 such post-strata in 2000 (Hogan 2003).

In order to expand the racial domain structure, I chose a recursive partitioning method to partition respondent records from both samples into new race/ethnicity cells because there was a desire that the cells be mutually exclusive and because such a method would not exclude higher order interactions between covariates. The goal was to create a model to partition the records into cells through the following steps:

-- Formulate a regression model using the match and CE rates as dependent variables, and Hispanic Origin and the six racial categories as potential covariates.
-- Add the most significant variable to the model using stepwise logistic regression (stopping the procedure after one step).
-- Use that variable to split the dataset into two groups in a branching structure.
-- Then use stepwise regression again to find the most significant remaining variable for each of *those* groups. This procedure is repeated until the regression model (or "tree") has been expanded as far as possible.
-- Finally, prune back the tree using a model selection criterion. The Bayesian Information Criterion (BIC) is used here, as it assesses a penalty for each parameter added to the model based on the log of the total number of records, and therefore favors more parsimonious models. This scaled back tree is the final model.

This procedure uses the basic concept of Classification and Regression Trees (CART) modeling, as discussed in Breiman et al. (1984). But unlike CART, which uses classification based on percentages, the procedure uses a likelihood-based selection. The likelihood mechanism used weighted observations in proportion to their sampling weight, adjusted to sum to the sample size. Any possibility that the model could be improperly influenced by a correlation between individual observations is ignored.

Initially, the project studied Hispanics and Non-Hispanics separately to address a question about the relationship between race and ethnicity identification for Hispanics. As a result, this analysis is divided into a Non-Hispanic and a Hispanic section. Hispanic ethnicity was coded as if it were a race, even though race and ethnicity are separate variables. This procedure does not study main effects beyond the first variable, but regression as classification is not in the scope of this work – the goal is to find classification cells to address multiracial populations.

Ultimately, the procedure created four trees (Match Rates and CE Rates for both Non-Hispanics and Hispanics). The procedures detailed in the next three sections use Non-Hispanic Match Rate results as a proxy for all four trees. The tree for that rate, identified from P-Sample data, is given as a flowchart on the next page.

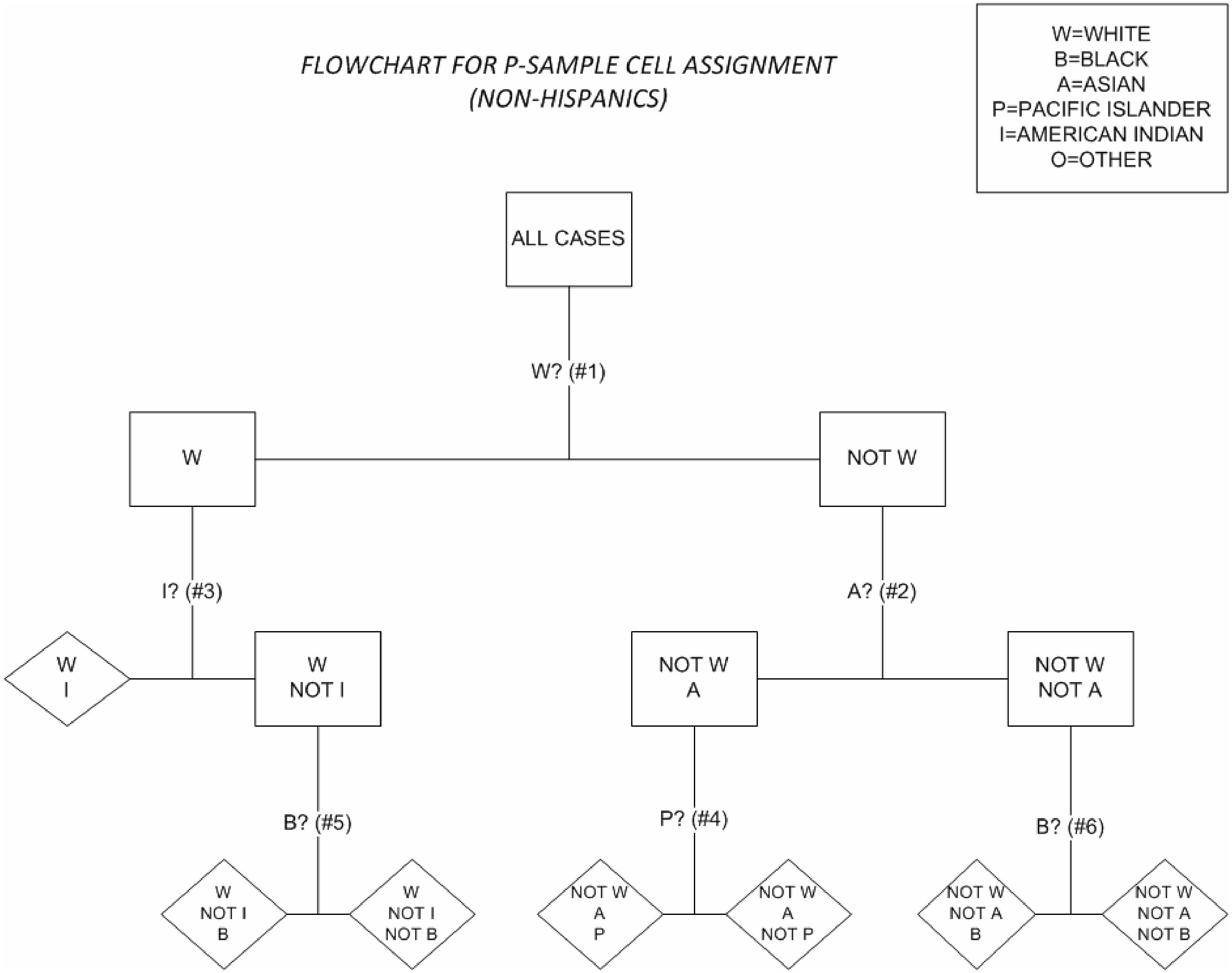
1.2: Creation of the Regression Tree

The flowchart illustrates the partitioning mechanism by which all the Non-Hispanic cases filter into one of the classification cells. It represents the tree after the BIC has been employed and the branches have been pruned. The process begins with all the cases in one bucket, at which point the model goes through all the covariates and picks out 'White' (W) as the most significant. The data is then separated into those records where 'White' was marked, and those where it was not. At the next step, the process repeats for those two distinct groups (White/Not White). The American Indian covariate (I) is found to be the most significant for the 'White' subgroup, but the Asian covariate (A) is found to be the most significant for the non-White subgroup. Then, each of the new subgroups is tested, and new covariates emerge. This process is extended for many stages, although it is ultimately pruned back through the use of a selection criterion to what is seen in the flowchart.

The final cells are represented by the diamond shape. The key in the upper right shows the letters used to denote each racial category. The numbers reflect the ranked statistical significance of each split: the first split, on the 'White' covariate, is the most significant.

All cases will fit into one of these cells based on their combination of racial responses. Racial combinations with similar match rates will be grouped together into the same cell. Some of the variables in this chart correspond to a higher match rate than the

FLOWCHART FOR P-SAMPLE CELL ASSIGNMENT
(NON-HISPANICS)



baseline (White, Asian); others correspond to a lower match rate (American Indian, Black, and Pacific Islander).

It is worth pointing out that the Match Rate tree is different from the Correct Enumeration tree for the same Non-Hispanic population and that match and CE rates can be poorly correlated for the different subgroups. Pacific Islander identifiers have a high CE rate and a very low match rate, while respondents who marked both White and Other have a low CE rate relative to match rate. White identifiers have higher match rates and CE rates than any other race.

1.3 Application of BIC to the model

After the framework for the tree was established, each variable addition to the model was evaluated using forward-selection BIC. The BIC used here selects a smaller model than Akaike's Information Criterion (AIC), due to its stricter penalty function. But while the criteria may vary, all of them will essentially measure the improvement in the model log-likelihood with an extra covariate added, and compare it to a penalty for that extra parameter.

1.4 Match Rate/Correct Enumeration Rate Cells

Similar procedures were carried out for Hispanic match rates and Hispanic and Non-Hispanic CE rates, resulting in four final trees. Appendix 1 shows the cells for those four trees. For the P-Sample, there are seven Non-Hispanic cells and two Hispanic cells. For the E-Sample, there are five Non-Hispanic cells and two Hispanic cells. The unweighted sample sizes, weighted sample sizes, and match rates are given. (Note that all totals and rates in this analysis use weighted data, and the total weights are in proportion to the sample size.)

The two trees are constructed differently, as the different racial variables do not have the same degree of importance for the two samples. They do share their first three Non-Hispanic nodes in common: White; Asian but not White; Neither White nor Asian.

I briefly attempted to combine the E-Sample and P-Sample trees into one unified tree, but this idea was abandoned because it involved creating too many splits of the data. For example, if one sample split Hispanic cases based upon White identification and the other split those cases upon Asian identification, the unified tree would split the cases based on both variables and the nodes would be too specific to be of use.

The Estimation team had somewhat mixed reaction to these findings. The procedure did create a group of heterogeneous cells that were informative and easy to interpret, but the team expressed concerns about inconsistency between 2000 and 2010, as well as how to address the effects of imputation on this data, if this procedure were to be implemented.

1.5 Conclusion

While the initial results provided a new framework for grouping the data, they were judged not to be entirely conclusive. They did not address whether this procedure could work in a localized way within the existing domains, or how consistent racial identification is between the two samples.

2. Challenges in Expanding Domains into Sub-Domains

After these results had been identified, the Estimation team and I decided on our next steps. First, I would apply the partitioning methods within each of the seven Census domains and split them further, so that we could develop a model while still retaining historic consistency with the previous Census. Meanwhile, the team wanted to investigate the stability of racial identification using “matched” records contained in both the Census and the P-Sample, to see how universal the models are across the two frames. The results of these two investigations, and the issues encountered, revealed the difficulties of this research project.

2.1 Cell Modeling Within the Domain Structure

The Census definitions of the domains are given in Appendix 2. These definitions come from a DSSD Memorandum (Haines 2001), but they are reorganized for this paper as an informal guide to each scenario.

There are two things to note here about this classification structure. First, the domain definitions affected the branching and kept the predictive models very small. For example, the Non-Hispanic Asian domain did not split at all, because there was nothing to split: an Asian respondent who marked any additional category in her response would have been automatically placed into a different domain. Second, there are geographical factors at play here in addition to race/ethnicity. Residence in Hawaii or on an American Indian Reservation contributed to the classification as well.

Although there were some similarities, the samples’ splits within the domains are based on different racial indicators. And the splits dictated by the recursive partitioning model, especially for the P-Sample, can be affected by either size or rate differential. In Domain 2, Off-Reservation American Indians were split by the ‘Asian’ covariate, based on a small population of Asian identifiers in that domain with a much lower match rate. In Domain 3, Hispanics were split by the ‘White’ covariate based on a very large population of White identifiers with a match rate that was only slightly smaller. Ultimately, these models are not very stable, and the splits being created do not lead to a clear-cut cell model.

2.2 Examining Consistency of Domain Classification

After the domain classification was finished, the project focused on measures of stability to gauge the consistency of racial classification for the two models. The study involves a different data set, comprised of the 578,300 P-Sample cases that matched to Census enumerations. The samples need to be linked to study consistency so that two race/origin responses are present on each record.

Appendix 3 shows the overall stability of the Census-defined domains by illustrating how E-Sample domain cases map back to the same P-Sample domain and vice versa. The data here is a little bit mixed. Over 96 percent of cases have consistent race domains for the two samples, but that number is somewhat inflated by the mostly White Domain 7; both Off-Reservation American Indians and Native Hawaiian/Pacific Islanders are under 76 percent for the two samples. This finding is consistent with the results in Farber (2001). It appears that the most stable ones are the most clear-cut, such as American Indians on a Reservation or Non-Hispanic Asians, but overall most domains are around the 90 percent level.

2.3 Explanations for Inconsistency

Data analysis of the inconsistent cases helped to uncover one source of the inconsistency problem. It was found that splitting the matched data into cases where the race/origin answer combinations do and do not match shows that 89 percent of respondents match their identification exactly between the E-Sample and the P-Sample. But the results of splitting that data into single-race and multi-race identifiers present a different picture:

- 88.4 percent of matched respondents marked one race on the Census, and matched that in the CCM follow-up;
- .55 percent of matched respondents marked multiple races on the Census, and matched those in the CCM follow-up;
- 9.3 percent of matched respondents marked one race on the Census, but changed at least one answer in the CCM follow-up;
- 1.7 percent of matched respondents marked multiple races on the Census, but changed at least one answer in the CCM follow-up.

To put it another way, the data shows that *less than 24 percent of respondent records self-identifying as more than one race on the Census later used those exact same categories to self-identify in the Census Coverage Measurement follow-up*. Even taking into account the aforementioned issue with 'Some Other Race' identifiers, this fact captures the problem in creating a single model – for those records that are multiracial identifiers, there is a limited effectiveness to these racial data. (Note: Hispanic Ethnicity does not count as a race here when assessing whether a Census respondent identifies with more than one race, but it does count as a category when assessing consistency versus inconsistency in the two samples.)

It should be noted that there are about 500 cases nationwide for each sample case in Census 2000, and that ratio is increasing for Census 2010 as the sample is being cut. The more weighting is used, the more the analysis is reliant on modeling assumptions. There was also an issue in the way Some Other Race cases were coded for Coverage data. The ACS Implementation Report explains that "detailed review of the edits used in Census 2000 led to the discovery of a difference on enumerator returns" (Griffin et al. 2004). As a result, the number of Census respondents identifying as Some Other Race was incorrectly inflated. But this error was limited in scope and should not affect broader conclusions.

2.5 Four Subgroups

Although the domain subgroups offered a new method for classifying the data, the inconsistency in race reporting between the samples largely undermined these findings. The Estimation team determined that the domains could not be reconciled, and the data was too inconsistent to rely on this type of modeling. However, there was still interest in examining the subgroups identified with the procedure to evaluate the efficacy of the domains as 2010 data arrived. The team and I proposed four race/origin subgroups to investigate more closely, based on unexpected results they had shown in the earlier research. Those groups were:

(A) 'Hispanic' and 'NHPI' identifiers in the HISPANIC domain (Domain 3). This group is possibly a specific ethnic population. All Hispanic identifiers will be put into Domain 3 unless there are geographic concerns, but the group's match and CE rates will be compared to Hispanic and Native Hawaiian/Pacific Islander (NHPI) baselines (from Domains 3 and 5, respectively).

(B) 'Black' and 'White' identifiers in the NON-HISPANIC BLACK domain (Domain 4). These cases are generally put into the Non-Hispanic Black domain. They are compared to the baselines from the Non-Hispanic White and Some Other Race domain (Domain 7).

(C) 'Asian' and 'NHPI' identifiers in the NATIVE HAWAIIAN/PACIFIC ISLANDER domain (Domain 5).If someone identifies with both of these races, they are

(D) *'Some Other Race' identifiers in the NON-HISPANIC WHITE OR "SOME OTHER RACE" domain (Domain 7).* The scheme for Domain 7 is fairly complex, but it basically includes 'White' identifiers and 'Some Other Race' identifiers (along with a few heterogeneous examples such as Black & Asian & NHPI). The 'Some Other Race' group is split from the rest of Domain 7, and studied separately.

The results are given in Appendix 4. The top table shows how the domains are actually arranged, including the weighted totals, correct enumeration rates and match rates (which are computed separately), and consistency rates. The consistency percentage given on the right hand side denotes the weighted percentage of records in the Census that were classified in the same domain in the P-Sample.

The bottom table shows the racial subgroups. The left column shows that A, B, C, and D have been removed from the existing domains and incorporated into new ones. The rates for these new domains are different from the initial domains in many instances, but the consistency percentages are much smaller. Each row is defined as a separate group for these charts, and consistency is defined as matching to this exact same row. Per this definition, adding more structure for each domain will decrease the consistency rate, as it would add more requirements for consistency.

On one hand, there is no real effect on the rates of the parent domains when the subpopulation is removed because of the large number of monoracial identifiers keeping the rates stable. (The Native Hawaiian/Pacific Islander domain is the lone exception, since it is much smaller than all the others which might be affected.) However, it should be noted that the subpopulations may have a rate substantially different from the parents. So if the group is of interest by itself, there is a reason to split it out systematically; but if not, splitting does not really matter to the larger domain. Splitting has different effects with each of the racial populations: some are larger than others, and some are more homogenous in their composition.

3. Conclusion/Recommendations

Self-reported race is a complex and thorny topic to study, and it requires a balance between quantitative conclusions and qualitative knowledge of the subject, while also taking into account historical definitions and peculiarities that exist within the topic.

The research project sought an empirically best approach to classifying different racial identifiers, but the dual criteria of forming cells with similar Match and CE rates produced different domains that were difficult to reconcile. An approach based on obtaining a single tree may be needed – perhaps based on a joint likelihood distribution – but this would require additional time for formulating the domains.

The scope of the project and the strength of its recommendations are also restricted by the nature of the Dual-System Estimation. There are limitations inherent in the data when examining racial consistency in reporting because only the matched sample is available; the racial consistency of reporting for the unmatched population is unknown. There is also a large amount of inconsistency between the E-Sample and P-Sample: often they will contain different records, or the same record may report race differently. A limitation of this study is that with a subjective measure such as race, there is no objectively correct answer against which to compare.

There were issues of classification variability, in terms of respondents interpreting the question of their racial ethnicity in different ways. This variability within the modeling may be due to the different modes of the survey: the Census uses a mail return questionnaire which is self-administered, while the P-Sample uses Computer Assisted Personal Interviewing (CAPI) in which an interviewer administers the questionnaire to respondents.

The groups created by the recursive partitioning models are homogenous for Correct Enumeration and Matching, but they are not very consistent, particularly when it comes to people who identify with multiple races. As a result, it might be difficult to expand the race domains as the group may have envisioned. The issue with poor racial consistency may be a matter of understanding what is being asked, or it may be a matter of weak identification with different groups. Either way, it was a concern to the Estimation Team that so many groups in this racial framework are difficult to identify.

Based on these results, the CCM Estimation Team and I recommend the examination of four groups for evaluation and sensitivity analysis as 2010 data arrives. Those groups are:

(A) Hispanic/NHPI identifiers in HISPANIC domain (3).
(B) Black/White identifiers in the NON-HISPANIC BLACK domain (4).
(C) Asian/NHPI identifiers in NATIVE HAWAIIAN/PACIFIC ISLANDER domain (5).
(D) "Some Other Race" identifiers in the NON-HISPANIC WHITE OR "SOME OTHER RACE" domain (7).

As the Census 2010 data does arrive, there is likely to be a noticeably larger population reporting more than one race, and researchers should be well equipped to measure it. Studying the data of multiracial populations doesn't provide any clear-cut answers, but it does illustrate how the national racial composition changes, and where any related research should focus.

Acknowledgements:

The author wishes to thank Don Malec of the National Center for Health Statistics (NCHS), and Tom Mule of the Census Bureau, for their supervision and guidance during this research and their assistance with this paper.

References:

Bell, Robert M., and Cohen, Michael L., eds. (2009). *Coverage Measurement in the 2010 Census*. Washington, DC: National Academies Press.

Breiman, L., Freidman, J. H., Olshan, R.A., and Stone, C. J. (1984).*Classification and Regression Trees.* Boca Raton: Chapman and Hall/CRC.

Farber, James (2001). *DSSD Census 2000 Procedures and Operations Memorandum Series B-10.*

Griffin, Deborah H., Broadwater, Joan K., Leslie, Theresa F., Love, Susan P., Obenski, Sally M., and Raglin, David A. (2004). *Meeting 21st Century Demographic Data Needs – Implementing the American Community Survey, Report 4: Comparing General Demographic and Housing Characteristics with Census 2000*.

Haines, Dawn (2001). *DSSD Census 2000 Procedures and Operations Memorandum Series #Q-48.*

Hogan, Howard (2003). *The Accuracy and Coverage Evaluation: Theory and Design*. Statistics Canada: Volume 29, Number 2, pp.129-138.

Wolter, Kirk M. (1986). *Some Coverage Error Models for Census Data*. Journal of the American Statistical Association: Volume 81, Number 394, pp.338-346.

Wright, Tommy, and Hogan, Howard (1999). *Census 2000: Evolution of the Revised Plan*. Chance: Volume 12, Number 4, pp.11-19.

Appendices:

Appendix 1(a): E-Sample Correct Enumeration Cells (Non-Hispanics)

CELL	MARKED	NOT MARKED	N	WGTD_N	RATE
1	WO	--	2,115	1,901	0.898
2	W	O	461,360	507,147	0.942
3	A	W	31,200	27,129	0.927
4	I	WA	18,688	4,937	0.919
5	--	WAI	98,185	83,719	0.904
TOTAL NON-HISPANIC			*611,548*	*624,833*	*0.936*

Appendix 1(b): E-Sample Correct Enumeration Cells (Hispanics)

CELL	MARKED	NOT MARKED	N	WGTD_N	RATE
1	P	--	503	291	0.840
2	--	P	100,849	87,776	0.926
TOTAL HISPANIC			*101,352*	*88,067*	*0.926*

TOTAL E-SAMPLE			***712,900***	***712,900***	***0.935***

Appendix 1(c): P-Sample Match Rate Cells (Non-Hispanics)

CELL	MARKED	NOT MARKED	N	WGTD_N	RATE
1	WI	--	5,033	4,587	0.916
2	WB	I	1,931	1,908	0.908
3	W	IB	408,903	447,600	0.936
4	AP	W	503	232	0.797
5	A	WP	25,937	22,888	0.906
6	B	WA	82,244	70,967	0.873
7	--	WAB	25,882	12,794	0.885
TOTAL NON-HISPANIC			*550,433*	*560,976*	*0.926*

Appendix 1(d): P-Sample Match Rate Cells (Hispanics)

CELL	MARKED	NOT MARKED	N	WGTD_N	RATE
1	W	--	33,225	31,613	0.887
2	--	W	56,919	47,989	0.874
TOTAL HISPANIC			*90,144*	*79,602*	*0.879*
TOTAL P-SAMPLE			***640,577***	***640,577***	***0.920***

W=WHITE
B=BLACK
A=ASIAN
P=PACIFIC ISLANDER
I=AMERICAN INDIAN
O=OTHER

Appendix 2: Summary of Census Domain Conditions

DOMAIN 1: AMERICAN INDIAN OR ALASKA NATIVE ON RESERVATIONS
*Marked AMERICAN INDIAN/ALASKA NATIVE
*Lives on a reservation

DOMAIN 2: OFF-RESERVATION AMERICAN INDIAN OR ALASKA NATIVE
*Marked AMERICAN INDIAN/ALASKA NATIVE
*Does not live on a reservation
*Either lives in Indian Country, OR did not mark any of these: HISPANIC, BLACK, ASIAN, WHITE, or OTHER

DOMAIN 3: HISPANIC
*Marked HISPANIC
*Does not live in Indian Country, OR did not mark AMERICAN INDIAN/ALASKA NATIVE
*Does not live in Hawaii, OR did not mark NATIVE HAWAIIAN/PACIFIC ISLANDER

DOMAIN 4: NON-HISPANIC BLACK
*Marked BLACK
*Does not live in Indian Country, OR did not mark AMERICAN INDIAN/ALASKA NATIVE
*Did not mark HISPANIC
*Does not live in Hawaii, OR did not mark NATIVE HAWAIIAN/PACIFIC ISLANDER
*Marked no more than 1 of the following races: NATIVE HAWAIIAN/PACIFIC ISLANDER, ASIAN, WHITE, or OTHER

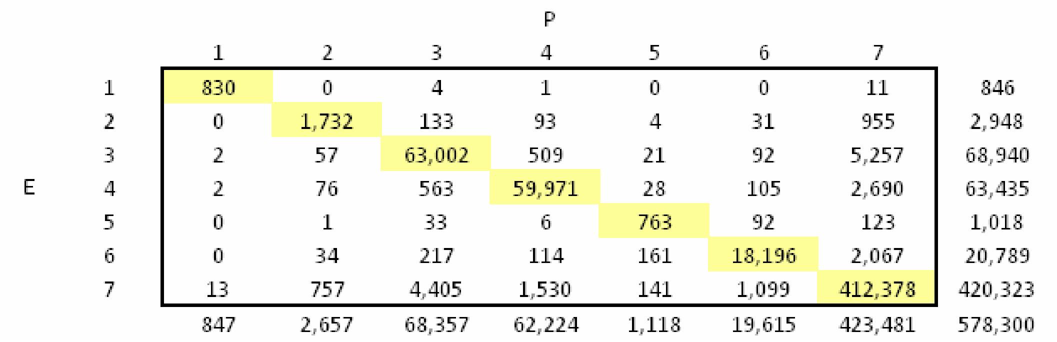
DOMAIN 5: NATIVE HAWAIIAN OR PACIFIC ISLANDER
*Marked NATIVE HAWAIIAN/PACIFIC ISLANDER
*Does not live in Indian Country, OR did not mark AMERICAN INDIAN/ALASKA NATIVE
*Either lives in Hawaii, OR did not mark any of these: HISPANIC, BLACK, WHITE, or OTHER

DOMAIN 6: NON-HISPANIC ASIAN
*Marked ASIAN
*Does not live in Indian Country, OR did not mark AMERICAN INDIAN/ALASKA NATIVE
*Did not mark any of these: HISPANIC, BLACK, NATIVE HAWAIIAN/PACIFIC ISLANDER, WHITE, or OTHER

DOMAIN 7: NON-HISPANIC WHITE OR “SOME OTHER RACE”
*Either marked 3 or more races, OR marked either WHITE or OTHER and did not mark BLACK.
*Does not live in Indian Country, OR did not mark AMERICAN INDIAN/ALASKA NATIVE
*Did not mark HISPANIC
*Does not live in Hawaii, OR did not mark NATIVE HAWAIIAN/PACIFIC ISLANDER

Appendix 3: General Consistency of Race/Origin Domains

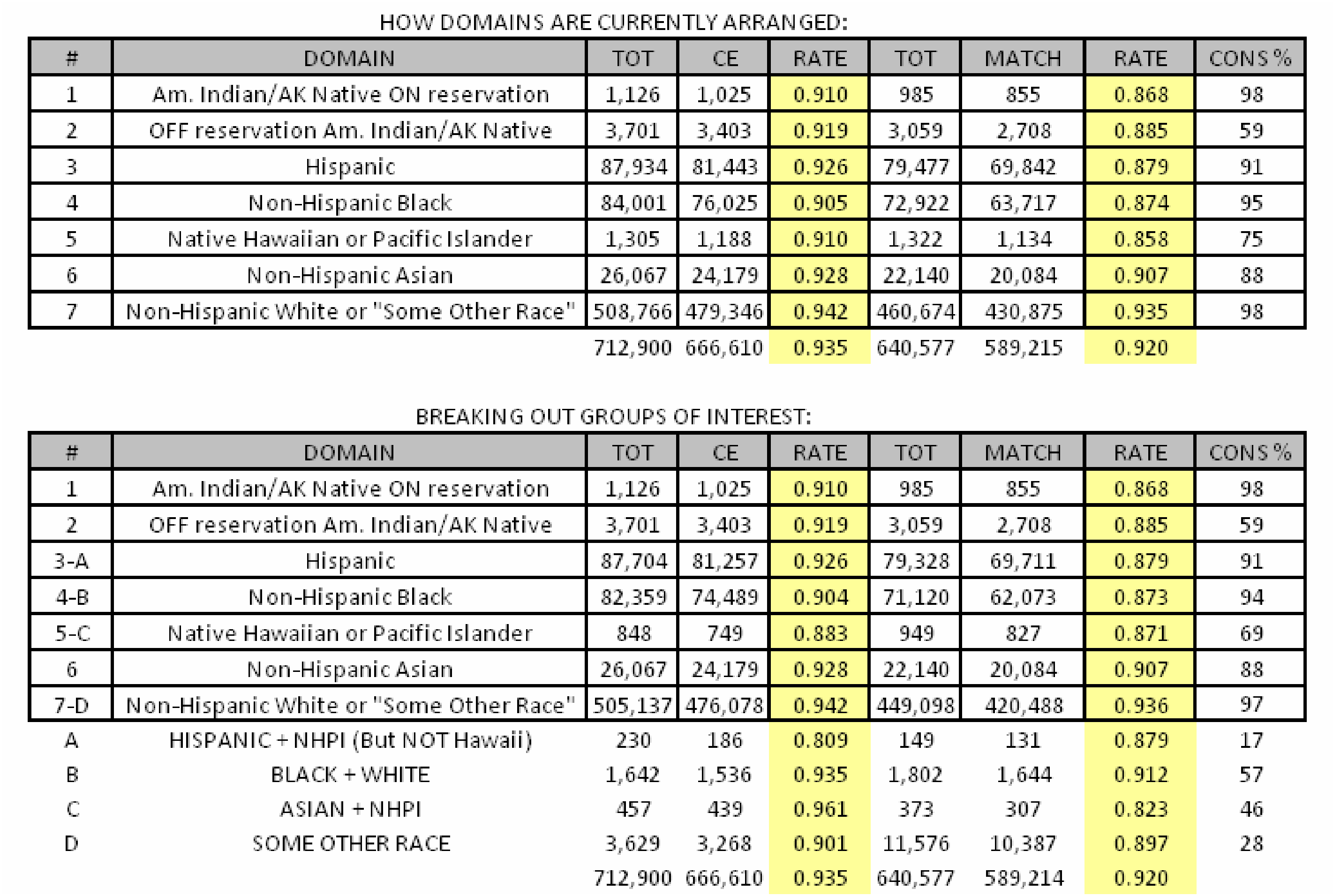
Domains for E-Sample and P-Sample Matched Cases (Weighted)



					P				
		1	2	3	4	5	6	7	
	1	830	0	4	1	0	0	11	846
	2	0	1,732	133	93	4	31	955	2,948
	3	2	57	63,002	509	21	92	5,257	68,940
E	4	2	76	563	59,971	28	105	2,690	63,435
	5	0	1	33	6	763	92	123	1,018
	6	0	34	217	114	161	18,196	2,067	20,789
	7	13	757	4,405	1,530	141	1,099	412,378	420,323
		847	2,657	68,357	62,224	1,118	19,615	423,481	578,300

(557,278 weighted cases -- 96.4% -- have consistent race domains)

Appendix 4: General Consistency of Race/Origin Domains



HOW DOMAINS ARE CURRENTLY ARRANGED:

#	DOMAIN	TOT	CE	RATE	TOT	MATCH	RATE	CONS %
1	Am. Indian/AK Native ON reservation	1,126	1,025	0.910	985	855	0.868	98
2	OFF reservation Am. Indian/AK Native	3,701	3,403	0.919	3,059	2,708	0.885	59
3	Hispanic	87,934	81,443	0.926	79,477	69,842	0.879	91
4	Non-Hispanic Black	84,001	76,025	0.905	72,922	63,717	0.874	95
5	Native Hawaiian or Pacific Islander	1,305	1,188	0.910	1,322	1,134	0.858	75
6	Non-Hispanic Asian	26,067	24,179	0.928	22,140	20,084	0.907	88
7	Non-Hispanic White or "Some Other Race"	508,766	479,346	0.942	460,674	430,875	0.935	98
		712,900	666,610	0.935	640,577	589,215	0.920	

BREAKING OUT GROUPS OF INTEREST:

#	DOMAIN	TOT	CE	RATE	TOT	MATCH	RATE	CONS %
1	Am. Indian/AK Native ON reservation	1,126	1,025	0.910	985	855	0.868	98
2	OFF reservation Am. Indian/AK Native	3,701	3,403	0.919	3,059	2,708	0.885	59
3-A	Hispanic	87,704	81,257	0.926	79,328	69,711	0.879	91
4-B	Non-Hispanic Black	82,359	74,489	0.904	71,120	62,073	0.873	94
5-C	Native Hawaiian or Pacific Islander	848	749	0.883	949	827	0.871	69
6	Non-Hispanic Asian	26,067	24,179	0.928	22,140	20,084	0.907	88
7-D	Non-Hispanic White or "Some Other Race"	505,137	476,078	0.942	449,098	420,488	0.936	97
A	HISPANIC + NHPI (But NOT Hawaii)	230	186	0.809	149	131	0.879	17
B	BLACK + WHITE	1,642	1,536	0.935	1,802	1,644	0.912	57
C	ASIAN + NHPI	457	439	0.961	373	307	0.823	46
D	SOME OTHER RACE	3,629	3,268	0.901	11,576	10,387	0.897	28
		712,900	666,610	0.935	640,577	589,214	0.920	