Statistical Small Area Estimation: Some examples and current projects at NCHS

Donald Malec, National Center for Health Statistics, CDC

Joint Roundtable on Health Data Needs for Community Driven Change
National Committee on Vital & Health Statistics
Subcommittee on Population Health
Subcommittee on Privacy, Confidentiality and Security
Subcommittee on Standards
Hyattsville, MD - May 1, 2013

The findings and conclusions in this presentation are those of the author and do not necessarily represent the official positions of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Definition: *Statistical* Small Area Estimation

1. **Estimation for a geographic area (or population domain) when the sample size does not provide precise (direct) estimates**

Definition: *Statistical* Small Area Estimation

1. **Estimation for a geographic area (or population domain) when the sample size does not provide precise (direct) estimates**

- **Required: the small area sample is part of a larger sample that does provide precise (direct)estimates for the larger area**

Definition: *Statistical* Small Area Estimation

1. **Estimation for a geographic area (or population domain) when the sample size does not provide precise (direct) estimates**

- **Required: the small area sample is part of a larger sample that does provide precise (direct)estimates for the larger area**
- **Optional: related covariates are available for all small areas**

Examples Of The Definition:

1. **Countries as Small Areas: Colombia, Costa Rica, Indonesia, etc.**

Examples Of The Definition:

1. **Countries as Small Areas: Colombia, Costa Rica, Indonesia, etc.**

2. **Census Block Groups as Large Areas**

Examples Of The Definition:

1. **Countries as Small Areas: Colombia, Costa Rica, Indonesia, etc.**

2. **Census Block Groups as Large Areas**

3. **Demographic group as small domain: Native Hawaiian or Pacific Islanders**

Examples Of The Definition:

1. **Countries as Small Areas: Colombia, Costa Rica, Indonesia, etc.**
   - **Context: The World Fertility Survey - contraceptive use**
   - **Small Area Method: Wong and Mason, JASA, (1985)**
     - **Within Country Covariates: Education Level, Rural Status**
     - **Between country Covariates: Gross National Product, Effectiveness rating of family planning program**
2. **Census Block Groups as Large Areas**
3. **Demographic group as small domain: Native Hawaiian or Pacific Islanders**

Examples Of The Definition:

1. **Countries as Small Areas: Colombia, Costa Rica, Indonesia, etc.**
    - **Context: The World Fertility Survey - contraceptive use**
    - **Small Area Method: Wong and Mason, JASA, (1985)**
        - **Within Country Covariates: Education Level, Rural Status**
        - **Between country Covariates: Gross National Product, Effectiveness rating of family planning program**
2. **Census Block Groups as Large Areas**
    - **Context: The American Community Survey - many estimates: more than 200,000 Areas in U.S and Puerto Rico**
    - **Small Area Method: None needed for 5-year estimates - block group sample large enough**
3. **Demographic group as small domain: Native Hawaiian or Pacific Islanders**

Examples Of The Definition:

1. **Countries as Small Areas: Colombia, Costa Rica, Indonesia, etc.**
   - **Context: The World Fertility Survey - contraceptive use**
   - **Small Area Method: Wong and Mason, JASA, (1985)**
     - **Within Country Covariates: Education Level, Rural Status**
     - **Between country Covariates: Gross National Product, Effectiveness rating of family planning program**
2. **Census Block Groups as Large Areas**
   - **Context: The American Community Survey - many estimates: more than 200,000 Areas in U.S and Puerto Rico**
   - **Small Area Method: None needed for 5-year estimates - block group sample large enough**
3. **Demographic group as small domain: Native Hawaiian or Pacific Islanders**
   - **Context: Diabetes prevalence measured in the National Health Interview Survey**
   - **Small Area definition requirements: NHIS diabetes prevalence can be precisely estimated for the U.S.**

What Are Small Area Estimates?

Typical form: $\hat{w} * \hat{y}(direct) + (1 - \hat{w})\hat{y}(model)$

- $\hat{y}(direct)$: Estimate using only data within the small area
- $\hat{y}(model)$: Estimate for small area using a model of the relationship across small areas
- $0 \leq \hat{w} \leq 1$: weight - estimated from data. Gets larger as the small area sample increases

What Are Small Area Estimates?

Typical form: $\hat{w} * \hat{y}(direct) + (1 - \hat{w})\hat{y}(model)$

- $\hat{y}(direct)$: Estimate using only data within the small area
- $\hat{y}(model)$: Estimate for small area using a model of the relationship across small areas
- $0 \leq \hat{w} \leq 1$: weight - estimated from data. Gets larger as the small area sample increases

Example: County per capita income, Fay & Herriot, JASA, (1979)

- $\hat{y}(direct)$: log of county PCI using county data from the Current Population Survey
- $\hat{y}(model)$: $\hat{a} + \hat{b} \times log(CensusPCI)$

Why Small Area Estimation?

1. **Policy decisions, funding allocation and interventions are often based on quantifiable needs**
   - **Typical estimates which use only data from each area may be suppressed due to small sample size**
   - **Small Area Estimates could fill this gap**

Why Small Area Estimation?

1. **Policy decisions, funding allocation and interventions are often based on quantifiable needs**
   - **Typical estimates which use only data from each area may be suppressed due to small sample size**
   - **Small Area Estimates could fill this gap**
2. **More demand for statistics on small areas or domains than are affordable through increased sample size**
   - **Small Area Estimates do not require additional sample - are labor (model based) intensive**
   - **Can provide a "first look" at data - use small area data to prioritize resources for additional sample**

Why Small Area Estimation?

1. **Policy decisions, funding allocation and interventions are often based on quantifiable needs**
   - **Typical estimates which use only data from each area may be suppressed due to small sample size**
   - **Small Area Estimates could fill this gap**
2. **More demand for statistics on small areas or domains than are affordable through increased sample size**
   - **Small Area Estimates do not require additional sample - are labor (model based) intensive**
   - **Can provide a "first look" at data - use small area data to prioritize resources for additional sample**
3. **Uniform quality - uniformity of definitions across Small Area Estimates**
   - **Input data from same survey**
   - **Covariates typically (can be required) to be from the same source (e.g. IRS)**
   - **Estimates and model constructed at one time - share the same assumptions**

Details

Assumptions Needed

- **Although data-based, the model may still not fit well for some small areas**
- $\hat{y}(direct)$ **usually needs further model assumptions to be implementable... e.g., Normality**
- **Estimates of** $Var(\hat{y}(direct))$ **are often imprecise but needed**

Details

Assumptions Needed

- **Although data-based, the model may still not fit well for some small areas**
- $\hat{y}(direct)$ **usually needs further model assumptions to be implementable... e.g., Normality**
- **Estimates of** $Var(\hat{y}(direct))$ **are often imprecise but needed**

More like....

-a way to fill in missing data

Details

Assumptions Needed

- **Although data-based, the model may still not fit well for some small areas**
- $\hat{y}(direct)$ **usually needs further model assumptions to be implementable... e.g., Normality**
- **Estimates of** $Var(\hat{y}(direct))$ **are often imprecise but needed**

More like....

-a way to fill in missing data

.... combine direct data with a model ... gets better with more direct data

Some current projects at NCHS - in order of maturity

1. **County estimates of smoking and cancer screening rates**
2. **State and sub-state estimates of people who use only wireless phones**
3. **Fast screening for outcomes that vary by small area**
4. **Small Area Estimates from the NHIS utilizing block-linked American Community Survey data**
5. **Some preliminary work on model-based estimates using Health care data**

County Estimates of Smoking and cancer screening rates

- **Combine NHIS county estimates by telephone status with BRFSS estimates**
- **Strengthen county estimate using associations with socio-demographic variables**
- **Estimates for 2000-2003 and 1997-1999 available online from the National Cancer Institute http://sae.cancer.gov**
- **Current estimates under development - modifying method to account for cell-phone only population**

State and sub-state estimates of people who use only wireless phones

- **Combine NHIS estimates of wireless rates with rates measured at other times**
- **Strengthen this component with state and substate estimate obtained form the American Community Survey**
- **Current method estimates:**
  2011:
  http://www.cdc.gov/nchs/data/nhsr/nhsr039.pdf
  2012:
  http://www.cdc.gov/nchs/data/nhsr/nhsr061.pdf
  - used to benchmark mixed-frame telephone surveys
- **Relatively new - possible improvement being investigated**

Fast screening for outcomes that vary by small area

- Small Area Estimation requires resources: analyst time, evaluation and review time
- Project based on premise that it is easier to estimate the small area variability than it is to estimate each individual small area
- Method uses simple model with no covariates
- Evaluation so far - discern among NHIS health insurance outcomes at the state level
- Method will break down if little data is available ANYWHERE - currently investigating when this happens

CDC

Small Area Estimates from the NHIS utilizing block-linked American Community Survey data

- **Working with the U.S. Census Bureau to create a NHIS/ACS file at the block-level and develop Small Area Estimates**
- **ACS: detailed estimates of health insurance, overall health, socio-economic variables**
- **Aim: use ACS estimates as covariates to create "NHIS like" estimates for small areas**
- **Targets: health insurance and access to care outcomes for states and the border counties of Mexico**

Other Uses of Small Area Methodology

- **Provide modeling ideas that can be used to analyze health outcomes and their interactions over geography**
- **The “synthetic data” approach to disclosure avoidance is often based on small area modelling. Small Area Estimates, themselves, will provide more disclosure avoidance than the original estimates.**
- **Some of the small area methodology research involves finding more accurate methods for incorporating the sample design into modelling**
- **Local Communities: Have options of using available small area estimates as an additional component to their local data**

Thank you.
dmalec@cdc.gov