Missing Unit Problem in Population Health (and Social Science) Research

Rockli Kim, ScD
Postdoctoral Research Fellow
Harvard T.H. Chan School of Public Health
rok495@mail.harvard.edu

S V Subramanian, PhD
Professor of Population Health and Geography
Harvard University
svsubram@hsph.harvard.edu

Harvard Geography Colloquium
March 15, 2018
Outline

• Single level perspective: conceptual overview
  • Epidemiology
  • Sociology
  • Geography

• Multi-level perspective: examples
  • Multiple hierarchical (nested) geographies
  • Areal and spatial geographies
  • Cross-classified levels

• Conclusion
Single Level Perspective
Two Distinct Types of Etiological Questions

1. “Why do some individuals have hypertension?”
   Seeks the causes of cases (individual inference)

   OR

2. “Why do some populations have much hypertension, whilst in others it is rare?”
   Seeks the causes of incidence (population inference)
Single Level Perspective in Epidemiology

Determinants of **between-population variability** (Fairly predictable)

Determinants of **within-population variability**

---

Single Level Perspective in Epidemiology

Determinants of **between-population variability** (Fairly predictable)

Determinants of **within-population variability** (Almost impossible to predict)

---

Single Level Perspective in Epidemiology

If we only care about explaining mean differences between populations then why not just do aggregate (ecological) analysis

• The Seven Countries Study
  • 1958 – 1970
  • 7 countries: Yugoslavia, Italy, Greece, Finland, Netherlands, USA, Japan
  • Inferential Unit: Populations
  • Population variability: of substantive interest
  • Unit of analysis: Sites/Countries
Single Level Perspective
Single Level Perspective in Sociology

ECOLOGICAL CORRELATIONS AND THE BEHAVIOR OF INDIVIDUALS

W. S. Robinson
University of California at Los Angeles

- 3rd most cited paper in ASR (>5000 citations)
- Dire warnings of “ecological fallacy” - a cornerstone of ALL epidemiologic textbooks
- Motivated collection of individual survey data

### Single Level Perspective in Sociology

- Ecological Correlations and the Behavior of Individuals

<table>
<thead>
<tr>
<th>Individual</th>
<th>Illiteracy</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illiteracy</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.203</td>
<td>1</td>
</tr>
</tbody>
</table>

Single Level Perspective in Sociology

- Ecological Correlations and the Behavior of Individuals

<table>
<thead>
<tr>
<th>Individual</th>
<th>Illiteracy</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illiteracy</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>0.203</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>State</th>
<th>% Illiteracy</th>
<th>% Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Illiteracy</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>% Black</td>
<td>0.773</td>
<td>1</td>
</tr>
</tbody>
</table>

Single Level Perspective in Sociology

• Ecological Correlations and the Behavior of Individuals

<table>
<thead>
<tr>
<th></th>
<th>Illiteracy</th>
<th>Foreign-born</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illiteracy</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Foreign-born</td>
<td>0.118</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>% Illiteracy</th>
<th>% Foreign-born</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Illiteracy</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>% Foreign-born</td>
<td>-0.526</td>
<td>1</td>
</tr>
</tbody>
</table>

Single Level Perspective in Sociology

• On the ecological relationship
  • The purpose of this paper will have been accomplished if it prevents the future computation of meaningless correlations.
Single Level Perspective in Sociology

• On the ecological relationship
  • The purpose of this paper will have been accomplished if it prevents the future computation of meaningless correlations.

• On the individual relationship
  • The purpose of this paper will have been accomplished if it stimulates the study of similar problems with use of meaningful correlation between the properties of individuals.

Single Level Perspective
The Modifiable Areal Unit Problem

"the areal units used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating."

Cited ~2600 times
Geographical Variances

“It is sometimes asserted that geographical processes operate at different scales”

Cited ~190 times
Multi-Level Perspective

• Critical re-thinking of any single-level analyses: ecological or individual

• No longer need to choose A level of analysis: an inductive approach to ascertaining at what level does action lie

• Multilevel modeling more appropriate when...
  • Observations that is being analyzed are correlated/clustered;
  • Causal processes is thought to operate at more than one level; and/or
  • Intrinsic interest in modeling the variability and heterogeneity in the population.

Multi-Level Perspective in **Neighborhoods and Health Research**

Multi-Level Perspective in Neighborhoods and Health Research

• But still confined to two-level structure (94%)

• The importance of extended environments, whether modeled by including spatial relationships among neighborhood units, or situating neighborhoods in higher-level geographies, is lost under the typical approach to modeling local contexts

Table 3
Neighborhood level characteristics in 256 empirical quantitative studies of neighborhood effects and health.

<table>
<thead>
<tr>
<th>Multiple level of geographies</th>
<th>No. of studies</th>
<th>% of total studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 level</td>
<td>241</td>
<td>94.14</td>
</tr>
<tr>
<td>2 or more levels</td>
<td>15</td>
<td>5.86</td>
</tr>
<tr>
<td>Neighborhood definition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Census tracts</td>
<td>137</td>
<td>53.52</td>
</tr>
<tr>
<td>Block groups</td>
<td>52</td>
<td>20.31</td>
</tr>
<tr>
<td>Neighborhood clusters</td>
<td>20</td>
<td>7.81</td>
</tr>
<tr>
<td>ZIP codes</td>
<td>19</td>
<td>7.42</td>
</tr>
<tr>
<td>Others</td>
<td>17</td>
<td>6.64</td>
</tr>
<tr>
<td>More than one definition</td>
<td>10</td>
<td>3.91</td>
</tr>
<tr>
<td>No description</td>
<td>1</td>
<td>0.39</td>
</tr>
<tr>
<td>Is neighborhood geographic vs spatial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic</td>
<td>205</td>
<td>80.08</td>
</tr>
<tr>
<td>Spatial</td>
<td>14</td>
<td>5.47</td>
</tr>
<tr>
<td>Both</td>
<td>37</td>
<td>14.45</td>
</tr>
<tr>
<td>Is neighborhood variable proximity vs prevalence</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prevalence</td>
<td>221</td>
<td>90.23</td>
</tr>
<tr>
<td>Proximity</td>
<td>5</td>
<td>1.95</td>
</tr>
<tr>
<td>Both</td>
<td>20</td>
<td>7.81</td>
</tr>
<tr>
<td>Neighborhood level variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Census-based aggregated</td>
<td>110</td>
<td>42.97</td>
</tr>
<tr>
<td>Survey-based aggregated</td>
<td>31</td>
<td>12.11</td>
</tr>
<tr>
<td>Non-aggregated</td>
<td>14</td>
<td>5.47</td>
</tr>
<tr>
<td>Combination</td>
<td>98</td>
<td>38.28</td>
</tr>
<tr>
<td>Not reported</td>
<td>3</td>
<td>1.17</td>
</tr>
<tr>
<td>Explicit mention of MAUP/UGP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>246</td>
<td>96.09</td>
</tr>
<tr>
<td>UGP</td>
<td>2</td>
<td>0.78</td>
</tr>
<tr>
<td>MAUP</td>
<td>8</td>
<td>3.13</td>
</tr>
</tbody>
</table>

Multi-Level Perspective

- Need to recognize and explicitly model the reality that individuals belong to multiple settings that can affect their health

- Not a “Modifiable Areal Unit Problem” but a “Missing Unit Problem”

1. Importance of considering multiple (nested) geographies
1. Importance of considering multiple (nested) geographies

Example: Life Expectancy Patterns in the United States

Eight Americas: Investigating Mortality Disparities across Races, Counties, and Race-Counties in the United States

Christopher J. L. Murray1,2,3, Sandeep C. Kulkarni2,4, Catherine Michaud2,3, Niels Tomijima3, Maria T. Bulzacchelli3, Terrell J. Iandiorio5, Majid Ezzati1,2,*

1 Harvard School of Public Health, Boston, Massachusetts, United States of America, 2 Harvard University Initiative for Global Health, Cambridge, Massachusetts, United States of America, 3 Center for Population and Development Studies, Harvard University, Cambridge, Massachusetts, United States of America, 4 University of California San Francisco, San Francisco, California, United States of America

The Reversal of Fortunes: Trends in County Mortality and Cross-County Mortality Disparities in the United States

Majid Ezzati1,2,*, Ari B. Friedman2, Sandeep C. Kulkarni2,3, Christopher J. L. Murray1,2,4

1 Harvard School of Public Health, Boston, Massachusetts, United States of America, 2 Initiative for Global Health, Harvard University, Cambridge, Massachusetts, United States of America, 3 University of California, San Francisco, California, United States of America, 4 Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, United States of America
1. Importance of considering multiple (nested) geographies
Example: Life Expectancy Patterns in the United States

Data

- **Response**: Life expectancy (stratified by gender)
- **Predictor**: Time (i.e., “technological progress”)
- **Structure**
  - Repeated cross-section
  - Three-level: years (1961-1999) at level-1 (n=122,850) nested within 3,150 counties at level-2 nested within 51 states at level-3.
- **Model**: Three-level random coefficient model

1. Importance of considering multiple (nested) geographies

Example: Life Expectancy Patterns in the United States

- From 1961-1999:
  - Life expectancy increased from 67 to 74 years for men,
  - 74 to 79 years for women

<table>
<thead>
<tr>
<th></th>
<th>State VE (SE)</th>
<th>State % VPC</th>
<th>County VE (SE)</th>
<th>County % VPC</th>
<th>Time VE (SE)</th>
<th>Time % VPC</th>
<th>Total VE</th>
<th>Total % VPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Model 1</td>
<td>-</td>
<td>-</td>
<td>3.978 (0.101)</td>
<td>84.5%</td>
<td>0.727 (0.003)</td>
<td>15.5%</td>
<td>4.705</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Model 1: Time (level-1) nested within county (level-2); Model 2: Time (level-1) nested within state (level-2); Model 3: Time (level-1) nested within county (level-2) and state (level-3)*

1. Importance of considering multiple (nested) geographies
Example: Life Expectancy Patterns in the United States

• From 1961-1999:
  • Life expectancy increased from 67 to 74 years for men,
  • 74 to 79 years for women

<table>
<thead>
<tr>
<th></th>
<th>State</th>
<th>% VPC</th>
<th>County</th>
<th>% VPC</th>
<th>Time</th>
<th>% VPC</th>
<th>Total</th>
<th>% VPC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VE (SE)</td>
<td></td>
<td>VE (SE)</td>
<td></td>
<td>VE (SE)</td>
<td></td>
<td>VE</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-</td>
<td></td>
<td>3.978 (0.101)</td>
<td>84.5%</td>
<td>0.727 (0.003)</td>
<td>15.5%</td>
<td>4.705</td>
<td>100%</td>
</tr>
<tr>
<td>Model 2</td>
<td>3.066 (0.604)</td>
<td>55.0%</td>
<td></td>
<td></td>
<td>2.510 (0.010)</td>
<td>45.0%</td>
<td>5.577</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Model 1: Time (level-1) nested within county (level-2); Model 2: Time (level-1) nested within state (level-2); Model 3: Time (level-1) nested within county (level-2) and state (level-3)*
1. Importance of considering multiple (nested) geographies
Example: Life Expectancy Patterns in the United States

- From 1961-1999:
  - Life expectancy increased from 67 to 74 years for men,
  - 74 to 79 years for women

<table>
<thead>
<tr>
<th>State</th>
<th>County</th>
<th>Time</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>VE (SE)</td>
<td>% VPC</td>
<td>VE (SE)</td>
<td>% VPC</td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-</td>
<td>-</td>
<td>3.978 (0.101)</td>
</tr>
<tr>
<td>Model 2</td>
<td>3.066 (0.604)</td>
<td>55.0%</td>
<td>-</td>
</tr>
<tr>
<td>Model 3</td>
<td>2.419 (0.494)</td>
<td>48.8%</td>
<td>1.814 (0.047)</td>
</tr>
</tbody>
</table>

Model 1: Time (level-1) nested within county (level-2); Model 2: Time (level-1) nested within state (level-2); Model 3: Time (level-1) nested within county (level-2) and state (level-3)
1. Importance of considering multiple (nested) geographies
   Example: Life Expectancy Patterns in the United States

   • From 1961-1999:
     • Life expectancy increased from 67 to 74 years for men,
     • 74 to 79 years for women

<table>
<thead>
<tr>
<th></th>
<th>State</th>
<th>% VPC</th>
<th>County</th>
<th>% VPC</th>
<th>Time</th>
<th>% VPC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VE (SE)</td>
<td></td>
<td>VE (SE)</td>
<td></td>
<td>VE (SE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-</td>
<td>-</td>
<td>3.978</td>
<td>0.101</td>
<td>0.727</td>
<td>0.003</td>
<td>4.705</td>
</tr>
<tr>
<td>Model 2</td>
<td>3.066</td>
<td>0.604</td>
<td>-</td>
<td>-</td>
<td>2.510</td>
<td>0.010</td>
<td>5.577</td>
</tr>
<tr>
<td>Model 3</td>
<td>2.419</td>
<td>0.494</td>
<td>1.814</td>
<td>0.047</td>
<td>0.727</td>
<td>0.003</td>
<td>4.961</td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model 1</td>
<td>-</td>
<td>-</td>
<td>2.290</td>
<td>0.058</td>
<td>0.625</td>
<td>0.003</td>
<td>2.915</td>
</tr>
<tr>
<td>Model 2</td>
<td>1.524</td>
<td>0.301</td>
<td>-</td>
<td>-</td>
<td>1.717</td>
<td>0.007</td>
<td>3.241</td>
</tr>
<tr>
<td>Model 3</td>
<td>1.226</td>
<td>0.252</td>
<td>1.112</td>
<td>0.029</td>
<td>0.625</td>
<td>0.003</td>
<td>2.962</td>
</tr>
</tbody>
</table>

Model 1: Time (level-1) nested within county (level-2); Model 2: Time (level-1) nested within state (level-2); Model 3: Time (level-1) nested within county (level-2) and state (level-3)
1. Importance of considering multiple (nested) geographies
Example: Life Expectancy Patterns in the United States

County Effects on Male Life Expectancy

(ignoring State)

(accounting for State)

1. Importance of considering multiple (nested) geographies
Example: Life Expectancy Patterns in the United States

State Effects on Male Life Expectancy

(ignoring County) (accounting for County)

Red: Significantly below average
Blue: Significantly above average
Gray: No different from average

1. Importance of considering multiple (nested) geographies
   Example: Life Expectancy Patterns in the United States

   • There is a tendency to assume that a finer resolution of geographic aggregation (e.g., counties) is more important than a coarser resolution (e.g., states).

   • Prior studies have implicitly suggested that research and policy efforts should focus on the county-level processes and causes that might be the only drivers of longevity and premature mortality. However, we found that when simultaneously considered, states are as important as (if not more than) counties.

   • When geographic processes are likely to occur at multiple scales, empirical assessments should expand the units of analysis to accurately understand the scale at which action lies.
1. Importance of considering multiple (nested) geographies
1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India
1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India

Is this a fair depiction of poverty distribution in India?
1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

- Data: The National Sample Survey (2009-10, 2011-12)
- Response: Household poverty (based on monthly per capita expenditure)
- Predictors: Household type of residence, Household size, Caste, Religion, Primary source of income, Household land ownership, Sex, age, education level and marital status of the head of the household, Mean age of the household, The proportion of dependents
- Model: Five-level random intercept logistic models
  - $\text{logit}(\pi_{ijklm}) = \beta_0 + \beta X_{ijklm} + (g_{0m} + f_{0tm} + v_{0kltm} + u_{0jklm})$
  - $g_{0m} \sim N(0, \sigma^2_{g0}), f_{0tm} \sim N(0, \sigma^2_{f0}), v_{0kltm} \sim N(0, \sigma^2_{v0}), u_{0jklm} \sim N(0, \sigma^2_{u0})$
  - Level 1 variation approximated as $\pi^2/3$

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

Variance estimates in logit scale (95% CI) and proportion of total variation in poverty attributable to village-, district-, region- and state- levels in fully adjusted five-level and two-level models

<table>
<thead>
<tr>
<th>Five level model</th>
<th>Variance estimates (95% CI)</th>
<th>% Variance attributable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Village</td>
<td>0.590 (0.522, 0.659)</td>
<td>12.10</td>
</tr>
<tr>
<td>District</td>
<td>0.190 (0.128, 0.252)</td>
<td>3.90</td>
</tr>
<tr>
<td>Region</td>
<td>0.159 (0.092, 0.225)</td>
<td>3.25</td>
</tr>
<tr>
<td>State</td>
<td>0.647 (0.271, 1.023)</td>
<td>13.27</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Two level models</th>
<th>Variance estimates (95% CI)</th>
<th>% Variance attributable</th>
</tr>
</thead>
</table>
1. Importance of considering multiple (nested) geographies
   Example: Geographies of Poverty in India

Variance estimates in logit scale (95% CI) and proportion of total variation in poverty attributable to village-, district-, region- and state- levels in fully adjusted five-level and two-level models

<table>
<thead>
<tr>
<th>Five level model</th>
<th>Two level models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Variance estimates (95% CI)</strong></td>
<td><strong>Variance estimates (95% CI)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>% Variance attributable</strong></td>
<td><strong>% Variance attributable</strong></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Village</td>
<td>Village</td>
</tr>
<tr>
<td>0.590</td>
<td>2.097</td>
</tr>
<tr>
<td>(0.522, 0.659)</td>
<td>(2.038, 2.157)</td>
</tr>
<tr>
<td>12.10</td>
<td>38.93</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>District</td>
<td>District</td>
</tr>
<tr>
<td>0.190</td>
<td>0.769</td>
</tr>
<tr>
<td>(0.128, 0.252)</td>
<td>(0.680, 0.858)</td>
</tr>
<tr>
<td>3.90</td>
<td>18.94</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Region</td>
<td>Region</td>
</tr>
<tr>
<td>0.159</td>
<td>0.709</td>
</tr>
<tr>
<td>(0.092, 0.225)</td>
<td>(0.498, 0.921)</td>
</tr>
<tr>
<td>3.25</td>
<td>17.74</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>State</td>
<td>State</td>
</tr>
<tr>
<td>0.647</td>
<td>0.786</td>
</tr>
<tr>
<td>(0.271, 1.023)</td>
<td>(0.416, 1.157)</td>
</tr>
<tr>
<td>13.27</td>
<td>19.29</td>
</tr>
</tbody>
</table>

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

State effects (five level)

State effects (two level)

Red: Significantly high poverty
Blue: Significantly low poverty
Gray: No different from average

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

Region effects (five level)

Region effects (two level)

Red: Significantly high poverty
Blue: Significantly low poverty
Gray: No different from average
1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

District effects (five level)

District effects (two level)

Red: Significantly high poverty
Blue: Significantly low poverty
Gray: No different from average

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

- The political unit at which federal polices operate since liberalizations in the early 1990s
  - State-level reforms on industrial policy and investment incentives
  - Expansion of infrastructure investments
  - Investments in agricultural growth
  - Quality of governance (corruption and inefficient administration)

- Our multilevel analysis further supports prior state-level studies that have emphasized the importance of this unit

---

States

Household poverty

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

Cross-tabulation of district-level poverty by state-level poverty based on the five-level fully adjusted model

| District-level poverty N (% of total district) | State-level poverty
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Low</td>
<td>5</td>
</tr>
<tr>
<td>(0.80%)</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>37</td>
</tr>
<tr>
<td>(5.91%)</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>6</td>
</tr>
<tr>
<td>(0.96%)</td>
<td></td>
</tr>
</tbody>
</table>

But, single-level analysis on states alone do not capture the wide divergence that exists within states.

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

- Groups of contiguous districts within states
- Associated with agro-ecological conditions, which can aid or hinder the spread of irrigation and agricultural development
- We found relatively small between-region variation after accounting for other geographic levels and household composition
1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

- The lowest administrative unit at which demographic data are consistently found
- District councils plan the provision and implementation of diverse services
- Prior district-wise analyses of poverty have over-estimated the importance of district level

1. Importance of considering multiple (nested) geographies
   Example: Geographies of Poverty in India

<table>
<thead>
<tr>
<th>States</th>
<th>Regions</th>
<th>Districts</th>
<th>Villages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household poverty</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• The primary sampling units representing relatively homogenous local environment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Heterogeneous service delivery and quality of implementation of social programs may contribute to the large between-village variation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• <strong>First</strong> analysis to assess the importance of this micro level</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Importance of considering multiple (nested) geographies
Example: Geographies of Poverty in India

• Poverty is an **ecological construct** largely driven **macro- (states)** and **micro- (villages)** environments.

• The relative importance of one contextual level is **highly sensitive** to other units simultaneously considered.

• **A single-level perspective should be avoided** when planning to reduce poverty and promote balanced regional development.
2. Importance of considering spatial and areal geographies


*Everything is related to everything else, but near things are more related than distant things* (Tobler, 1970)
2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

• Both space and membership in geographically-embedded administrative units can produce variations in health, resulting in geographic clusters of good and poor health

• Objective:
  • Highlight that multiple forms of dependence may exist in geographically-referenced, hierarchical data, and that such clustering must be considered
  • Propose one method of integrating place and space perspectives when such clustering is present
2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

Model 1. Single-level
\( i = \text{county}, 1, \ldots, 2063 \)
\[ LE_i = \beta + (e_i), e_i \sim N(0, \sigma_e^2) \]

2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

Model 2. Two-level accounting for hierarchy

\( i = \text{county}, 1, \ldots, 2063 \) 
\( j = \text{state}, 1, \ldots, 49 \) 
\[ LE_{ij} = \beta + (u_j + e_{ij}), \]
\[ u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2) \]

Fig. 5. Model 2 county-level residuals exhibit spatial clustering after accounting for membership in states (Moran’s I = .289, 95% CI:.281 – .298).

2. Importance of considering spatial and areal geographies

Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

Model 2. Two-level accounting for hierarchy

(i = county, 1, ..., 2063)

(j = state, 1, ..., 49)

\[ LE_{ij} = \beta + (u_j + e_{ij}), \]

\[ u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2) \]

Fig. 6. Model 2 state-level residuals exhibit spatial clustering \((\text{Moran’s } I = .641, 95\% \text{ CI: .561 – .707})\).
2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

Model 3. Two-level accounting for spatial clustering

\[ LE_i = \beta + (e_{\text{neighbors}}(i) + e_i), \]
\[ e_{\text{neighbors}}(i) \sim N(\bar{e}_{\text{neighbors}}(i), \sigma^2_{(e_{\text{neighbors}})}), \]
\[ e_i \sim N(0, \sigma^2_e) \]

- Counties within “spatial patches”
- 2063 x 2063 spatial weight matrix
- First-order neighborhood structure using Queen-based contiguity
  - 0 indicates county pairs do not share a border and do not directly influence each other
- Conditional autoregressive was used to allow counties to be “cross-classified”

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>f</th>
<th>g</th>
<th>h</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

Model 3. Two-level accounting for spatial clustering

\[ LE_i = \beta + (e_{neighbors(i)} + e_i), \]
\[ e_{neighbors(i)} \sim N(\bar{e}_{neighbors(i)}, \sigma_{e_{neighbors}}^2/n_i), \]
\[ e_i \sim N(0, \sigma_e^2) \]

Fig. 7. Model 3 county-level residuals are spatially independent (Moran’s I = .008, 95% CI: -.021 – .038).

2. Importance of considering spatial and areal geographies

Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

**Model 4. Three-level accounting for hierarchy and spatial clustering**

\[ LE_{ij} = \beta + (u_j + e_{neighbors}(i) + e_{ij}), \]

\[ u_j \sim N(0, \sigma_u^2), \]

\[ e_{neighbors}(i) \sim N(\bar{e}_{neighbors}(i), \sigma_{(e,neighbors)}^2/n_i), \]

\[ e_i \sim N(0, \sigma_e^2) \]

---

**Fig. 8.** Model 4 county-level residuals are spatially independent (Moran’s I = .006, 95% CI: −.022 − .036).
2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

Model 4. Three-level accounting for hierarchy and spatial clustering

\[ LE_{ij} = \beta + (u_j + e_{neighbors(i)} + e_{ij}), \]
\[ u_j \sim N(0, \sigma_u^2), \]
\[ e_{neighbors(i)} \sim N(\bar{e}_{neighbors(i)}, \sigma_{(e,neighbors)}/n_i), \]
\[ e_i \sim N(0, \sigma_e^2) \]

Fig. 9. Model 4 state-level residuals are spatially independent (Moran’s I = -.028, 95% CI: -.180 – .152).
2. Importance of considering spatial and areal geographies
Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US

- Which model is the best?

| Table 2 |
|-----------------|-----------------|
| Bayesian deviance information criterion of model complexity and fit. | DIC |
| Model 1 | 8696.5 |
| Model 2 | 7367.0 |
| Model 3 | 6126.6 |
| Model 4 | 6106.2 |

- Integrating geographic membership and space adds enough information about the structure of county-level life expectancy to warrant Model 4’s complexity.

- Life expectancy is associated with both within-state features that are distinct from spatial proximity, and by separate spatial processes.
3. Importance of considering cross-classified levels

Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

- Cross-classification occurs when the two higher-level units are non-hierarchical
- Failing to account for cross-classified data structures will produce biased variance estimates, such that the variance associated with the omitted level will be attributed to the included level

- Objective:
  - How important are schools relative to neighborhoods in terms of the risks they confer on adolescents’ health?
  - Are the effects of neighborhoods (or schools) meaningful after adjusting for the other level?

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

• Data: National Longitudinal Study of Adolescent Health (AddHealth) Wave 1
  • 16,070 youth who attended 128 schools and lived in 2,111 census tracts

• Response: Smoking
  • Continuous measure (the number of days in past month, ranging from 1 to 30)
  • Binary measure (had ever smoked in the past month, yes/no)

• Predictors: Individual age, sex, public assistance, parental education, and race, School-level proportion on public assistance, Neighborhood-level proportion on public assistance

• Model: Cross-classified multilevel model
  \[ y_{ij(k)} = \beta_0 + \beta_{10i} + \beta_{20j} + \beta_{30k} + (u_{0j} + u_{0k} + e_{0i(jk)}) \]

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

- Average of 20.1 census tracts per school and 1.24 schools per census tract.

- 2,647 unique combinations of school and neighborhood, suggesting extensive cross-classification between the two levels.

- Fair amount of discordance between where students went to school and where they neighborhood, with respect to poverty.
  - ~25% live in incongruent settings (p<0.001 from McNemar χ2)
  - 13.87% lived in low poverty neighborhood, but attended a high poverty school
  - 10% lived in high poverty neighborhood, but attended a low poverty school

---

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

Outcome: the number of days smoked in past month

<table>
<thead>
<tr>
<th>Random effect estimates</th>
<th>Null Models</th>
<th>Adjusted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood</td>
<td>-</td>
<td>4.58 (3.66, 5.65)</td>
</tr>
<tr>
<td>School</td>
<td>5.44 (4.04, 7.19)</td>
<td>-</td>
</tr>
<tr>
<td>Individual</td>
<td>83.1 (81.3, 85.0)</td>
<td>84.0 (82.1, 85.9)</td>
</tr>
</tbody>
</table>

- Similar magnitude of variation across schools and neighborhoods in separate hierarchical models.
- Variance components significant in both models.

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

<table>
<thead>
<tr>
<th>Random effect estimates</th>
<th>Null Models</th>
<th>Adjusted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood</td>
<td>-</td>
<td>4.58 (3.66, 5.65)</td>
</tr>
<tr>
<td>School</td>
<td>5.44 (4.04, 7.19)</td>
<td>-</td>
</tr>
<tr>
<td>Individual</td>
<td>83.1 (81.3, 85.0)</td>
<td>84.0 (82.1, 85.9)</td>
</tr>
</tbody>
</table>

- Neighborhood variance no longer significant in cross-classified null model.
- Indicate that including only neighborhood substantially over-estimates the neighborhood-level variance.

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

Outcome: the number of days smoked in past month

<table>
<thead>
<tr>
<th>Random effect estimates</th>
<th>Null Models</th>
<th>Adjusted Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighborhood</td>
<td>-</td>
<td>4.58 (3.66, 5.65)</td>
</tr>
<tr>
<td>School</td>
<td>5.44 (4.04, 7.19)</td>
<td>-</td>
</tr>
<tr>
<td>Individual</td>
<td>83.1 (81.3, 85.0)</td>
<td>84.0 (82.1, 85.9)</td>
</tr>
</tbody>
</table>

- Variation at the school level persists even after adding in all covariates (age, parent on public assistance, and race) though the magnitude is reduced.

3. Importance of considering cross-classified levels
Example: Schools and Neighborhoods Effect on Adolescents’ Smoking Behaviors

• Particularly in the neighborhood-only models, variance associated with the omitted level (school) is incorrectly attributed to the included level (neighborhood).

• Results indicative of a “missing” level demonstrating that one context cannot substitute for another when individuals are simultaneously nested in multiple settings.

• Without concurrently examining the potential importance of one setting relative to another (cross-classification), we may be misestimating contextual effects and investing resources in the “wrong” context.

Conclusion

• Traditional analyses: a dilemma
  • EITHER Individual/Micro OR Aggregate/Macro

• Multilevel analysis: a solution
  • Consider the two simultaneously

• Unit of analysis matters
  • Omitting relevant level(s) can lead to errors in causal inference and misspecification of the model
  • Conceptual rigor in justifying the different units of analysis
Conclusion

• Multilevel perspective to understanding and explaining variability in outcomes

• Geographic clustering is not a “nuisance” to be simply corrected for, but a substantively important pattern that merits careful exploration
  • Modeling intra-class correlation
  • Modeling heterogeneity
  • Modeling complex data analytic structure
Rockli Kim
rok495@mail.harvard.edu

S V Subramanian
svsubram@hsph.harvard.edu