

LOCALITY POPULATION ESTIMATES METHODOLOGY

Rebecca Tippett, rebecca.tippett@virginia.edu or (434) 982-5861

Between decennial census years, the Demographics & Workforce Group at the University of Virginia's Weldon Cooper Center produces the official population estimates for Virginia's counties and independent cities. These estimates are used in funding formulae based on per capita allocations, in planning, in budgeting, in applications for grants, in approving and setting salaries for certain public officials, and in all manner of state agencies from VDOT to VDOE.

The locality estimates are for the mid-year, July 1, population and are released on the last Monday of the following January. For example, July 1, 2011 estimates were released on January 30, 2012. The seven month period between the estimate date and release date is the time required to collect and clean input data from multiple state agencies, to produce the estimates, and to prepare for the release on the website and to the public.

For the state population, we use the Census Bureau's estimate, released in December of the estimating year. Using the state population as a control total, we allocate population to each locality using a regression methodology known as ratio-correlation.

Ratio-Correlation Methodology

The ratio-correlation method uses linear regression to estimate population based on changes in a set of symptomatic or indicator variables that capture population dynamics. All variables in the regression are expressed as ratios. The theory underlying the method is that the change in a locality's share of an indicator variable reflects changes in its share of the state's total population.

First, *single ratios (SR)* are constructed for both base and estimating years. The base year is the most recent decennial census and the estimating year is the year for which estimates are being produced:

$$SR_{indicator} = \frac{Locality\ Total_{ind}}{State\ Total_{ind}}$$

Next, a *double ratio (DR)* is constructed to compare the locality's share of the indicator in the estimating year to its share of the state total in the base year.

Estimates vs. Projections

The difference between estimates and projections are that the former are for the present or the recent past, while the latter are for the future. The approaches for producing estimates and projections are therefore different.

Population estimates are typically based on a variety of observed administrative record data, such as births, deaths, school enrollment, and residential housing construction to detect population changes since the most recent decennial census. **Population projections**, on the other hand, predict future population change based on prior patterns. Due to unknown future natural, social, economic, and political events, population projections have higher levels of uncertainty than estimates.

In 2012, under a contract with the Virginia Employment Commission, the Cooper Center is producing population projections for 2020, 2030, and 2040 for Virginia and its counties, cities, and large towns.

$$DR_{indicator} = \frac{\textit{Estimating Year } SR_{ind}}{\textit{Base Year } SR_{ind}}$$

These double ratios measure change in the locality's share of the state total for each indicator. A double ratio less than one indicates that the locality's share of the state total for that variable has fallen relative to its share in the base year; a double ratio greater than one indicates the locality's share of that variable has grown relative to the base year.

These double ratios are used in the following general model

$$DR_{Total Pop} = \beta_0 + \beta_1 DR_{ind_1} + \beta_2 DR_{ind_2} \dots + \beta_n DR_{ind_n}$$

The Cooper Center estimates are produced using five indicator variables for each locality: total housing stock; school enrollment in grades 1-8; three-year aggregate of births; three-year aggregate of deaths; total licensed drivers. The resulting estimate is not a population total. Rather, it is the percentage of the state's total household population that will be allocated to the locality. The Cooper Center uses the Census Bureau's state population estimate minus statewide Census group quarter (GQ) population as the statewide control total. The final step in obtaining an estimate of the total population for a county is the addition of the GQ population to the estimated household population.

Detailed model deliberations are presented in Appendix A. The Cooper Center's estimating equation for total household population for 2011 to 2020 is

$$DR_{Total Pop} = -0.088 + 0.638DR_{Housing} + 0.147DR_{School} + 0.073DR_{Births} - 0.031DR_{Deaths} + 0.251DR_{Drivers}$$

Helping the U.S. Census Bureau

In 2010, the Demographics & Workforce Group was awarded two federal grants to evaluate estimates and to advise the Census Bureau on how to improve their estimates program.

The Cooper Center's research provided a detailed evaluation of the accuracy of a variety of small-area estimating methodologies. Findings revealed that the Cooper Center ratio-correlation estimates performed better than estimates produced by the Census Bureau over the past decade.

Census Bureau Estimates

The Census Bureau uses a different methodology than the Cooper Center to produce population estimates. They use administrative records data, such as vital statistics, federal tax returns, and Medicare enrollment, to track births, deaths, and migration. This methodological difference leads to different population estimates. While both estimate series are highly accurate, an evaluation showed that the Cooper Center's estimates slightly outperform those produced by the Census Bureau.

Appendix A. Developing 2010-2020 Estimating Equation

Three models were considered for the post-censal population estimates:

Model 1: Housing Stock, School Enrollment, Births, Driver Licenses, Tax Exemptions

Model 2: Housing Stock, School Enrollment, Births, Driver Licenses

Model 3: Housing Stock, School Enrollment, Births, Deaths, Driver Licenses

Results from these regressions are shown in Table A1.

Table A1. Estimating 2010 Total Population for Cities and Counties, OLS Regression Results and Associated F-Values

	Model 1	Model 2	Model 3
Housing Stock	0.601 (13.34)**	0.619 (14.00)**	0.638 (14.27)**
School Enrollment	0.133 (5.95)**	0.151 (7.60)**	0.147 (7.41)**
Births	0.064 (3.98)**	0.073 (4.78)**	0.073 (4.80)**
Deaths			-0.031 (2.00)*
Driver Licenses	0.222 (5.58)**	0.234 (5.94)**	0.251 (6.29)**
Tax Exemptions	0.056 (1.74)		
Constant	-0.087 (4.55)**	-0.088 (4.58)**	-0.088 (4.63)**
F	759.03	933.45	764.99
R^2	0.97	0.97	0.97
N	134	134	134

* $p < 0.05$; ** $p < 0.01$

Each model was further evaluated with respect to the major assumptions of ordinary least squares regression.

1. Linearity

The relationship between each of the considered predictor variables and the outcome variable (household population) was plotted and was strongly linear. In addition, plots of the residuals from each model against each of the predictor variables in the regression model behaved appropriately.

2. Normality of Residuals

Visual analysis through kernel density, P-P, and Q-Q plots showed no major deviations from assumptions of normality. However, the Shapiro-Wilk test for normality found that **Model 1** violated assumptions of normally distributed residuals.

Shapiro-Wilk test for Normal Data

Residuals from...	W	V	z	Prob>z
Model 1	0.97833	2.29	1.867	0.031
Model 2	0.98971	1.087	0.188	0.425
Model 3	0.98367	1.726	1.23	0.109

3. Homoscedasticity of Residuals

Two separate tests, White's and Breusch-Pagan, reveal that only Model 1 residuals exhibit heteroscedasticity.

Prob>chi2 for...

Residuals from...	White's Test	Breusch-Pagan Test
Model 1	0.0331	0.5975
Model 2	0.2860	0.4517
Model 3	0.4232	0.4989

Based on these results, **Model 1** was eliminated from consideration as a potential estimating equation.

A comparison of estimation performance, predicting 2010 population from 2000, revealed that **Model 3** slightly outperformed **Model 2** (Table A2-2). While Model 3 had nine localities with greater than 5 percent estimating error, compared to seven in Model 2, it had a larger number of cases predicted within 1% of their actual value. In addition, it had a lower overall MAPE and the average error of localities with prediction errors greater than 5 percent was lower than in Model 2.

Table A2. Comparing Estimation Performance, Model 2 vs. Model 3

Estimation Error	Model 2	Model 3
Within 1%	46	52
Within 2%	84	90
Within 3%	107	109
Within 4%	118	117
Within 5%	127	125
>5%	7	9
MAPE	1.92	1.90
Average Error of those >5%	6.46	6.28

As a result of these comparisons, **Model 3** was selected for the post-2010 population estimates.