Geospatial Data Methods for Estimating Population Health Outcomes

James B Holt, PhD
Xingyou Zhang, PhD

Epidemiology and Surveillance Branch
Division of Population Health/CDC

CSTE Webinar
Atlanta, GA
August 21, 2013
Outline

- Need for sub-state population health data
- General approaches to obtain sub-state population health data
- Model-based small area estimation
- Two examples of how this may work
- Limitations and appropriate uses
- Future developments and conclusion
- Questions
Need for Sub-state Population Health Data

- Epidemiology and surveillance
- Policymaking and evaluation
- Programmatic prioritization, resource allocation, and evaluation
Epidemiology and Surveillance

- **Small area population health data could help**
  - Evaluate local disease burden
  - Monitor local population health trends
  - Understand local geographic variations in population health status

- **Example**
  - Estimate neighborhood (census block group level) childhood obesity prevalence
  - A request from a national, nonprofit, land conservation organization, The Trust for Public Land
Policy

- Small area population health data could help
  - Inform policy making
  - Facilitate policy deliberation, formulation and delivery

- Example
  - Estimate the prevalence of chronic conditions and preventive service coverage for 435 US Congressional Districts (CDs)
  - A request from HHS/ASPE/Office of Health Policy for informing the potential impact of some human resource provisions, July 2009.
Programmatic

- Small area population health data are often needed for
  - Health program planning
  - Health program monitoring and evaluation

- Example
  - Estimate county-level population health measures for the Center for Medicare and Medicaid Services (CMS) $1 billion State Innovation Model initiative with an aim to advance community prevention and population health
General Approaches to Obtain Sub-state Population Health Data

- **National/state vital statistics system**
  - Birth data
  - Mortality data

- **National/state medical claims databases**
  - Medicare and MarketScan (private medical claims database)
  - Healthcare Cost and Utilization Project (HCUP) data

- **National/state health surveys [examples]**
  - Behavioral Risk Factor Surveillance System (BRFSS)
  - National Health Interview Survey (NHIS)
  - National Health and Nutrition Examination Survey (NHANES)
  - National Survey of Children's Health (NSCH)
  - California Health Interview Survey
Approaches to Obtain Sub-state Data Using National/State Health Surveys

- **Direct survey estimates**
  - Large sample size via survey
  - Temporal aggregation
  - Spatial aggregation
  - Spatial smoothing

- **Model-based small area estimation (SAE)**
Large Sample Size via Survey Design

- Obtain sufficient sample sizes in the Survey

- Examples
  - Selected Metropolitan/Micropolitan Area Risk Trends (SMART)
    - BRFSS selected metropolitan and micropolitan statistical areas with 500 or more respondents.
      - BRFSS 2011 has 213 counties (out of 3143 US counties) with 500 or more respondents.
  - The Youth Risk Behavior Survey (YRBS) district data covering 21 counties or cities (2011)
    - Boston, MA
    - Charlotte-Mecklenburg, NC
    - Dallas, TX
    - District of Columbia
    - Houston, TX
    - Memphis, TN
    - Milwaukee, WI
    - Orange County, FL
    - Philadelphia, PA
    - San Diego, CA
    - Seattle, WA
    - Broward County, FL
    - Chicago, IL
    - Detroit, MI
    - Duval County, FL
    - Los Angeles, CA
    - Miami-Dade County, FL
    - New York City, NY
    - Palm Beach County, FL
    - San Bernardino, CA
    - San Francisco, CA
Temporal Aggregation

- Combining survey data from different times (years)

Example

- Estimate a county’s obesity prevalence via combining the survey respondents from multiple years of BRFSS data.
Spatial Aggregation

- Combining survey data from adjacent areas

Examples

- Estimate a county’s obesity prevalence via including the survey respondents from its neighboring counties in BRFSS data. OR…
- To construct new units of analysis based on counties: “8 Americas”

3141 -> 2072 individual or merged counties units:
- Total pop at least 10,000 males and 10,000 females
- Account for county boundary changes since 1980
Spatial Smoothing

Spatial Smoothing of Geographically Aggregated Data by borrowing information from adjacent areas

- **Parametric**
  - Bayesian hierarchical spatial modeling

- **Nonparametric**
  - Direct spatial averaging
  - Spatial kernel density
  - Weighted Headbanging (a median smoother)

**Example**
- Estimate a county’s obesity prevalence via averaging those of its neighboring counties.
  - Often used when individual survey data are not available.
Model-based Small Area Estimation

- Small area estimation brief review
- Multilevel small area estimation
- Two examples
  - Congressional districts COPD prevalence by Congressional districts (CDs) using BRFSS
  - Childhood obesity prevalence by census blockgroup using National Survey of Children’s Health
What is small area?

- “Any domain for which direct estimates of adequate precision cannot be produced” (Rao, 2003)

- Domain: demographic or geographic or both
  - State is a small area for NHANES but not for BRFSS
  - County is a small area for national/state health surveys.

- Small sample size or no sample in the domain of interest

SAE Review (2)

– Approaches for Small Area Estimation

• Survey design-based
  – Combining the same surveys temporally, spatially, and spatiotemporally, even combining different type surveys.

• (Spatial) Microsimulation

• Explicit small area models
  – borrow “strength/information” from related areas through linking models based on survey data and auxiliary data such as census data and administrative records
  – This can lead to more precise and more stable indirect estimates for the various small areas of interest.
– **Small area models**
  
  * Area level models
    
    link direct survey estimates for areas of interest with area-level auxiliary data
    
    \[
    y_i = \theta_i + e_i = x_i^\prime \beta + v_i + e_i
    \]

  * Unit level models
    
    link the survey unit-level (person or household) outcome with unit/area-specific auxiliary data
    
    \[
    y_{ij} = \theta_{ij} + e_{ij} = x_{ij}^\prime \beta + x_i^\prime \eta + v_i + e_{ij}
    \]

  * Generalized linear mixed models (GLMM)
    
    \[
    y = X\beta + Z\alpha + \varepsilon
    \]

    – Multilevel logisitic model
    – Multilevel Poisson model
Multilevel small area models

- Why multilevel?
  - Population health conditions and health behaviors

- Individual sociodemographics
  - Age, gender, race/ethnicity, education, income

- Contextual characteristics (e.g., state, county and local neighborhood)
  - Social, economic, political and cultural environments
  - Physical environment
Multilevel Regression and Poststratification (MRP): a Geospatial Perspective

- Multilevel Regression and Poststratification (MRP): theory and application
The very basic idea of extended MRP for SAE

**Health Survey**
- Health outcomes?
  - Observed

**Census**
- Health outcomes?
  - Missing

**MRP**

**SAE**
- Health outcomes?
  - Predicted

All have:
- Age
- Gender
- Race
- Ethnicity
- Location
Multilevel Regression and Poststratification (MRP)

Four basic steps

1. Model construction
2. Model prediction
3. Poststratification with census data
4. Model-based estimates internal and external validation

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model construction</td>
<td>Model prediction</td>
<td>poststratification</td>
<td>internal</td>
</tr>
<tr>
<td>survey data</td>
<td>Census population data</td>
<td>small area estimates</td>
<td>external</td>
</tr>
<tr>
<td>Prevalence model</td>
<td>Prediction model</td>
<td>Aggregation</td>
<td>Validation</td>
</tr>
</tbody>
</table>
SAE Example 1

- Estimate prevalence of chronic obstructive pulmonary disease (COPD) for 435 US Congressional Districts (CDs) using BRFSS 2011
Is a Congressional District a small area?

<table>
<thead>
<tr>
<th>Domain</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>population (2010)</td>
<td>526,283</td>
<td>705,974</td>
<td>989,415</td>
<td>708,377</td>
</tr>
<tr>
<td>land area (square miles)</td>
<td>10</td>
<td>2,092</td>
<td>570,641</td>
<td>8,119</td>
</tr>
</tbody>
</table>
Standard Hierarchy of Census Geographic Entities

http://www.census.gov/geo/reference/pdfs/geodiagram.pdf
Data Sources

- **BRFSS (2011)**
  - Population health outcome
    - Chronic obstructive pulmonary disease (COPD)
  - Demographics
    - Gender (male vs female)
    - 13 age group (18-24, 25-29, ..., 75-79, 80+)
    - 8 race/ethnicity group (Non-Hispanic (white, black, AIAN, Asian, NHPI, other single race, two more races) and Hispanic)
      - 2x13x8=208 subpopulation groups
  - Geographic Location
    - State and county of residence

- **American Community Survey (ACS) 2007-2011**
  - County and tract level poverty

- **Census2010**
  - Block population by age, gender, race/ethnicity matching BRFSS 208 demographic groups
Step 1 for SAE using BRFSS (prevalence model)

- Multilevel logistic regression model for COPD
  - COPD ~ gender + age + race/ethnicity
    + county poverty
    + county-level random effects
    + state-level random effects

  - The structure of multilevel statistical model needs to be determined on strong epidemiological grounds.

  - The multilevel model could borrow information from the whole sample as well as from other data sources

  - The estimation of multilevel model could be implemented in traditional likelihood-based approach or a full Bayesian approach.
proc glimmix data=brfss6 noclprint;
  class county state sex age race;
  model copd (descending)=sex age race poverty
                   /dist=binary solution;
  weight _LLCPWT_scaled;
  random state county(state) /solution;
  ods output
  parameterEstimates=out.beta_fixed
  SolutionR=out.beta_random;
run;
Step 2 for SAE using BRFSS (prediction model)

- Model prediction of census block subpopulation by gender, age, race and ethnicity
  
  - COPD $\sim$ gender + age + race/ethnicity
    + block (tract) poverty
    + county-level random effects
    + state-level random random effects

  - Block (tract) poverty was used in the prediction model to further adjust local poverty influence on population health outcomes.

  - Non-sampled county random effects were obtained from spatial smoothing its adjacent counties with random effects.

  - The expected COPD risk could be obtained for all demographic groups in all census blocks.
Step 3 for SAE using BRFSS

- **Poststratification with census data**
  - A COPD prevalence estimate for a census block is the population weighted prevalence of the predicted COPD prevalence for all 208 subpopulation groups within a census block.
  - Aggregate census block estimates to congressional districts
  - Generate uncertainties associated with small area estimates
    - Monte Carlo simulation could be used to estimate the standard errors, confidence intervals for all SAEs.
%MACRO SIMU(outcome=, year=);
%do i=1 %to 1000;
data temp1; set beta&year._&outcome;
  xbeta=int+int_se*rannor(ABS(INT(RAND('UNIFORM')*100000)))+age+age_se*rannor(ABS(INT(RAND('UNIFORM')*100000)))+sex+sex_se*rannor(ABS(INT(RAND('UNIFORM')*100000)))+race+race_se*rannor(ABS(INT(RAND('UNIFORM')*100000)))+county+county_se*rannor(ABS(INT(RAND('UNIFORM')*100000)))+state+state_se*rannor(ABS(INT(RAND('UNIFORM')*100000)))+poverty_rate*(poverty+poverty_se*rannor(ABS(INT(RAND('UNIFORM')*100000))));
  p=exp(xbeta)/(1+exp(xbeta));
keep block pop p;
run;

*summarize obesity by block (or county, or congressional district);
proc sql;
  create table temp2 as
    select distinct block, sum(pop) as pop, (sum(p*pop)/sum(pop))*100 as &outcome
    from temp1
    group by block;
quit;
proc append base=temp3 data=temp2 force;run;
%MEND;
Step 4 for Model-based SAE Validation

- **Internal validation**
  - The model-based estimates, when aggregated, should be consistent with the direct survey estimates at the geographic levels supported by original survey design.
  - The model-based estimates, whenever possible, should be expected to be consistent with the direct survey estimates for those geographic areas with large sample sizes and adequate precision.

- **External validation**
  - Compared to the direct more reliable and accurate estimates from local health surveys
  - Compared to the direct estimates from census (if possible)
Internal Validation: State-level COPD Prevalence

State-level Pearson Correlation Coefficient (rho=0.99)

State-level descriptive sample statistics

<table>
<thead>
<tr>
<th>Method</th>
<th>state</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-based</td>
<td>51</td>
<td>4.10</td>
<td>6.18</td>
<td>9.73</td>
<td>6.40</td>
<td>5.63</td>
</tr>
<tr>
<td>Survey-based</td>
<td>51</td>
<td>3.95</td>
<td>6.14</td>
<td>9.77</td>
<td>6.37</td>
<td>5.82</td>
</tr>
</tbody>
</table>
Internal Validation: County-level COPD Prevalence

County-level Pearson Correlation Coefficients

<table>
<thead>
<tr>
<th>County</th>
<th>rho</th>
<th>Sample limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>525</td>
<td>0.95</td>
<td>N&gt;=50 and CV&lt;=0.3</td>
</tr>
<tr>
<td>208</td>
<td>0.98</td>
<td>N&gt;=500 and CV&lt;=0.3</td>
</tr>
</tbody>
</table>
SAE Example 2

- Prevalence of local neighborhood (census block group level) childhood obesity
  - A request from a national, nonprofit, land conservation organization: The Trust for Public Land, 2009

- Multilevel Regression and Poststratification (MRP) using geocoded national health surveys
Geocoded National Health Surveys

- National Health and Nutrition Examination Survey (NHANES)

- National Health Care Surveys
  - National Ambulatory Medical Care Survey (NAMCS) and National Hospital Ambulatory Medical Care Survey (NHAMCS)
  - National Hospital Discharge Survey (NHDS)
  - National Nursing Home Survey (NNHS) and National Nursing Assistant Survey (NNAS)
  - National Home and Hospice Care Survey (NHHCS) and National Home Health Aide Survey (NHHAS)
  - National Survey of Residential Care Facilities (NSRCF)

- National Health Interview Survey (NHIS)
- National Immunization Survey (NIS)
- National Survey of Family Growth (NSFG)

- State and Local Area Integrated Telephone Survey (SLAITS)
  - National Survey of Children’s Health (NSCH)
  - National Survey of Children with Special Health Care Needs (CSHCN)

- See details at http://www.cdc.gov/rdc/B1DataType/Dt122.htm
Data Sources for Childhood Obesity Estimates

- **NSCH (2007)**
  - Population health outcome
    - A child was considered obese if his or her body mass index (kg/m²) was equal to or greater than the sex- and age-specific 95th percentile on the CDC 2000 growth charts
  - Demographics
    - Gender (male vs female)
    - 2 age group (10-14, 15-17)
    - 8 race/ethnicity group (NH (white, black, AIAN, Asian, NHPI, other single race, two more races) and Hispanic)
  - Geography of sampling
    - 50 states and DC
      - Sample size ranges from 736 (NV) to 947 (ND) with a median of 876 (VT) and a mean of 865
    - 2,618 counties
    - 13,129 ZIP Codes
Data Sources for Childhood Obesity Estimates (continued)

- **ACS 2007-2011**
  - County and block group (tract) level poverty

- **ESRI Demographics 2010**
  - Median household income
  - Block group, ZIP Code, and county

- **ESRI Tapestry Segmentation Dataset 2010**
  - Lifestyle and urbanization
  - Block group, ZIP Code, and county

- **Census2010**
  - Block group level population by age, gender, race/ethnicity
Step 1 for SAE using NSCH (prevalence model)

- **Model construction and comparison**
  - NSCH child obesity status (yes or no) ~
  - sex + age + race/ethnicity (individual level)
  - + income + lifestyle + urbanization (ZIP Code level)
  - + income + urban-rural (county-level)
  - + random effects (ZIP Code, county and state levels)
### Proportion of Area-Level Variance for Childhood Obesity in Null Models Explained by Fixed Effects in Full Multilevel Models by Type of Geographic Area

<table>
<thead>
<tr>
<th>Model</th>
<th>Random Effects</th>
<th>Area Variance Null Model (SE)</th>
<th>Area Variance Full Model (SE)</th>
<th>Area-Level Variance Explained, %*</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>ZIP Code</td>
<td>0.1950 (0.03)</td>
<td>0.0134 (0.02)</td>
<td>93.1</td>
</tr>
<tr>
<td>II</td>
<td>State</td>
<td>0.0526 (0.01)</td>
<td>0.0211 (0.01)</td>
<td>60.0</td>
</tr>
<tr>
<td></td>
<td>County (state)</td>
<td>0.0456 (0.01)</td>
<td>0.0061 (0.01)</td>
<td>86.5</td>
</tr>
</tbody>
</table>

Null models are the models with random effects only.

Full models are the models with random effects as well as fixed effects including age, sex, race/ethnicity, ZIP Code level median household income, lifestyle and urbanity, and county level urban-rural status and median household income.

* Area-level variance explained = 1 - [area-level variance (null model) ÷ area-level variance (full model)]
Final prevalence model

- NSCH child obesity status (yes or no) ~
- sex + age + race/ethnicity (individual level)
- + income + lifestyle + urbanization (ZIP Code level)
- + income + urban-rural (county-level)
- + state-level random effects
Step 2 for SAE using NSCH (prediction model)

- **Model prediction**
  - A child’s *predicted* obesity risk (probability) ~
  - sex + age + race/ethnicity (individual level)
  - + income + lifestyle + urbanization (*block group* level)
  - + income + urban-rural (*county-level*)
  - + random effects (*state level*)

The expected childhood obesity risk could be obtained for all 32 demographic groups in all census block groups.
Step 3 for SAE using NSCH

- **Poststratification with census data**
  - The childhood obesity prevalence for all census block groups is the population weighted prevalence of the predicted childhood prevalence for all 32 subpopulation groups within a census block group.
  - Aggregate block group level estimates to larger geographic units
    - census tract, county, and state
  - Generate uncertainties associated with small area estimates
    - Monte Carlo simulation could be used to estimate the standard errors, confidence intervals for all SAEs.
Step 4 for SAE using NSCH

- Internal Validation (national- and state-level)

  - The national model-based childhood obesity estimate of 16.8% obesity among children aged 10 to 17 years was a nonsignificant 0.4 percentage points higher than the estimate based on direct survey (16.4%, <2.5% difference).

  - At the state level, the observed childhood obesity prevalence ranged from 9.6% (Oregon) to 21.9% (Mississippi). Compared with these direct state-level estimates, the model-based estimates for each state fell within the 95% confidence intervals (CIs).

  - Paired \( t \)-tests showed no significant difference between direct-survey and model-based estimates.
Step 4 for SAE using NSCH

- Internal Validation (county level)

The relationship between correlation coefficients of model-based and direct survey estimates and minimum county sample size.
Step 4 for SAE using NSCH

- **External Validation (state)**

  - When we compared state-level model-based estimates for children aged 15 to 17 years with the observed prevalence of obesity found by YRBS for schoolchildren in grades 9 through 12, the average model-based SAEs of obesity prevalence for states with YRBS estimates was 12.6% compared with 12.2% for YRBS.

  - A paired $t$ test showed no significant difference between these 2 sets of state-level estimates.
Multilevel Regression and Poststratification (MRP) : Advantages and Limitations

• Advantages
  o Reliable estimates for areas with small, or no samples
    • Estimates with high precision
  o Flexible in combining both individual and area-specific information relevant to small area estimation of outcomes of interest

• Disadvantages
  o Potential bias from model misspecification
    • Could-be lower accuracy
  o Model selection and validation could be challenging
Conclusions

- The model-based estimates in our studies were consistent with those direct survey estimates at both state and county levels.

- Our extended multilevel regression and poststratification (MRP) approach could be applied to geocoded national health surveys to estimate population health outcomes at different geographic domains in a scalable framework:
  - Census tracts, local neighborhoods, ZIP Codes, county, Congressional Districts, public health districts, voting districts and others
Conclusions

- This is one approach presented as an illustration of the feasibility of using national/state health surveys to estimate small area population health outcomes.

- Other methods for SAE exist and may be equally suitable or better, depending on specific topic, data source or needs.
Small area estimation for Using BRFSS (webinars)

- April 2013, Bayesian Small Area Estimation of Diabetes Prevalence by U.S. County, Ted Thompson
- May 2013, A SAS Small Area Estimation System for the BRFSS, Mike and Martin
- July 2013, Approaches on conducting small area estimation, Haci Akcin
- July 2013, Diabetes and Obesity Prevalence Estimates in Missouri Counties: Comparison of Missouri County-level Study and CDC’s Bayesian Model-based Approach, Shumei Yun
- August 2013, Rapid Response Health Surveillance and the Utility of Small Area Estimates, Haomiao Jia
Appropriate Uses

- **In application**
  - Understand the potential bias and precision of small area estimates

- **In presentation**
  - Always label “model-based”, particularly in maps
Future research

- **Multilevel analysis of complex survey data**
  - How construct and estimate multilevel model with complex survey data

- **Cross-scale inference with multilevel modeling**
  - What is potential bias from cross-scale inference for area-level variables

- **Geocoded national health surveys for small area estimation**
Acknowledgements

- Janet Croft
- Stephen Onufrak
- Hua Lu
- Yong Liu
- Earl Ford
- Anne Wheaton
- Kurt Greenlund
- CDC BRFSS
- CDC Research Data Center
Thank you!

James B. Holt
jgh4@cdc.gov

Xingyou Zhang
gyx8@cdc.gov

For more information please contact Centers for Disease Control and Prevention
1600 Clifton Road NE, Atlanta, GA 30333
Telephone, 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
E-mail: cdcinfo@cdc.gov  Web: www.cdc.gov

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.
Questions?
For more information please contact Centers for Disease Control and Prevention

1600 Clifton Road NE, Atlanta, GA 30333
Telephone: 1-800-CDC-INFO (232-4636)/TTY: 1-888-232-6348
E-mail: cdcinfo@cdc.gov Web: http://www.cdc.gov

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.