VRA Section 203 Determinations: Statistical Methodology Summary

November 30, 2016

According to Section 203(b) of the *Voting Rights Act of 1965*, as later amended in 1982 and 2006, in certain circumstances states and political subdivisions must make voting materials available in languages other than English. Section 203(b) defines these circumstances in terms of specific determinations, made by the Director of the Census Bureau, involving the sizes and proportions of designated population subgroups. In order to make the determinations, it is necessary to estimate the total population of voting age persons who are citizens, of citizens who have limited English proficiency, and of citizens with limited English proficiency who are illiterate in approximately 8000 jurisdictions, 570 American Indian and Alaska Native Areas (AIA/ANAs), and 12 Alaska Native Regional Corporations (ANRCs), separately for 68 Language Minority Groups (LMGs). Section 203(b) specifies that the most recent decennial census and the most current available American Community Survey (ACS) are the permitted data sources used for these estimates. Many of the mandated determinations require estimates for geographic areas with a small population and extremely small subpopulations of interest. As a result, the sampling variability of direct estimates from the 5-year ACS data is large.

In order to improve the reliability of the estimates needed to make the Section 203(b) determinations, mathematical statisticians at the U.S. Census Bureau developed statistical models based on “small area estimation” techniques that allow more accurate subpopulation estimates for LMGs within political subdivisions (jurisdictions or AIA/ANAs) by synthesizing information from similar political subdivisions using a statistical model that “shares parameters” and thereby “borrows information.” These techniques have been used for decades in such programs as the Small Area Income and Poverty Estimates (SAIPE). In the Section 203(b) determinations, small area estimation uses the relations between population ratios and measured variables defined from educational level, average age, and population proportions born outside the US.

The United States is partitioned into approximately 8,000 Jurisdictions, Counties (in most states) and Minor Civil Divisions (MCDs) (in the other states). In addition to the Counties and MCDs, the law mandates estimates for approximately 570 unique AIA/ANAs and 12 ANRCs.

For purposes of Section 203(b), persons self-identify (in the Census or ACS) into 1 to 8 racial/ethnic categories that are then used to define 68 “Language Minority Groups” (LMGs), of which 16 are Asian, 51 American Indian or Alaska Native (AIAN), and one is Hispanic. (Note: only the 51 AIAN LMG categories are maintained within AIA/ANAs and ANRCs.)

Special tabulations of weighted sample survey direct estimates (henceforth, direct estimates) of state, jurisdiction, and AIA/ANA voting-age populations cross-classified by citizenship, limited English proficiency, illiteracy, and LMG are produced from American Community Survey 5-year data. These tabulations could be used to create direct estimates of all of the elements of the Section 203(b) determinations. Such direct estimates were used to provide Section 203(b) determinations prior to 2011; however, the counts of ACS-sampled voting-age persons by jurisdiction (or AIA/ANA) and LMG on which these weighted sums are based are often quite small. Consequently, the standard errors of the direct estimates are often large compared to the estimates themselves. Moreover, the standard errors estimated by current ACS methodology are also very unreliable for population counts in such small domains. For these reasons, beginning with the 2011 Section 203(b) determinations, statistical models were used to improve the reliability of estimates. In 2011, for each LMG, the Census Bureau modeled the number of voting age persons in categories defined by citizenship, limited-English proficiency, and illiteracy using a hierarchical stacked beta-binomial model. In addition to 5-year ACS data, the statistical models in 2011 made use of 2010 Census data.

For the 2016 Section 203(b) determinations, after preliminary investigations comparing direct estimates with several forms of modeled estimates similar to those used in 2011, Census Bureau statisticians again used small area estimation models, separately within each LMG, and separately for jurisdictions and for AIA/ANAs, based on the ACS 5-year 2010-2014 data. The form of model is Dirichlet-multinomial regression, a random-effects generalization of logistic regression, for the incidence of citizenship among voting-age persons within a domain (i.e., the intersection of LMG with jurisdiction or AIA/ANA), and for the incidence of LEP among voting-age citizens within the domain. Under this model, the observed data for the voting-age ACS-sampled persons within each domain, conditional on covariates and on the random-effect parameters, are modeled as independent multinomial trials. That is, each voting-age person in the domain is treated as falling into one of the mutually exclusive categories of non-citizen, LEP illiterate citizen, LEP literate citizen, or non-LEP citizen, randomly and independently of all others once covariates and domain-level random effects are taken into account. The underlying domain fixed-effect rates of citizenship and of LEP among citizens, are each modeled within LMG across jurisdiction (or AIA/ANA) as logistic regressions with the following inputs: computed covariates consisting of the corresponding rates directly estimated from the ACS at the level of the state containing the domain, and also several other covariates such as educational level, age, proportion foreign born, and average time in US, separately calculated for all adults in the domain and also for the adults in the LMG within the domain. The data used to fit this model are the direct ACS domain-level proportions. The parameters of these models are shared across geographic areas within each fixed LMG. Models are estimated using maximum likelihood, and produce estimates of citizenship and LEP rates that are weighted combinations of the direct ACS survey-weighted ratio estimates and the logistic regression model-predicted values. The weights heavily favor the direct estimators in large-population domains, where the direct estimates are very precise, and give substantial weight to the logistic regression values when the direct estimates are unstable.

The models considered were evaluated extensively using ACS 5-year 2008-2012 data in the same way that the selected model ultimately employed the ACS 5-year 2010-2014 data. Specific choices of the numbers of predictive variables used in the different LMGs and different geographic small areas (jurisdictions versus AIA/ANAs) were motivated by the richness of data (the numbers of domains with ACS sample, and the size of the samples within those domains) and the possibility of finding convergent maximum likelihood estimates within large, reasonable ranges. Generally, these criteria led to more predictor variables being used in larger LMGs. In the smallest LMGs, models were intercept-only, but still provided some stabilization of estimates across domains due to the sharing of the intercept and random-effect dispersion parameters). Finally, in some of the smallest LMGs, it was found (based on the ACS 5-year 2008-2012 data) that no model of the selected form could be fit with parameters in reasonable ranges. In those LMGs, the direct survey-weighted ACS estimates were used. For the same reasons, direct rather than modeled estimates were used in small areas defined by ANRCs.

Variances of the domain population estimates used in the determinations were computed in different ways, according to whether the estimates themselves were the direct survey-weighted ACS values or were based on estimates from the Dirichlet-multinomial models. For the direct estimates, the variance estimates were calculated, as in ACS standard and special tabulations, by a Balanced Repeated Replication method called Successive Difference Replication. In the LMGs where estimates were based on models, the variances were computed by a novel hybrid method combining replicate weights as used in ACS with parametric-bootstrap generation of pseudo-samples of survey data.

Various topics mentioned in this summary will be elaborated in depth in the forthcoming, publicly available, technical documentation. These topics include:

- comparisons of direct estimators with their standard errors,
- model assessments used to select LMG-specific, predictive (regression) covariates for the Dirichlet-multinomial models used in computing estimates,
- justification and computational details concerning the criteria used to determine when models of specific predictive complexity could be fitted by maximum likelihood with parameters in reasonable ranges,
- specific comparisons between domain population estimates generated by direct ACS methods versus the model-based estimates for those LMGs and geographic areas where the Dirichlet-multinomial-model was used,
- comparisons between direct-ACS-method variances for direct ACS population estimates versus replicate-weight and bootstrap variances for model-based estimates for those LMGs and geographic areas where the Dirichlet-multinomial-model was used,
- overall predictive parametric-bootstrap model assessments.