MATHEMATICA
Policy Research

Small Area Estimates Produced by the U.S. Federal Government: Methods and Issues

Small Area Estimation Conference
Maastricht, The Netherlands

August 17-19, 2016

John L. Czajka
Mathematica Policy Research

Outline

- Overview of U.S. government applications of small area estimation (SAE)
- Review of illustrative examples
- Issues for federal agencies in producing small area estimates
- Responding to changes in the available data
  - Use of the American Community Survey
  - Health care reform

Overview of U.S. Government Applications of SAE

Production applications within HHS

- State estimates from the National Survey on Drug Use and Health (SAMHSA)
- County estimates of diabetes prevalence, incidence, and risk factors (CDC)
- State and county estimates of cancer risk factors and screening (NCI, NCHS, U. Michigan, U. Pennsylvania)
- State estimates of wireless substitution (NCHS)
- State estimates from the Medical Expenditure Panel Survey (AHRQ)
- State and substate estimates of vital statistics for five single race groups (NCHS)

Production applications outside HHS

- **Small Area Income and Poverty Estimates—SAIPE (Census Bureau)**
- **Small Area Health Insurance Estimates—SAHIE (Census Bureau)**
- **State and local area employment and unemployment statistics (Bureau of Labor Statistics)**
- **Persons eligible for two programs that provide food assistance to low-income families (Food and Nutrition Service)**

Research and development

- National Center for Health Statistics
  - Diabetes prevalence among 11 small domains based on race
  - Health insurance and access-to-care for states and counties
  - County estimates of the adoption of electronic medical record systems by office-based physicians
- Agency for Healthcare Research and Quality
  - Extension of limited state and metropolitan area estimates to more states, areas, and variables
  - County estimates of selected health conditions, combined with hospital discharge data to measure quality of care

Research and development cont'd

- National Cancer Institute
  - County estimates of tobacco-related indicators
  - State estimates of cancer-related knowledge
- Bureau of Justice Statistics
  - State and substate estimates of crime victimization based on the National Crime Victimization Survey
- U.S. Department of Agriculture
  - County agricultural estimates to expand or replace the traditional program
- National Center for Education Statistics
  - State and county estimates of adult literacy
  - State and county estimates of mathematics and science knowledge and skills

Illustrative Examples

Drug use

- Substance Abuse and Mental Health Services Administration (SAMHSA) Conducts National Survey on Drug Use and Health
- Uses SAE to produce age-group-specific state prevalence estimates for 25 behaviors, including use of illicit drugs, alcohol, and tobacco
- Methodology is survey-weighted hierarchical Bayes
- Modeling is at the person level
- Substate predictors used in regression models—to reflect the geographic location of respondents—are drawn from 7 federal agencies and a private vendor
- To improve precision, estimates are two-year moving averages

Cancer risk factors and screening

- Collaborative effort of NCI, NCHS, and Universities of Michigan and Pennsylvania
- State and county prevalence estimates of:
  - Current and past smoking of adult males and females
  - Mammography screening for women 40 and older
  - Pap smear tests for adult women
- Methodology combines estimates from two surveys—Behavioral Risk Factor Surveillance System (state-based telephone surveys with very large sample) and National Health Interview Survey (in-person)—with county-level demographic and socio-economic variables
- Uses hierarchical Bayesian model
- Role of NHIS is to correct for undercoverage and non-response bias

Income and poverty estimates

- Census Bureau developed methodology with assistance of a National Academy of Sciences panel
- State and county estimates of:
  - People in poverty
  - Children under 5 in poverty (states only)
  - Related children 5 to 17 in families in poverty
  - Children under 18 in poverty
  - Median household income
- School district estimates of:
  - Total population; children 5 to 17; related children 5 to 17 in families in poverty

Income and poverty estimates cont'd

- **Bayesian methodology based on Fay and Herriot (1979) combines direct survey estimates with regression predictions of these direct estimates, with weights a function of relative precision of survey estimates and model-based estimates**
- **Many areas have no direct estimates; their estimates are based on the model alone**
- **As implemented originally, direct estimates were drawn from the Current Population Survey, and predictors were drawn from administrative records (taxes and food stamps), Census Bureau population estimates, per capita personal income, and income and poverty statistics from the most recent decennial census**
- **Revised methodology using the American Community Survey is discussed below**

Participation in Supplemental Nutrition Assistance Program (SNAP)

- SNAP participation is underreported in surveys, so participation rates use administrative counts of participants in the numerator
- Denominator (eligible persons) derived from a microsimulation model, but direct estimates too imprecise at the state level
- Empirical Bayesian shrinkage estimator used to derive more precise estimates of four variables
  - SNAP participation rate among all eligible persons
  - Number of eligible persons
  - SNAP participation rate among the eligible working poor
  - Number of eligible working poor

Participation in SNAP cont'd

- **Predictors in regression models include three based on administrative data—percentage of state population receiving SNAP, percentage of school age children certified to receive a free lunch, and a child poverty rate calculated from tax data—and four measures obtained from a three-year roll-up of American Community Survey data published by the Census Bureau**
- **To borrow strength over time, estimates are produced for three years at a time, and equations are estimated jointly**
- **Shrinkage estimate for a given year is not constrained to fall between the direct and indirect estimates *for that year* but generally do so**

Issues for Federal Agencies in Producing Small Area Estimates

Observations about U.S. applications

- Relatively few applications produce annual estimates but show a high level of sophistication and cutting edge methods
- Development of an SAE program requires significant time and resources, but these are small relative to data collection
- Software limitations are an especially challenging part of development and implementation
- Auxiliary variables present a number of challenges
- Comparative evaluations of competing approaches are rare

Observations cont'd

- Varied approaches to validation have been used
- Interpretation and communication of results require careful attention
- Collaboration presents both opportunities and challenges
- The American Community Survey changes the landscape for SAE in a number of ways

Requirements to establish an SAE program

- Technically qualified staff—with modeling expertise in particular
- Strong programming staff
- Resources and time for development and evaluation
- Communication skills—for presentation and interpretation to users
- Hiring qualified staff to do production work can be difficult; alternative options exist:
  - Collaboration (NCI and NCHS with university faculty)
  - Contracting (SAMHSA with RTI, FNS with Mathematica)
  - Assemble an expert panel (Census Bureau)

Software issues

- Software packages reduce the need for high-level programming staff, but users report issues with convergence and excessive run time
- Limited model diagnostics—for example, in detecting over-specification
- Some found it necessary to program the entire application in C or R or even FORTRAN
- Writing one's own software requires higher level programming staff and greater statistical expertise—not always an option

Auxiliary variables present challenges

- Program microdata have played key roles in SAE, but access and use are restricted; aggregate data are more readily available but also more limited
- Quality is important, and relying on evaluations performed by the data producer may be insufficient
- Potential impact of data anomalies underscores importance of consulting with program staff
- Too much time may be spent finding potential covariates when basic variables may work just as well
- Variables used in a model are removed from future analyses of area variation
- If small area estimates are to be used to measure change, choice of variables may need to reflect this

Comparative evaluations of methods are rare

- Literature search identified few examples of comparative evaluations of methods
- Typically, agencies explore alternative in literature and identify an approach that is consistent with their objectives, data, and resources
- Effort required to develop and test two competing approaches discourages researches from empirical evaluation of alternatives
- Consequently, we know less than we might like about the comparative strengths and weaknesses of major approaches

Varied approaches to validation

- Validation of estimates can present a significant challenge because reliable estimates are generally limited—if they exist at all
- Simulations using an artificial population have been used in several programs; subsamples are drawn, estimates created, and compared to "truth"
- One program used a large survey with related measures and constructed maps for comparison
- Preservation of known correlations was another approach
- Cross-validation has also been used: removing a subset of the areas, re-estimating the model, and applying it to the areas removed

Interpreting and communicating results

- No less than with direct estimates, producers of small area estimates need to consider how best to communicate variability of estimates
- Users interested only in point estimates present a particularly difficult challenge
- Meeting with stakeholders can be invaluable
- Maps can be exceedingly useful in validation, interpretation, and communication
- Some producers have developed informative graphics to assist users in making comparisons across areas

Improving cross-agency collaboration

- Collaboration between agencies can be very useful
- Challenges to collaboration can be significant—particularly across departments
- Common interest in a successful collaboration is critical
- The idea of a group that would meet periodically to share experiences was appealing to agency staff involved in SAE

Using estimates for policy purposes

- **Variability of estimates from year to year can present a problem when estimates are used for funding allocation of other policy purpose**
- **Variability becomes a more pronounced issue when estimates are used as thresholds to trigger or cut off funding**
- **Congress may address this by limiting year-to-year change in estimates or allowing areas to choose between current and prior year estimates (hold harmless provisions)**

The American Community Survey

- Eliminates long-form variables previously used in models and validation
- Provides direct estimates for many characteristics and areas that required SAE previously
- Provides wide range of new auxiliary variables
- Eliminates the need to deal with a variable time lag in incorporating prior census results into models
- Expands opportunities for SAE through pairing with other surveys
- Shows limits of even large scale surveys
  - Precision requires 3 or 5-year averages for most substate areas
  - SAE can provide more timely estimates for these areas

Responses to Changes in the Available Data

Use of the ACS in SAIPE

- The ACS replaced the CPS as the source of direct survey estimates of income and poverty for the SAIPE program
- A direct estimate is obtained for every county in which the survey estimate of poverty is nonzero
- The regression model is estimated over counties with direct estimates, but predictions are developed for all counties
- Because the 2010 census does not provide estimates of income and poverty, predictors based on census poverty are drawn from the 2000 census

Overview of health care reform in the U.S.

- The Affordable Care Act had two main objectives with respect to increasing health insurance coverage:
  - Make private nongroup coverage more affordable through a combination of tax credits and reduced cost sharing
  - Expand eligibility for public coverage to fill the gap between current coverage and the subsidies for private insurance
- Provisions were phased in over time
- Health insurance coverage has increased in both public programs and private plans
- Increased public coverage is reflected in program administrative data, available at the county level

Adapting SAHIE to health care reform

- SAHIE combines current survey estimates with regression predictions using auxiliary variables at the county level
- Key auxiliary variables on public coverage lag the survey data; changing coverage trends induced by health care reform make the lag problematic
- Survey estimates are current, but is that sufficient to overcome the lag in the auxiliary variables?
- Census Bureau introduced aggregate administrative measures available without a lag to address trend issue
- Initial evaluation showed little difference between the alternatives
- How to interpret this finding not yet clear

For More Information

- John L. Czajka
  - JCzajka@mathematica-mpr.com