MODIFICATIONS TO CONNECTICUT SMALL AREA POPULATION ESTIMATES – 2018

Connecticut Department of Public Health, Health Statistics and Surveillance Section,

Statistics Analysis & Reporting Unit

Mueller L., Hayes L, Backus K

Annual Small Area Populations Estimates (SAPEs) for the state of Connecticut have been made available for the period 2011-2014 by the Connecticut Department of Public Health (CTDPH) Surveillance, Analysis, and Reporting (SAR) Unit as of December, 2017. A description of the data sources and statistical modeling approach for the initial development of the SAPEs are provided in the document "Small Area Population Estimates Project Summary Report" [1], with figures and tables summarizing early project results. This update provides an overview of changes made to the production the SAPEs between the time of writing of the initial project summary report referenced above and the release of the estimates by CTDPH in December, 2017. We have also added a description of the background events that motivated us to undertake this project.

Background

The motivation for undertaking this project derived from the general need for population data to support the calculation of population-based health indicator rates for small-geographic areas within Connecticut for the Connecticut's State Innovation Model (SIM) program[2,3], for CTDPH, and for others in the community. While the Census Bureau has published local area estimates through the American Community Survey (ACS), CTDPH came to the specific realization that currently available ACS data were not adequate to meet these needs. The SAPEs project is a result of those motivating factors and is focused on improving the stability and accuracy of Connecticut annual population estimates by town, age, sex, and race/ethnicity. Our early use of the the 5-year town population estimates from the American Community Survey (ACS) resulted in the production of some health statistics that varied erratically over time. Evaluation of the components of the erratic rates revealed that ACS estimates themselves were erratic over time for many of the towns in Connecticut. For example, using ACS data, the all-cause age-adjusted mortality rates were estimated to have doubled between 2002-2006 and 2007-2011 for males in the town of Bozrah, jumping from 881 to 1,981 deaths per 100,000. In this instance, the shift in mortality rates occurred without a comparable change in the death counts and with somewhat erratic shifts in the ACS population figures by age and sex. Similarly, we found that town-level birth rates for women 15-19 years old, when based on ACS populations estimates, changed erratically over time. As the evidence accumulated, we became concerned that publishing statistics based on the ACS town-level estimates might send people off to "chase windmills" rather than address real health problems. These problems prompted the SAR unit to suspend the use of town-level ACS population estimates in publications and to examine the reliability of the ACS figures in greater detail.

The ACS population statistic are based on a representative sample of the population in each state that responds to a survey. Since the ACS estimates are based on a small random sample of a larger population, the ACS figures are measured with a degree of variability that can be estimated (sampling error). The Census reports margin-of-error (MOE) estimates along with the ACS population statistics to give users a clear sense of the reliability of their estimates. In contrast, the decennial Census figures are based on a complete enumeration of the whole US population. Consequently, those figures are not subject to sampling error, though other sources of measurement error may be present. The published ACS MOE figures are for 90% confidence intervals, meaning that the true value for that statistic will lie within the interval range 90% of the time. The ACS MOE information for Connecticut population estimates was used was used to assess the general degree of precision of these estimates. (equal to +/- 1.645 * Standard Error) .

A common approach to evaluating the precision of survey and non-survey statistics s to compare the magnitude of the point estimate (e.g. a population count) with a measure of its variability, its standard error (SE). The SE/Estimate ratio is referred to as the coefficient of variation (CV). Ideally the SE should be small in comparison to the estimate, and therefore the CV should be small. To evaluate the precision of the ACS data for Connecticut towns, we adopted a liberal SE/Estimate ratio threshold of ½ to distinguish between poor quality estimates (CV ratios > ½) and adequate quality estimates (CV ratios <= ½). For comparison, CDC often recommends that a CV value no higher than .30 be deemed adequate for publication of counts. We applied our classification dichotomy (CV<.50 and CV>.50) to the 5-year, 2010-2014 ACS population estimates by town, age and sex (T*A*S) for Connecticut. Overall, only about two-thirds of the 2010-2014 ACS estimates by T*A*S met our threshold for adequate quality (CV ratio <= ½). Furthermore, only five of Connecticut's 169 towns had adequate CVs across all of the town's 36 population estimates by age and sex..

During our review of the ACS estimates, we discovered that other researchers shared our concerns. A summary from the 2011 Workshop on the ACS[4] sounded a cautionary note about the large errors terms associated with some ACS small-area estimates.

> "There are concerns about the large standard errors associated with the ACS estimates for small geographic areas and small population groups. Data users are dealing with these errors in different ways, depending on their particular applications and audiences. Some present margins of error along with every ACS estimate, while others ignore sampling error under certain circumstances." [4]

Concerns about using the ACS 5-year population data for census tracts and block groups may be more significant. A recent review by Spielman[5] states the problem plainly: "While there are clear advantages to working with fresh data, the ACS margins of error are so large that for many variables at the census tract and block group scales, the estimates fail to meet even the loosest standards of data quality." As a result of all these factors, we had a strong sense of the need for improved Connecticut SAPE figures. The work described below and in the original report was made possible as a result of the Center for Medicare & Medicaid Innovation (CMMI) SIM initiative grant to the state of Connecticut[2].

Methodological Improvements

The same overall statistical modeling methodology was used from beginning to end of the project. The population sizes based on the 2010 decennial census published by the United States Census Bureau (USCB) were modeled as a function of input variables specific to that same year. Data for important input variables identified from modeling work were then gathered for subsequent years (2011-2014) and used in conjunction with the optimal models identified using the 2010 decennial census data to estimate population sizes in these subsequent years. Within this overall methodology, however, two changes to the SAPE production protocol did occur between earlier modeling attempts and the final SAPE production. The first change to the methodology was a modification to model inputs such that different inputs were utilized based on their predictive utility. The second change was adopting a enhanced model selection method, based on ten-fold cross-validation, with the goal of improving the reliability of the final predictive models.

Changes to Model Inputs

The original project report describes the development of five statistical models—one for each of the five age groups of 0 - 4, 5 - 14, 15 - 19, 20 - 64, and 65 and older. As described in the original report, individual statistical models for each of these age groups were developed for final SAPE production using the statistical modeling tool Multivariate Adaptive Regression Splines (MARS). However, for three of the five models, the specific data inputs themselves changed, due either to a change in data availability or to the omission of data from a specific source based on evidence of limited predictive utility of the variable in one or more of the models. In all five age group models, the revised SAPE model inputs allowed for modeling of the total population size by T*A*S*RH:

Ages 0 – 4: Model inputs for final SAPE production are as described in the original project report, i.e. natural increase totals by T*A*S*RH.

Ages 5 – 14: Model inputs for final SAPE production for this age group were the Connecticut State Department of Education (SDE)-based public and private school enrollment totals, as described in the original project report. However, SDE public school enrollment data described in the original report were censored for low cell counts, such that the RH details from the latter dataset had to be estimated. After revision of the data use agreement with the SDE, public school enrollment totals by T*A*S*RH with no censoring for low cell counts became available. These detailed files provided the public school enrollment data for final SAPE production. Private school enrollment data availability remained unchanged from the original project report, whereby enrollment totals were only available by T*A. Thus, calculations of private school enrollment totals by T*A*S*RH were possible after splitting T*A totals equally into male and female subgroups and then applying public-school enrollment RH proportions to the private school enrollment data by T*A*S.

Ages 15 – 19: Model inputs for this age group were as described in the original report whereby total private and school enrollments were used for ages 15-17 and total driver's license counts from

the Connecticut Department of Motor Vehicles (DMV) were used for ages 18-19. However, modifications to inputs include those pertaining to public and private school enrollment data as described for the ages 5-14 model. Additionally, RH proportions from school enrollment data were applied to the T*A*S DMV data breakdown such that final model input of the sum of school enrollments and DMV license issuances was by T*A*S*RH. The size of the total town population under the poverty level, a described input for this model in the original report, was not used as a model input in the final SAPE production for this age group model due to only a negligible increase in predictive ability when included in the model with the school enrollment-DMV license total input.

Ages 20 – 64: This age group used, for final SAPE production, the inputs of DMV license totals as well as variables pertaining specifically to the group quarters populations, i.e. university and correctional facility data.

For the DMV data, license counts were available only by T*A*S, so RH proportions for DMV inputs were estimated based on weighted averages from the natural increase and Medicare enrollments inputs used for the ages 0-4 and 65+ models, respectively. Under this approach, DMV license counts for each T*A*S group between 20-64 years were distributed to RH groups according to proportions calculated as the weighted averages of RH proportions in the same T*S group. For example, the 60-64 age group was assigned RH proportions that were (12/13) of the RH proportions from the Medicare enrollee count for the 65-69 age group and (1/13) of the RH proportions for the 0-4 population based on natural increase.

For correctional facilities data, total population sizes by facility are available to the public but are insufficient for modeling demographics subgroups. In order to create total population size estimates for correctional facilities by T*A*S*RH to be used as model inputs, additional information sources were used to partition these data. A*S*RH proportions were based on tabulated data in CT Department of Corrections Monthly Reports published on August 1, 2010. Correctional facility population sizes by A*S*RH were then aggregated by town to create model inputs by T*A*S*RH.

Finally, for the university population variable, total population sizes for each university in Connecticut are obtained annually by CTDPH. In order to create population sizes by A*S*RH, proportions by university were derived from the National Center for Education Statistics (sex proportions for Fall enrollment 2016) and the U.S. Department of Education (race-ethnicity proportions for enrollment year 2015-2016) and the University of Connecticut's Fall 2015 enrollment tables (age proportions based on total enrollment of full-time students ages 20-24 years). A*S*RH proportions derived by combining data from these three sources were then applied to CTDPH Group Quarters totals for university populations and then counts by A*S*RH were aggregated by town. Additionally, group quarters total population sizes by T*A*S obtained from 2010 decennial census in the towns of Groton and New London, with RH proportions from the U.S. Coast Guard Academy, were added to university population totals to account for the sector of the military population living in group quarters between of ages 20-24.

Utilities data, as well as the below-poverty level population size, as described in the original report as model inputs, were not used in final SAPE production due to negligible increases in predictive ability for total population size by T*A*S*RH for this age group model. The DMV and Group Quarters counts overlap and identify some of the same people. However, the modeling adjusts for this overlap. Notably, the model performed better using total population as the dependent variable rather than modeling separate Household and Group Quarters population components.

Ages 65+: Model inputs for final SAPE production are as described in the original project report, i.e. Medicare enrollment totals by T*A*S*RH.

In addition to the continuous inputs specific to each age group model described throughout the section above, variables pertaining to subgroup identification also were entered as model inputs. For age groups models that contained more than one five-year age group, the five-year age group entered the models as a continuous input. Categorical inputs for all models were sex (either male of female), race-ethnicity (either non-Hispanic white, non-Hispanic black, non-Hispanic American Indian or Alaska Native, non-Hispanic Asian or Pacific Islander, or Hispanic), and town (corresponding to each of the 169 towns in Connecticut). This feature allows the models to be more flexible and to reflect different prediction patterns by T*A*S*RH.

Changes to Model Selection

The second change to the SAPE production protocol pertained to model selection. Significant changes were made to improve the ability of the final, selected models to accurately predict annual population figures. Model development was done using the Multivariate Adaptive Regression Spline (MARS) technique, as described in the original project report. While model selection in preliminary analyses using MARS relied solely upon use of the Generalized Cross Validation criterion (GCV), final SAPE production relied on a more comprehensive assessment aimed at improving the reliability of the final predictive models. It was based on inferences derived from several competitive models that were selected based on minimization of the 10-fold cross-validation error score repeated with 50 random iterations.

Ten-fold cross-validation is a procedure by which a dataset is first split into 10 partitions, each of which serves once as a hold-out sample. A model is derived using 9 other partitions, and the model parameters are used to make independent predictions in the hold-out sample. The model's fit is assessed by calculating a mean-squared-error (MSE) in the hold-out sample. This process is repeated ten times such that the hold-out sample role rotates through all 10 partitions. The optimal number of model predictors is determined first. The average MSE over 10 models is calculated by model size. The model size associated with the lowest average MSE for 10 hold-out samples is considered optimal[6]. Once the optimal number of model components, or basis functions is determined, then a final model using the full dataset is developed. [7].

Selection of an optimal model through 10-fold cross-validation is subject to variability based on the random number seed used to partition data into the individual ten datasets used in cross-validation. In other words, the "optimal" model selected can change depending on how the dataset used for modeling was partitioned due to the random number seed used. To incorporate such variability into the SAPE production, a series of 50 random number seeds were used, such that an optimal model was selected 50 times. Final SAPE production was thus based on a suite of 50 models derived from each iteration of the 10-fold process. Unraked population estimates for 2011-2014 SAPEs were based on the average of model-based predictions for corresponding years across all 50 models. Variation across estimates from the 50 models provides a way to the examine consistency or precision of this modeling approach. Ultimately, the 2020 Census will be used to assess the accuracy of our population model predictions.

Software used for MARS modeling for final SAPE production was Salford Predictive Modeler® software suite 8.2. The maximum number of basis functions was specified as 100 and the maximum number of interactions between basis functions was set to 3. Use of the CVREPEATED battery allowed for use of 50 random number seeds for 10-fold cross validation. Model output files from SPM software were exported to SAS® 9.4 for averaging of model estimates, model evaluation, and scoring datasets for the years 2011-2014. Final steps in SAPE production for the years 2011-2014 were raking the model-averaged estimates using iterative proportional fitting in SAS so that population estimate subtotals by town (total population) and county (A*S*RH subgroups) conformed with those released by the USCB annually[8].

Future SAPE Production

Production of annual SAPEs for the period 2015-2016 are currently underway and CTDPH will continue to release annual SAPEs for post-2016 years as model input data become available. Note that all SAPEs provided by SAR until release of the official 2020 decennial census population estimates by the United State Census Bureau (USCB) will be reviewed. This is due to the fact that, despite rigorous methodologies applied for data screening and model development, final validation of the estimation production protocol cannot occur until a validation dataset becomes available via release of the 2020 USCB estimates. SAPEs released by SAR from 2011 forward will be finalized retroactively following analysis using this 2020 USCB validation dataset.

Citation:

Mueller L., Hayes L, Backus K (2018). Modifications to Connecticut Small Area Population Estimates – 2018; Connecticut Department of Public Health, Health Statistics and Surveillance Section, Statistics Analysis & Reporting Unit, Hartford, CT.

References

1. Hu Q, Li X, Zhang W, and M Howser. 2016. Small Area Population Estimates Project Summary Report.

2. SIM Program Management Office: http://www.healthreform.ct.gov/ohri/site/default.asp .

3. SIM and the Office of Health Strategy: http://portal.ct.gov/OHS/Content/State-Innovation-Model-SIM

4. Summary of a Workshop on the American Community Survey; Population Reference Bureau, June 28, 2011. http://www.prb.org/Publications/Articles/2011/acs-workshop-2011.aspx

5. Spielman SE, Folch D, Nagle N. Patterns and causes of uncertainty in the American Community Survey. Applied Geography. 2014;46:147–57. doi: 10.1016/j.apgeog.2013.11.002

6. James G, Witten D, Hastie T, and R Tibshirani. 2013. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

7. Milborrow S. 2017. Notes on the earth package. http://www.milbo.org/doc/earth-notes.pdf.

8. Backus K, and L Mueller. 2016. Population Estimates for Connecticut. The Connecticut Department of Public Health. http://www.portal.ct.gov/DPH/Health-Information-Systems--Reporting/Population/Population-Statistics

Table 1: Connecticut 2010 Population Estimates Model: Results Summary

	Final Model Characteristics				
	Continuous Input Variables		Fitted 2010 Model*		
Age Cohort Covered	Short Name	Description	Test R^2	R^2	N Basis functions
Ages 0-4	Natural Increase	Natural Increase Accrued Over Past 5 Years	0.991 (0.983 - 0.995)	0.998	33 (8 - 66)
Ages 5-14	School Totals	Total Public and Private School Enrollments	0.998 (0.997 - 0.998)	0.999	37 (21 - 65)
Age 15-19	School Totals + DMV	Total Public and Private School Enrollments (ages 15-17) + DMV Driver's License and ID Totals (ages 18 - 19)	0.977 (0.946 - 0.983)	0.996	34 (6 - 65)
Ages 20-64	DMV	DMV Driver's License and ID Totals	0.985 (0.982 - 0.987)	0.991	68 (35 - 76)
	Correctional Facilities	Correctional Facilities Population Sizes			
	Universities	University Population Sizes			
Ages 65+	Medicare	Medicare Enrollment Totals	0.997 (0.997 - 0.998)	0.999	58 (39 - 65)

*Model Characteristics:

Test R^2	"R-squared " is equal to the ratio of [Explained variation / Total variation] , and varies between 0.0 and 1.0. Values shown reflect the median and range of R^2-values on hold-out, i.e. test, samples across all 50 optimal models identified during model development.
R^2	Values in this column indicate R2 for a simple linear regression model of decennial population sizes by T-A-S-R-H regressed on model-based estimates of population sizes, as averaged aross all 50 models identified as optimal during model development and prior to raking.
N Basis functions	Multivariate adaptive regression splines (MARS) use basis functions as components. Basis functions typically define the relationship between a predictor or a set of predictors and the dependent variable for a specific segment or subset of the predictors range. A single basis function can include multiple predictors. The median and range of the number of model basis functions is shown, as identified from the 50 optimal models identified during model development.

Table 2: Age-Cohort Specific Model Details

Model for Ages 0 - 4		
Input Variable Short Name*	Input Variable Type	Input Variable Importance**
Natural Increase	Continuous	100.0 (100.0-100.0)
Town	Categorical	15.8 (10.9 - 18.0)
Race-Ethnicity	Categorical	12.7 (7.3 - 14.6)
Sex	Categorical	1.7 (0.6 - 1.9)

Model for Ages 5 - 14		
Input Variable Short Name*	Input Variable Type	Input Variable Importance**
School Totals	Continuous	100.0 (100.0-100.0)
Race-Ethnicity	Categorical	11.8 (11.5 - 12.5)
Town	Categorical	10.8 (10.5 - 11.7)
5-year Age Group	Continuous	1.8 (1.5 - 1.9)
Sex	Categorical	0.5 (0.4 - 0.5)

Model for Ages 15 - 19		
Input Variable Short Name*	Input Variable Type	Input Variable Importance**
School Totals + DMV	Continuous	100.0 (100.0-100.0)
Town	Categorical	58.8 (28.8 - 59.9)
Race-Ethnicity	Categorical	42.6 (38.4 - 44.0)
Sex	Categorical	7.2 (6.5 - 7.6)

Model for Ages 20 - 64		
Input Variable Short Name*	Input Variable Type	Input Variable Importance**
DMV	Continuous	100.0 (100.0-100.0)
Town	Categorical	34.1 (33.3 - 35.0)
Race-Ethnicity	Categorical	24.3 (23.5 - 24.8)
Universities	Continuous	21.6 (20.8 - 21.9)
5-year Age Group	Continuous	15.1 (14.5 - 15.4)
Sex	Categorical	4.8 (4.3 - 5.8)
Correctional Facilities	Continuous	4.8 (4.1 - 5.7)

Model for Ages 65 +		
Input Variable Short Name*	Input Variable Type	Input Variable Importance**
Medicare	Continuous	100.0 (100.0-100.0)
Race-Ethnicity	Categorical	15.8 (15.4 - 15.8)
Town	Categorical	9.1 (8.7 - 9.1)
5-year Age Group	Continuous	7.4 (7.3 - 7.6)
Sex	Categorical	2.8 (2.5 - 2.8)

***** See text and Table 1 for detailed descriptions of Input Variables.

****** The most important predictor is given an arbitrary score of 100. The values for other predictor are scaled relative to 100. Values shown reflect the median and range of variable importance values across all 50 optimal models identified during model development.