**UNITED STATES DEPARTMENT OF COMMERCE**
**Economics and Statistics Administration**
**U.S. Census Bureau**
Washington, DC 20233-0001

October 30, 2008

DSSD 2010 CENSUS COVERAGE MEASUREMENT MEMORANDUM SERIES #2010-E-19

MEMORANDUM FOR Donna L. Kostanich, Co-Chair
Census Coverage Measurement Operational Integration Team

Magdalena Ramos, Co-Chair
Census Coverage Measurement Operational Integration Team

From: Richard A. Griffin *(signed)*
Chair, Census Coverage Measurement Estimation Subgroup

Prepared by: Douglas Olson
Decennial Statistical Studies Division

Subject: Correlation Bias Adjustment by Individual Year of Age

The attached document on correlation bias adjustment by individual year of age is being provided as background material for the Census Coverage Measurement Estimation Workshop to be held in January 2009.

Attachment

cc:
DSSD CCM Contacts List

# Correlation Bias Adjustment by Individual Year of Age


## Summary

The Census Bureau is expecting to adjust its Census Coverage Measurement (CCM) Dual System Estimates (DSEs) for correlation bias using sex ratios from Demographic Analysis (DA). The form of the calculation will be similar to the adjustment made for Accuracy and Coverage Evaluation (A.C.E.) Revision II, except the methodology will need to be adapted for our plans to estimate DSEs for individual ages. Demographic Analysis has population estimates for each individual age 0-84 and for 85+, by sex and by Black/non-Black. To create a sex-ratio adjustment, we will estimate both a DA and a CCM sex ratio for each adult age for both Blacks and non-Blacks, to create an adjustment factor for males of each age and race. That factor will be applied to the DSE of the male population at each age.


## The Issue of Correlation Bias

Dual system estimates are said to contain bias if they systematically underestimate or overestimate the true population. Biases in the sample estimates of the components of the DSE formula can lead to biases in the DSEs, as discussed in Mulry (1991) and Mulry and Spencer (1991,1993). Even in the absence of any of these biases, DSEs can still be subject to another form of bias called correlation bias, resulting from failure of a general independence assumption that underlies the DSEs. This independence assumption can fail due to either:

- causal dependence – the act of being included in the Census makes someone more likely or less likely to be included in the CCM, or

- heterogeneity – Census and CCM inclusion probabilities vary over persons with identical estimated Census and CCM inclusion rates

When heterogeneity exists it is generally suspected to be of the form where persons more likely to be missed in the Census are also more likely to be missed in the CCM. Correlation bias resulting from this form of heterogeneity is negative, reflecting systematic underestimation of true population by the DSEs. Reducing bias due to heterogeneity is the principal reason that DSEs have traditionally been calculated using post-stratification or, in 2010, modeling. The adjustment using DA sex ratios attempts to reduce the effect of heterogeneity that remains after post-stratification or modeling adjustment.

### Evidence for Correlation Bias

The Census Bureau has traditionally estimated correlation bias by examining the sex ratios (ratio of males to females) by age and race, the latter broken down only by Black vs. non-Black, which is the level of detail available through traditional DA. In the 2000 A.C.E. Revision II and 1990 PES estimates, those sex ratios suggested the existence of correlation bias. The ratio of males to

females in almost all the breakdowns suggests that too few males had been estimated by the coverage measurements (A.C.E. and PES) compared to the estimates of DA, consistent with the theory of correlation bias. (Shores 2002)

Sex Ratios from DA and A.C.E. Revision II* – 2000

| Age | Black A.C.E. Rev II | Black DA | Non-Black A.C.E. Rev II | Non-Black DA |
|---|---|---|---|---|
| 18-29 | .83 | .90 | 1.05 | 1.04 |
| 30-49 | .81 | .89 | .99 | 1.01 |
| 50+ | .72 | .76 | .85 | .86 |

* Ratio before correlation bias adjustment applied

Sex Ratios from DA and the PES – 1990

| Age | Black PES | Black DA | Non-Black PES | Non-Black DA |
|---|---|---|---|---|
| 18-29 | .83 | .90 | 1.02 | 1.02 |
| 30-49 | .84 | .91 | .99 | 1.01 |
| 50+ | .72 | .78 | .81 | .82 |

**Demographic Analysis**

Demographic Analysis represents a macro-level approach for measuring coverage. Estimates of net undercount are obtained by comparing census counts to independent estimates of the population derived from other measures (mostly administrative data). In general, DA population estimates are developed for the census date by combining various types of demographic data that are independent of the census and are highly reliable, such as administrative statistics on births, deaths, and Medicare data and estimates of immigration and emigration. (Robinson 2004)

For people born since 1935, DA estimates for the population are based on the compilation of historical estimates of the components of population change: births since 1935 (B), deaths to persons born since 1935 (D), immigrants born since 1935 (I), and emigrants born since 1935 (E). Presuming that the components are accurately measured, the population estimates are derived by the basic demographic accounting equation applied to each birth cohort:

$$P = B - D + I - E$$

For persons born before 1935, administrative data on aggregate Medicare enrollments are used to estimate the population:

$$P = M + m$$

Here M is the aggregate Medicare enrollment and m is the estimate of underenrollment in Medicare. The DA estimate of the population born before 1935 is based on Medicare enrollments. Medicare is an administrative data set from the Health Care Financing Administration.

Reliability of DA Estimates

Demographic Analysis estimates are subject to certain errors and limitations. Errors in the administrative data used for DA, particularly uncertainty about the level of unauthorized immigration to the U.S. and uncertainty about the level of emigration from the U.S., lead to errors in the DA estimates. For this reason, DA population estimates (DA totals) are thought to be relatively less accurate than DA *sex ratios* (number of males over number of females). This reflects an assumption that errors in migration estimates are not grossly different for males than for females.

Demographic Analysis estimates are also limited by a lack of detail. Difficulties in using administrative data to construct estimates of subnational migration mean that subnational DA estimates, while providing useful indicators, are of significantly lesser accuracy than DA national estimates. Also, limited racial detail in the administrative data sources, along with differences in racial classification from the census, limits separate DA estimates by race to simply Black and non-Black. This limitation was somewhat more pronounced beginning in 2000 than it had been in 1990 and previously because the allowance of multiple race responses to Census 2000 creates some uncertainty about appropriate definitions of the Black and non-Black groups for comparability of DA and CCM results. This limitation has small but appreciable effects on DA totals, but effects on DA sex ratios are believed negligible. (U.S. Census Bureau 2001)

Assumptions of DA Sex Ratio Adjustment

There are obviously some basic assumptions about DA that are implicit in the decision to use it. We see no reason to doubt the accuracy of those tabulations within the limits imposed by their methodology. The birth and death components are considered reliable (Robinson 2004). The migration component, especially as regards undocumented immigration, is more suspect and tabulations from Census 2000 motivated a change in the way DA estimates this group. Still, we are willing to assume that DA will provide usable data.

The most basic assumption implicit in the correlation bias adjustment is that females of all ages and both race groups have less bias than males. When the estimated sex ratio of 18-29 year old

non-Black males to females in the Census 2000 A.C.E. Revision II was higher than the ratio in DA, the Census Bureau did not apply the adjustment, which would have lowered the estimated number of males. Although the calculation methodology implicitly assumes that females show zero correlation bias, which is probably not correct, a DA sex-ratio adjustment should at least represent an improvement versus not making any adjustment.

Methodologically, adjusting for DA requires that the CCM be able to create population groups consistent with the data available in DA. Since DA produces population estimates by sex and by Black/non-Black for every individual year of age up to 85, the CCM will need to duplicate those groups. The mathematical adjustments necessary to do so are described in the next section, such as using a consistent definition of "Black." Implementation of single year adjustment factors is motivated by the assumption that CCM will create population estimates by individual years of age. If this decision is reversed, and age groups are used instead, sex ratio adjustment factors will be calculated within the age groups so defined.

**Implementation Requirements**

First, the DA and CCM counts have to be adjusted for definitional differences:

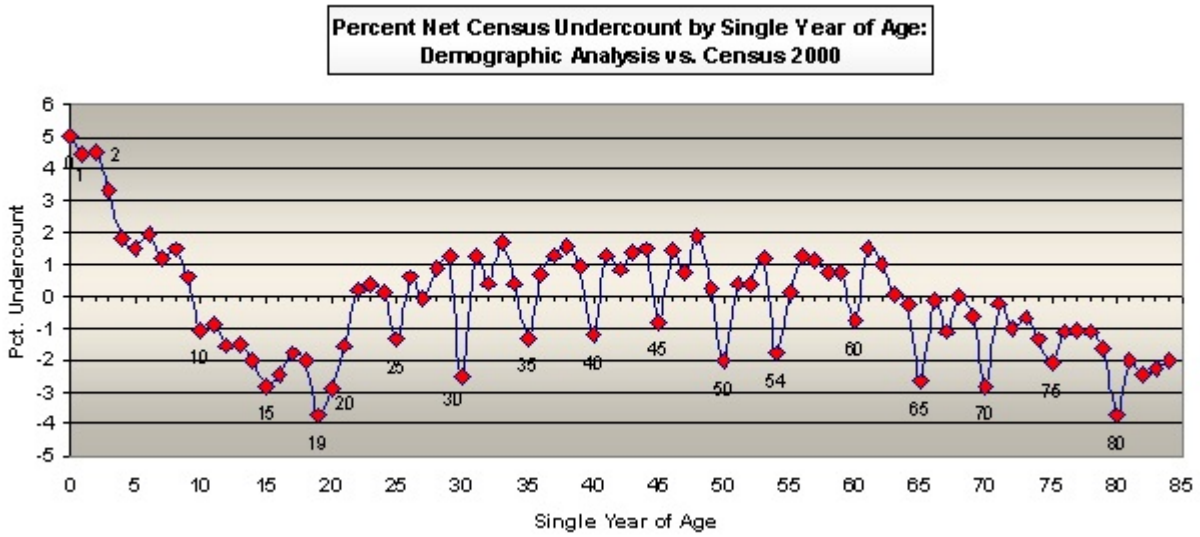
- A definition of Black in the CCM has to be constructed. Although Black is one choice of race variable in the Census, some persons are Black in combination with one or more other races. For purposes of sex ratio adjustment, we intend to include every person who is Black either alone or in combination with other races in the Black group.

- Since CCM only covers the household population, the census count of the group quarters (GQ) population has to be subtracted from each DA estimate group. Notice that in subtracting the census GQ counts from the DA totals we are ignoring possible under-coverage or overcoverage of the GQ population in the census. While it might be desirable to allow for census coverage errors of the GQ population in these calculations, separate estimates of census GQ population coverage are not available from either DA or CCM.

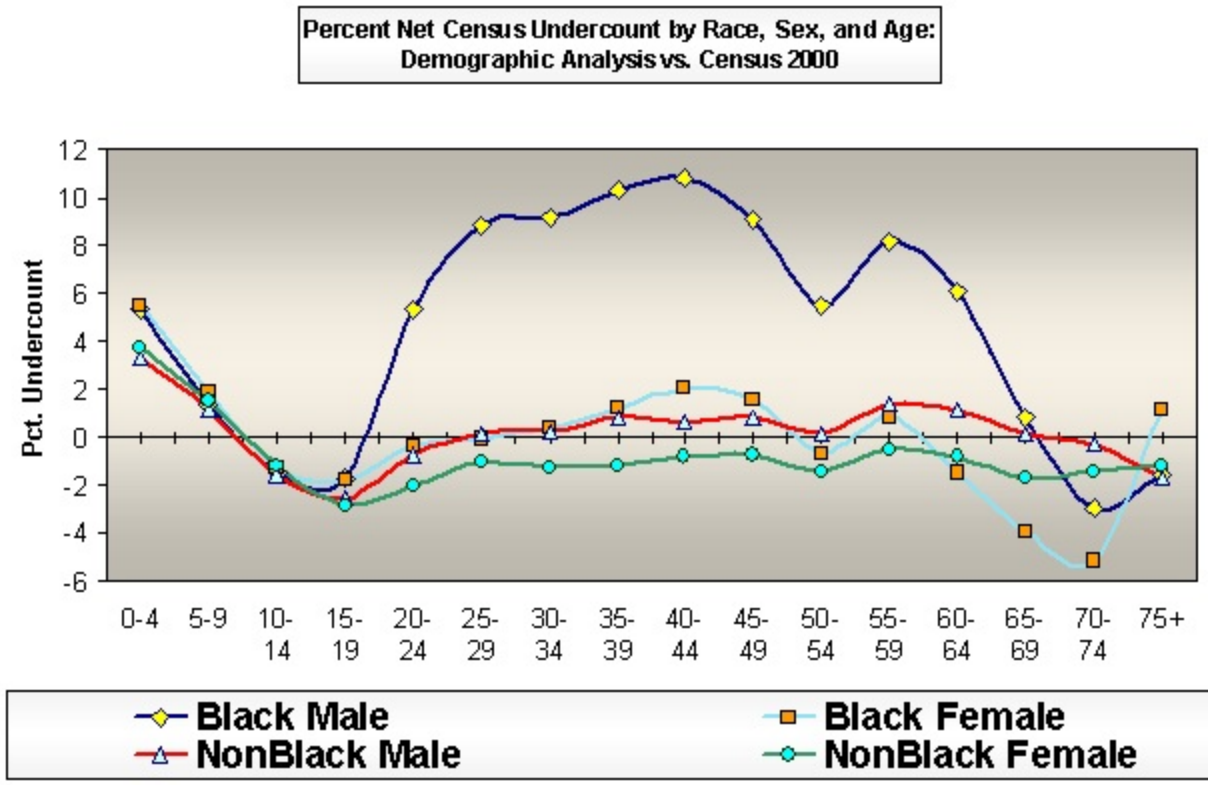
- The population in Remote Alaska (which is excluded from CCM) should be subtracted from the DA counts. This population is small relative to other population groups so its role in sex ratio estimation could be assumed minimal, and was not performed in Census 2000 A.C.E. Rev. II correlation bias adjustment, but may be used in 2010 CCM.

<u>Experience with DA Sex Ratio Adjustment</u>

Demographic Analysis methods have been used to estimate Census undercount since at least 1940 (Robinson 2000). In 2000 DA suggested significant Census undercount for the population under 10 and overcount of the population over 65. The estimates of undercount were also significantly different for the different sex/race groups, which motivated the use of correlation bias adjustment (both graphs from Bruce 2002):




The pattern of estimated undercounts for most ages between 25 and 65, except those divisible by five, is believed to be a manifestation of "age heaping," the tendency to report an age divisible by five by a respondent who does not know the census person well, such as a neighbor (West 2005). This could have important implications when applied to population estimates for individual ages.



Correlation bias adjustment was applied to the ACE Revision II estimates for Black males 18+ and non-Black males 30+. Non-Black males 18-29 were not adjusted because their adjustment factor was deemed unreliable. (Shores 2002) No adjustment was made for children.

Correlation Bias Adjustment for A.C.E. Revision II

| Race/Age Categories | DSE Totals Without Corr Bias Adj | DSE Totals With Corr Bias Adj | Difference Between DSE Totals | Percent Increase |
|---|---|---|---|---|
| Black males | | | | |
| 18 - 29 | 2,582,683 | 2,790,833 | 208,150 | 8.1 |
| 30 - 49 | 4,385,308 | 4,832,212 | 446,904 | 10.2 |
| 50+ | 2,746,088 | 2,895,811 | 149,723 | 5.5 |
| Non-Black males | | | | |
| 18 - 29 | 19,047,719 | 19,047,719 | 0 | 0 |
| 30 - 49 | 36,735,919 | 37,312,620 | 576,701 | 1.6 |
| 50+ | 30,311,508 | 30,632,206 | 320,698 | 1.1 |

Adjusting for Individual Ages

The 2010 CCM is intending to estimate all ages individually, instead of forming our traditional population groups (18-29, 30-49 and 50+) for three main reasons:

- The use of logistic regression modeling makes the use of continuous covariates possible that was not under post-stratification.

- There are detectable differences in coverage within some population groups, most noticeably between the younger and older members of the 18-29 group.

- The use of post-stratified groups creates artificial jumps in coverage adjustment at the borderline ages.

Since sex ratios vary by age, a population group defined by age cut-offs would impose the artificial jumps at the discontinuity points. Since the smoothness of the age graph before correlation bias adjustment is worth maintaining after adjustment, adjustment factors need to be constructed to maintain the smooth line.

Additionally, a maximum age value (85 is currently assumed) is necessary as sex ratios above that number tend to become extreme, and are hampered by small population size due to natural population attrition, and by inaccuracies cause by unrecorded deaths.

**CCM 2010 Proposed Methodology**

Regardless of which form of correlation bias adjustment is employed, it starts by calculating a Preliminary DSE for each person in the census, using the three predictions from the applicable modeling results:

$$PREDSE_j = \pi_{dd(j)} \times \frac{\pi_{ce(j)}}{\pi_{m(j)}}$$

The next step would be to calculate a correlation bias adjustment factor $C_k$ from the Preliminary DSE above, the DA population estimate and the Census Group Quarters population for each group k of interest:

$$c_k = \frac{\sum_{j \in Female\,k} PREDSE_j}{\sum_{j \in Male\,k} PREDSE_j} \div \frac{DA_{Female\,k} - GQ_{Female\,k}}{DA_{Male\,k} - GQ_{Male\,k}}$$

Where:
$C_k$ is the correlation bias adjustment for males in the $k^{th}$ race domain/age group,
$PREDSE_j$ is the preliminary DSE prior to correlation bias adjustment,
$DA_{sex,k}$ is the Demographic Analysis population estimate for group k, split by males and females,
$GQ_{sex,k}$ is the Census Group Quarters population for group k, split into males and females

Then, for each census person j, their final DSE is the product of the Preliminary DSE, multiplied by the adjustment factor above, where applicable:

$$DSE_j = PREDSE_j \times CB_j$$

The correlation bias adjustment factor (CBj) for each census person record is equal to one of the following:

$C_k$ if the census case is a male in the $k^{th}$ race/age group or
1 for all other census cases.

The above could be performed for single years of age, by treating the groups k as being the cross of Black or non-Black with each individual age-year. Concerns about heaping for adult ages divisible by five, as well as irregularity caused by minor errors in age reporting, motivates the desire to smooth factors through the use of a moving average. In the $C_k$ calculation above, all the population counts and estimates (PREDSE, DA and GQ) will be calculated from a five-year moving average, centered around age k in most cases:

$$PREDSE_k = \sum_{k-2}^{k+2} PREDSE_i$$

And analogously for $DA_k$ and $GQ_k$.

To reduce concerns about including ages outside the range of ages (0-85), the moving average near the limits will double-count certain ages to maintain the operative spectrum:

| Age | Use ages (count double if repeated): |
|---|---|
| 0 | 0, 0, 1, 1, 2 |
| 1 | 0, 1, 1, 2, 3 |
| 2 | 0, 1, 2, 3, 4 |
| 3 | 1, 2, 3, 4, 5 |
| ... | ... |
| 81 | 79, 80, 81, 82 ,83 |
| 82 | 80, 81, 82, 83, 84 |
| 83 | 81, 82, 83, 84, 85 |
| 84 | 82, 83, 84, 85, 85 |
| 85 | 83, 84, 84, 85, 85 |

In this way, each age k=[0, 85] is calculated from five donors (including repetitions) as well as donating to the calculation of five ages k, again including repetitions.

<u>An Example</u>

To study the likely effect of correlation bias adjustment on population estimates, a simplified model was run using the 2000 A.C.E. Revision II research data file. This simplified model used independent variables from that data set (described in U.S. Census Bureau 2004):

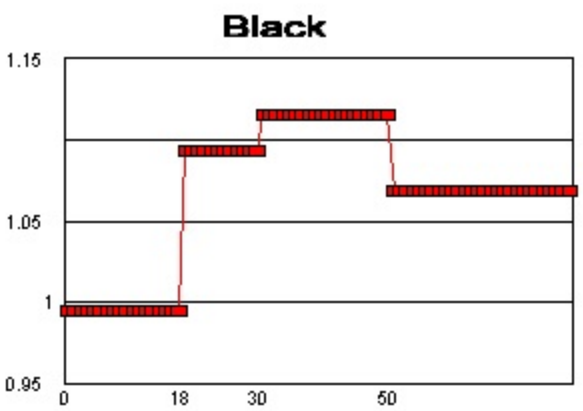
- Race/Ethnicity Domains
- Tenure
- Sex
- Region
- Age, modeled as splines in six ranges: 0-17, 18-20, 21-49 and 50+

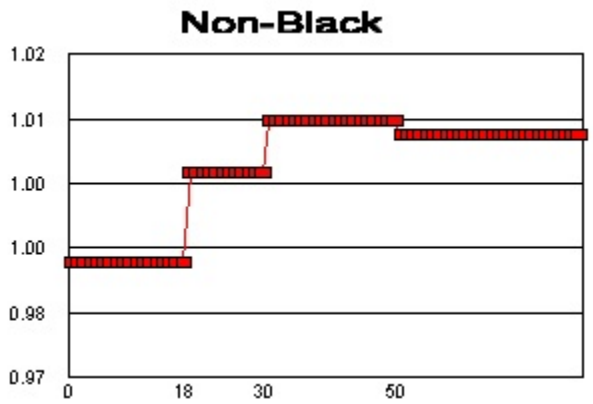
Because this approach used data from the 2000 A.C.E., it was not targeted toward duplicating every feature planned for the 2010 CCM. Among the simplification employed were:

- age was topcoded to 80 for modeling purposes
- Black was defined as the black domain used for poststratification

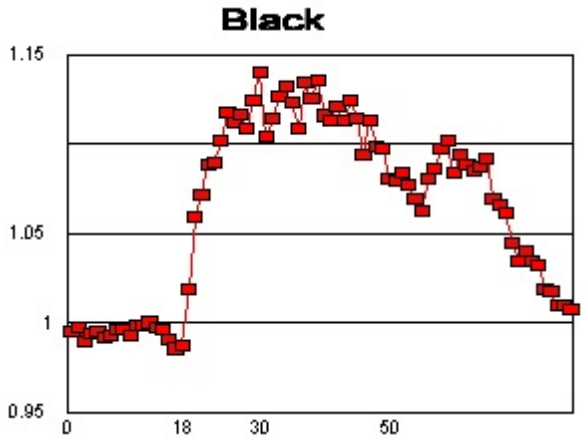
Since these are potentially important differences versus true population counts, this paper will not include actual population totals used under these simplified assumptions.

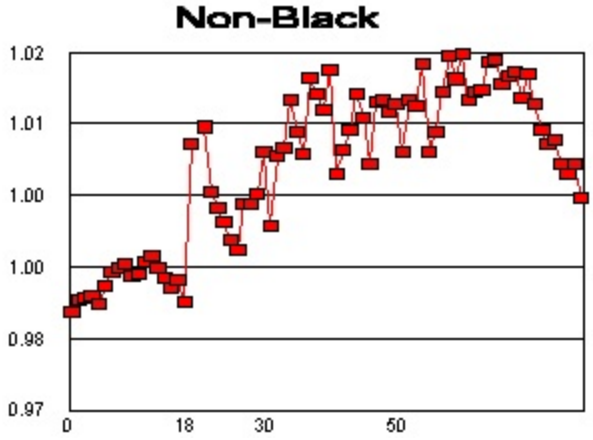
The age-group methodology used up until 2000 would produce correlation bias adjustment factors ($C_k$ from above) for males with jump-discontinuities at the border ages (0-17, 18-29, 30-49, 50+):



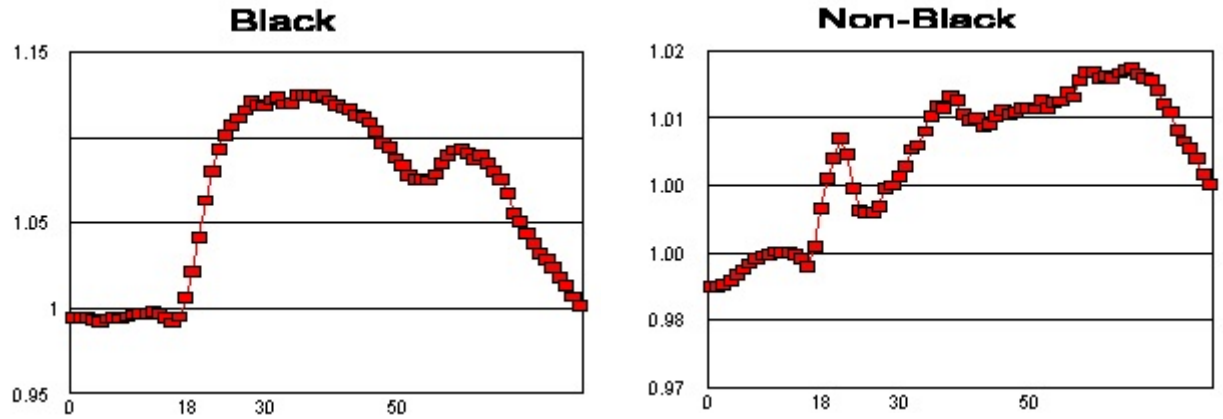




Application of the individual years model captures some important differences within the age groups, but shows noise and age-heaping at some years divisible by five:



The smoothed line produces the desired adjustment:




**Future Work**

The sections above outline the current plans for implementing a correlation bias adjustment for 2010 CCM. Other adjustment methods are possible, and further developments in DA allow the possibility of research for future adjustments, although it is not known at this time if they will be available or possible to implement for use in 2010 CCM.

The above method is known as the "two group model" and is based on the assumption of consistent bias within each of the two groups (male and female), the ratio of which would be applied consistently to each member of the male group. Other models based on other assumptions are possible, and are listed in Bell (2001). Those could be studied during the 2010 post-production research cycle for comparison.

Some research underway or planned in DA include (Judson 2008):

- expanding coverage of Hispanics as a distinct population group
- using other sources of administrative records to better capture sub-national estimates and net migration flows
- developing measures of uncertainty

Whether the CCM investigates any potentially enhanced DA estimates will depend on timing and available resources.

**References**

Bell, William. 2001. "Accuracy and Coverage Evaluation: Correlation Bias." DSSD Census 2000 Procedures and Operations Memorandum Series B-12. U.S. Bureau of the Census.

Bruce, Antonio, Arjun Adlakha, Peter Johnson, and J. Gregory Robinson. 2002. "Historical Profile of Demographic Analysis and Population Estimates: Components across Time (1935 to 2000)," Population Association of America, Atlanta, Georgia, May 9-11, 2002

Judson, Dean, J. Gregory Robinson, Antonio Bruce, Kirsten West; 2008; "Proposals for a Demographic Analysis Research Program;" DART proposal to FSCPE

Mulry, Mary, 1991; "1990 Post Enumeration Survey Evaluation Project P16: Total Error in PES Estimates for Evaluation Post-strata," 1990 Coverage Studies and Evaluation Memorandum Series R-6, U.S. Bureau of the Census, July 11, 1991.

Mulry, Mary and Bruce Spencer, 1991; "Total Error in PES Estimates of Population (with discussion)," Journal of the American Statistical Association, 86, pp. 839-863, 1991.

Mulry, Mary and Bruce Spencer, 1993; "Accuracy of the 1990 Census and Undercount Adjustments," Journal of the American Statistical Association, 88, pp. 1080-1091, 1993.

Robinson, J. Gregory and Kirsten K. West, 2000; "Demographic analysis". Pp 164-168 in Encyclopedia of the U.S. Census. Washington, DC: Congressional Quarterly Press, 2000.

Robinson, J. Gregory, 2001; "Accuracy and Coverage Evaluation Survey: Demographic Analysis Results," DSSD Census 2000 Procedures and Operations Memorandum Series B-4*, February 28, 2001

Robinson, J. Gregory, 2004, "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology;" Section 1, Appendix A, "Demographic Analysis"

Shores, Roger (2002) "Adjustment for Correlation Bias". DSSD A.C.E. Revision II Memorandum Series #PP- 53.'

U.S. Census Bureau, 2001. "ESCAP II. Accuracy and Coverage Evaluation: Demographic Analysis Results", by J. Gregory Robinson. Executive Steering Committee For A.C.E. Policy II, Report No. 1, October 13.

U.S. Census Bureau 2004, "Accuracy and Coverage Evaluation of Census 2000: Design and Methodology"

West, Kirsten, 2005; "'Did Proxy Respondents Cause Age Heaping in Census 2000." Paper presented at the Annual Meeting of the American Statistical Association, August 7-11, 2005.