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

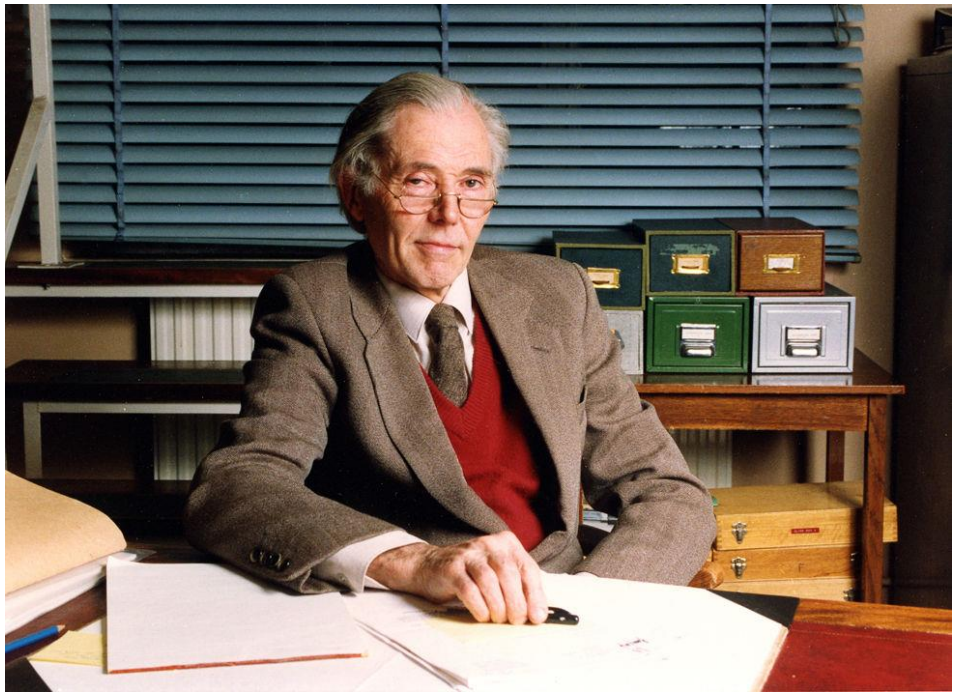
Harvard Geography Colloquium
March 15, 2018

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

Single Level Perspective in Epidemiology



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)

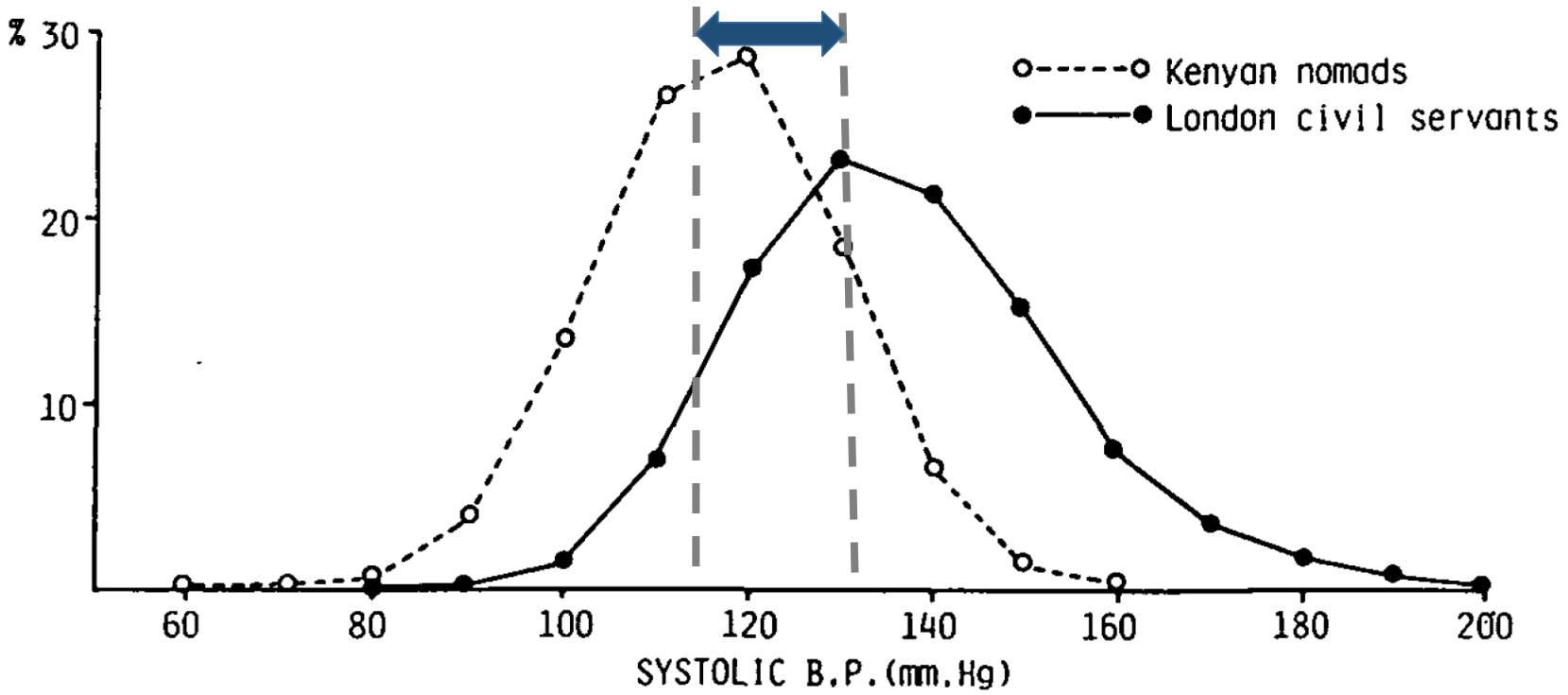
Rose, G. (1985). Sick Individuals and Sick Populations. International Journal of Epidemiology. 14 (1).

Single Level Perspective in Epidemiology

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

≠

Determinants of **within-population variability**



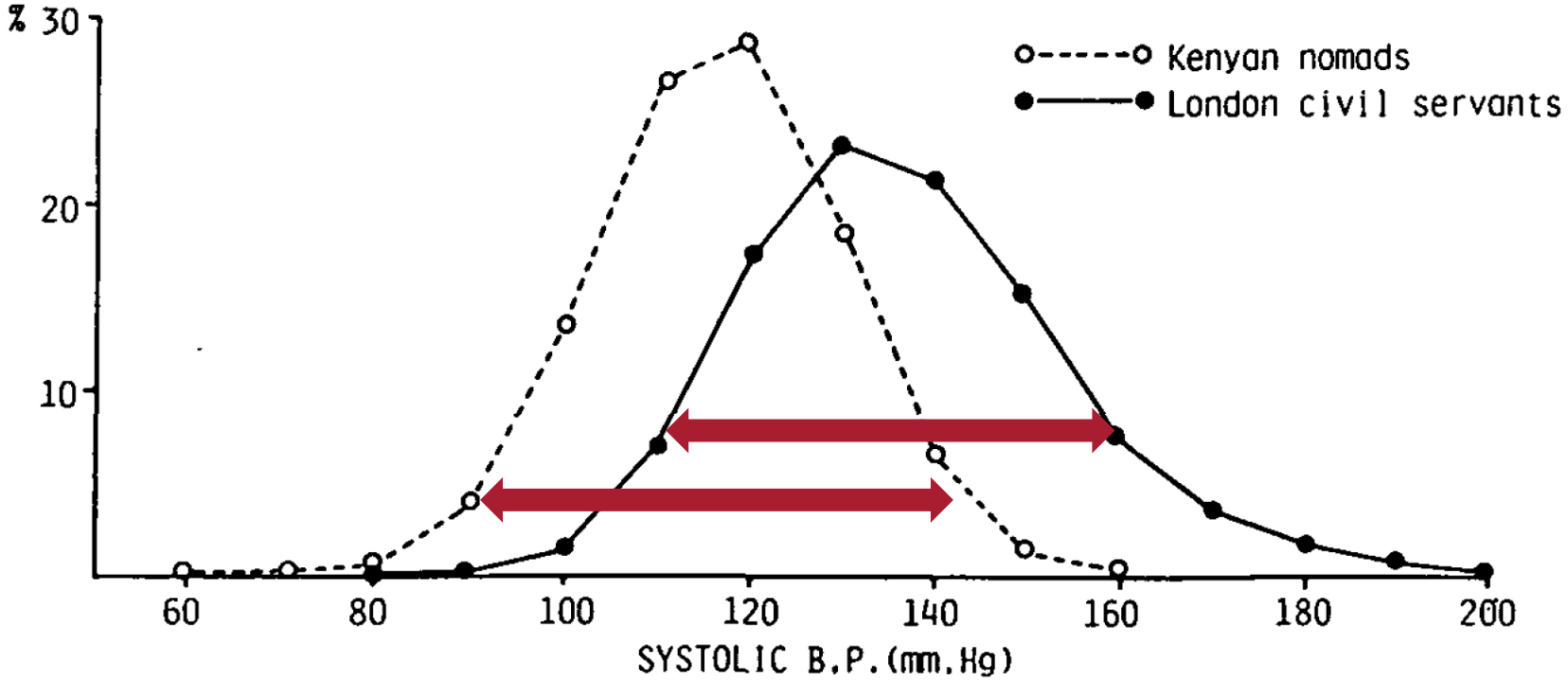
Rose, G. (1985). Sick Individuals and Sick Populations. International Journal of Epidemiology. 14 (1).

Single Level Perspective in Epidemiology

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

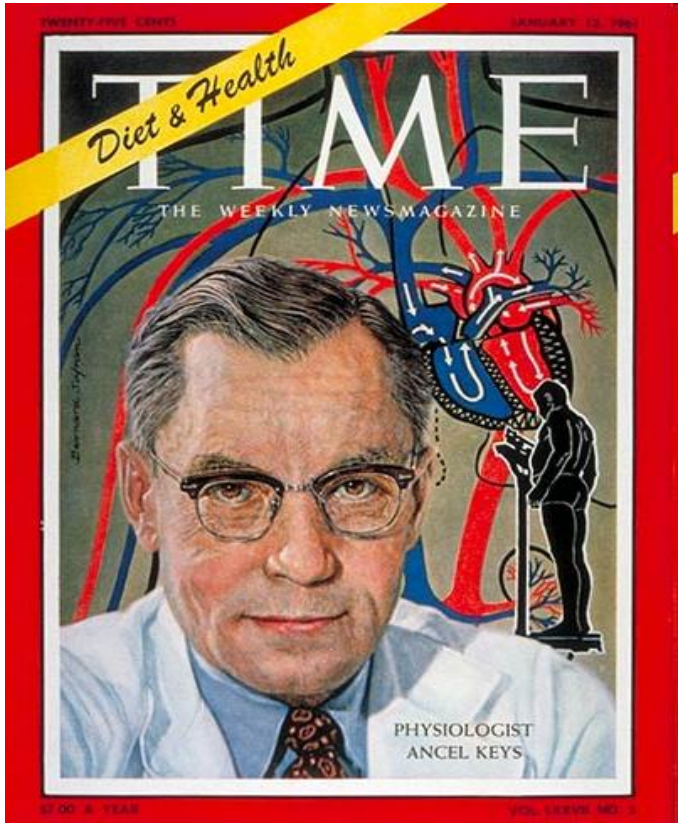
≠

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



Rose, G. (1985). Sick Individuals and Sick Populations. International Journal of Epidemiology. 14 (1).

Single Level Perspective in Epidemiology

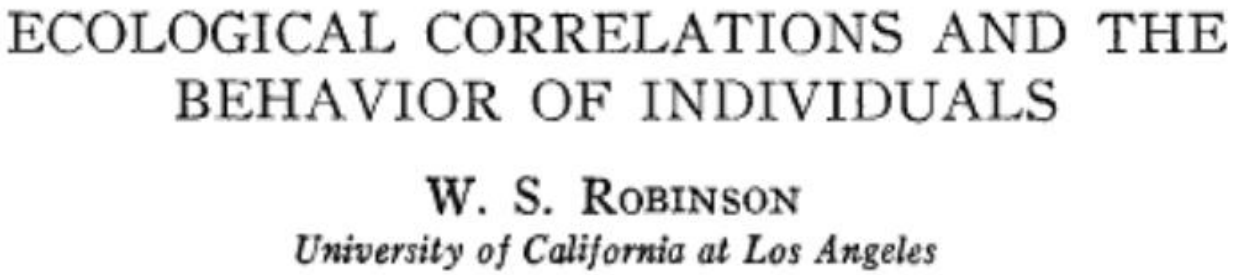


If we only care about explaining mean differences between populations then why not just do aggregate (ecological) analysis

- 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

Robinson. W.S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15(3), p 351-357

Single Level Perspective in Sociology

- Ecological Correlations and the Behavior of Individuals

Individual	Illiteracy	Black
Illiteracy	1	
Black	0.203	1

Robinson. W.S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15(3), p 351-357

Single Level Perspective in Sociology

- Ecological Correlations and the Behavior of Individuals

Individual	Illiteracy	Black
Illiteracy	1	
Black	0.203	1

State	% Illiteracy	% Black
% Illiteracy	1	
% Black	0.773	1

Robinson. W.S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15(3), p 351-357

Single Level Perspective in Sociology

- Ecological Correlations and the Behavior of Individuals

Individual	Illiteracy	Foreign-born
Illiteracy	1	
Foreign-born	0.118	1

State	% Illiteracy	% Foreign-born
% Illiteracy	1	
% Foreign-born	-0.526	1

Robinson. W.S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15(3), p 351-357

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.

Robinson. W.S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15(3), p 351-357

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.

Robinson. W.S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15(3), p 351-357

Single Level Perspective

Single Level Perspective in **Geography**

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

Openshaw, S. (1984). The modifiable areal unit problem. Concepts and techniques in modern geography.

Single Level Perspective in **Geography**

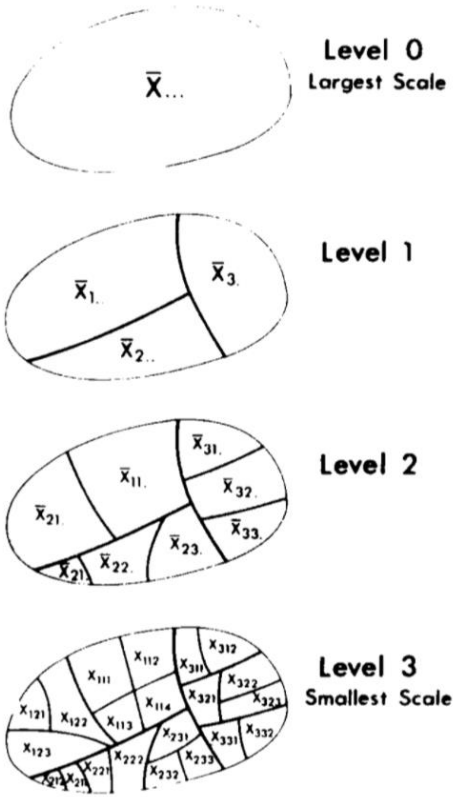


FIG. 2. An Illustration of the Nesting of an Irregular Hierarchy. The notation corresponds to that used in the text.

Geographical Variances

"It is sometimes asserted that geographical processes operate at different scales"

Cited ~190 times

Moellering, M. & Tobler, W. (1972). Geographical Variances. 4(1). p 34-50 Geographical Analysis.

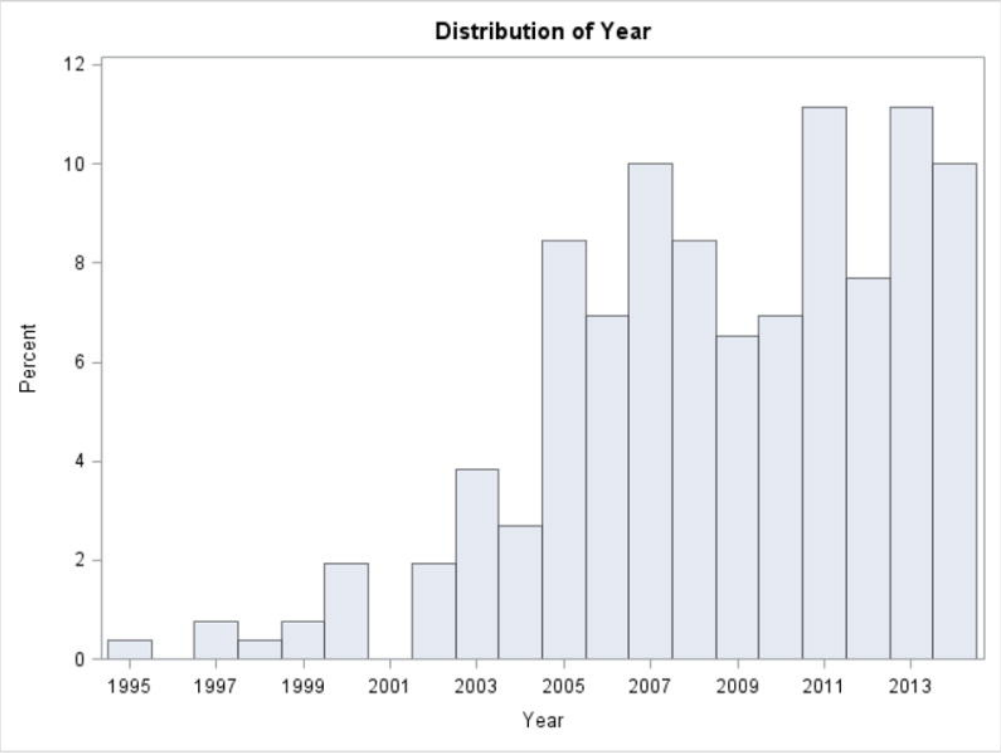
Multi-Level Perspective

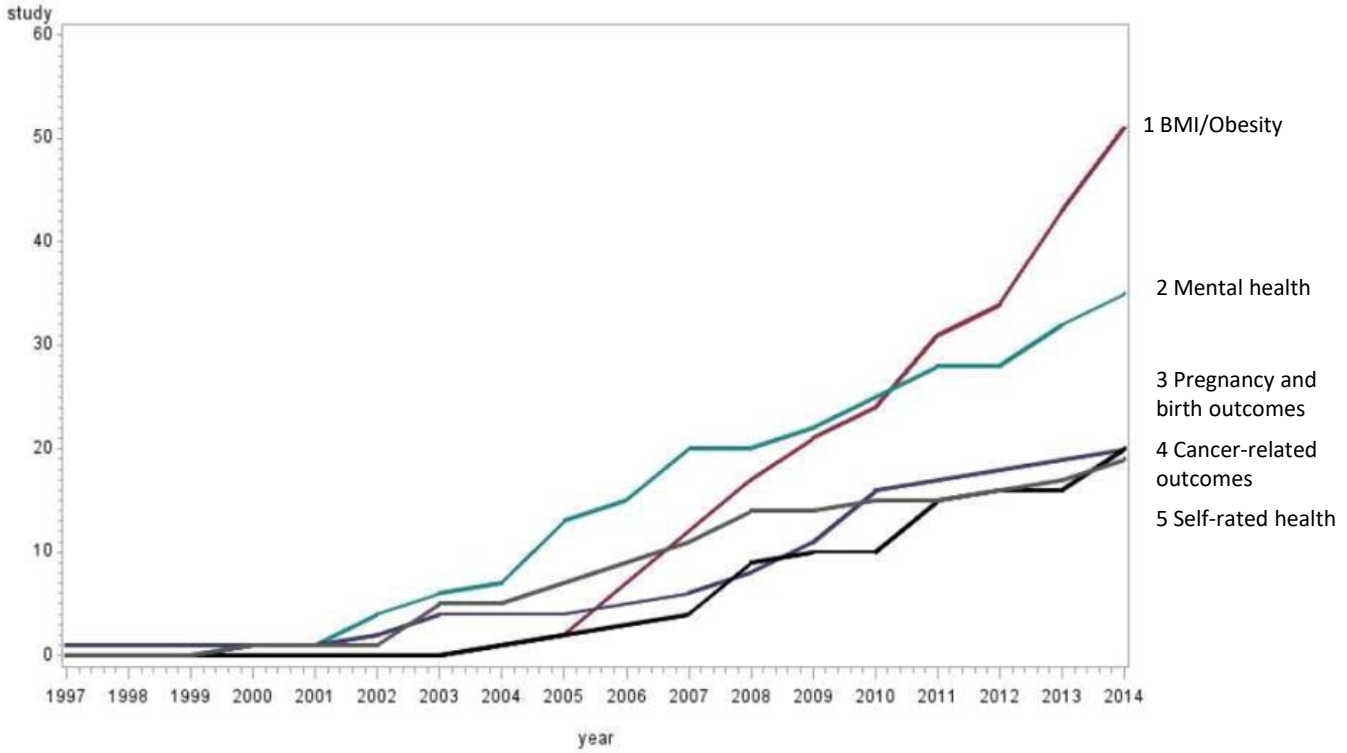
- Critical re-thinking of any single-level analyses: ecological or individual
- No longer need to choose **A** level of analysis: an inductive approach to ascertaining at what level does action lie
- Multilevel modeling more appropriate when…
  - Observations that is being analyzed are correlated/clustered;
  - Causal processes is thought to operate at more than one level; and/or
  - Intrinsic interest in modeling the variability and heterogeneity in the population.

Subramanian, S.V., Jones, K., Kaddour, A. and Krieger, N. (2009). Revisiting Robinson: the perils of individualistic and ecologic fallacy. International journal of epidemiology, 38(2), pp.342-360.

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

- Systematic review of 259 neighborhood effects on health research on US population (1995-2004)





Arcaya, M.C., Tucker-Seeley, R.D., Kim, R., Schnake-Mahl, A., So, M. and Subramanian, S.V. (2016). Research on neighborhood effects on health in the United States: a systematic review of study characteristics. Social Science & Medicine, 168, pp.16-29.

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

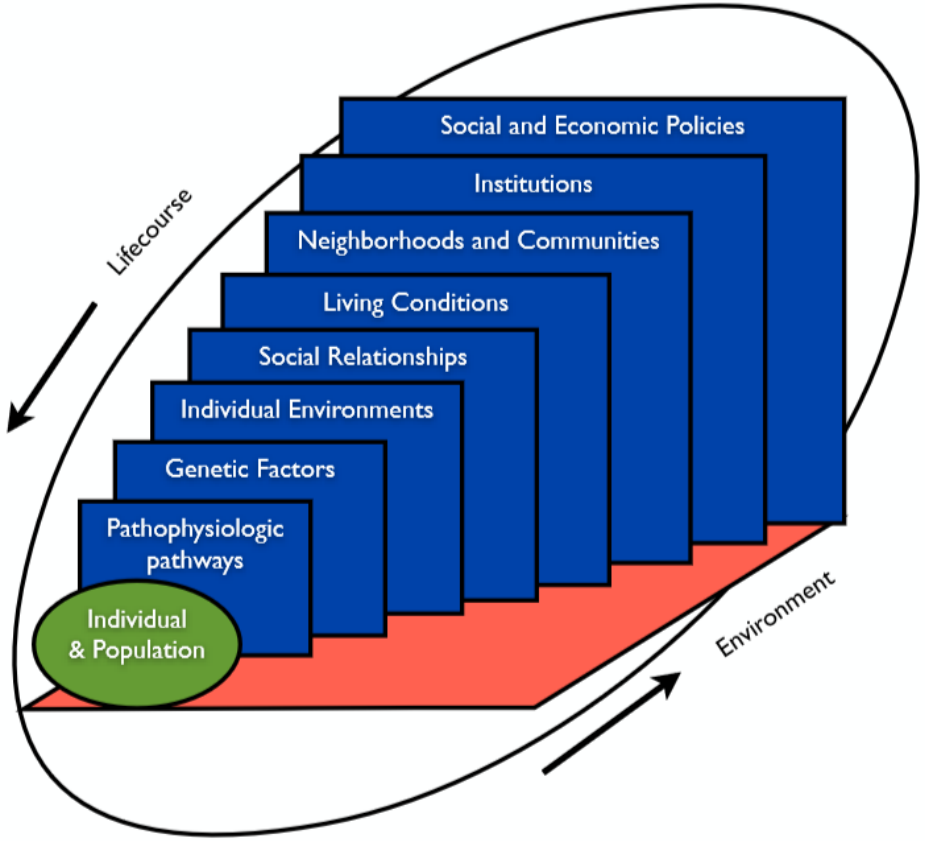
- But still confined to two-level structure (94%)
- The importance of extended environments, whether modeled by including spatial relationships among neighborhood units, or situating neighborhoods in higher-level geographies, is lost under the typical approach to modeling local contexts

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

	No. of studies	% of total studies
Multiple level of geographies		
1 level	241	94.14
2 or more levels	15	5.86
Neighborhood definition		
Census tracts	137	53.52
Block groups	52	20.31
Neighborhood clusters[a]	20	7.81
ZIP codes	19	7.42
Others[b]	17	6.64
More than one definition[c]	10	3.91
No description	1	0.39
Is neighborhood geographic vs spatial		
Geographic	205	80.08
Spatial	14	5.47
Both	37	14.45
Is neighborhood variable proximity vs prevalence		
Prevalence	231	90.23
Proximity	5	1.95
Both	20	7.81
Neighborhood level variables		
Census-based aggregated	110	42.97
Survey-based aggregated	31	12.11
Non-aggregated[d]	14	5.47
Combination[e]	98	38.28
Not reported	3	1.17
Explicit mention of MAUP/UGP		
None	246	96.09
UGP	2	0.78
MAUP	8	3.13

Arcaya, M.C., Tucker-Seeley, R.D., Kim, R., Schnake-Mahl, A., So, M. and Subramanian, S.V. (2016). Research on neighborhood effects on health in the United States: a systematic review of study characteristics. Social Science & Medicine, 168, pp.16-29.

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"

Kaplan, G. (2004). What's wrong with social epidemiology, and how can we make it better? Epidemiologic Reviews, 26, pp.124-135.

1. Importance of considering multiple (nested) geographies

1. Importance of considering multiple (nested) geographies

Example: Life Expectancy Patterns in the United States

OPEN ACCESS Freely available online

PLOS MEDICINE

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

Christopher J. L. Murray[1,2,3], Sandeep C. Kulkarni[2,4], Catherine Michaud[2,3], Niels Tomijima[3], Maria T. Bulzacchelli[3], Terrell J. Iandiorio[3], Majid Ezzati[1,2*]

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

ble online

PLOS MEDICINE

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

Majid Ezzati[1,2*], Ari B. Friedman[2], Sandeep C. Kulkarni[2,3], Christopher J. L. Murray[1,2,4]

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

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

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

Kim, R. & Subramanian, S.V. (2016). What’s Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

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

	State		County		Time		Total	
	VE (SE)	% VPC	VE (SE)	% VPC	VE (SE)	% VPC	VE	% VPC
Male								
Model 1	-	-	3.978 (0.101)	84.5%	0.727 (0.003)	15.5%	4.705	100%

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)

Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

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

	State		County		Time		Total	
	VE (SE)	% VPC	VE (SE)	% VPC	VE (SE)	% VPC	VE	% VPC
Male								
Model 1	-	-	3.978 (0.101)	84.5%	0.727 (0.003)	15.5%	4.705	100%
Model 2	3.066 (0.604)	55.0%	-	-	2.510 (0.010)	45.0%	5.577	100%

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)

Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

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

	State		County		Time		Total	
	VE (SE)	% VPC	VE (SE)	% VPC	VE (SE)	% VPC	VE	% VPC
Male								
Model 1	-	-	3.978 (0.101)	84.5%	0.727 (0.003)	15.5%	4.705	100%
Model 2	3.066 (0.604)	55.0%	-	-	2.510 (0.010)	45.0%	5.577	100%
Model 3	2.419 (0.494)	48.8%	1.814 (0.047)	36.6%	0.727 (0.003)	14.7%	4.961	100%

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)

Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

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

	State		County		Time		Total	
	VE (SE)	% VPC	VE (SE)	% VPC	VE (SE)	% VPC	VE	% VPC
Male								
Model 1	-	-	3.978 (0.101)	84.5%	0.727 (0.003)	15.5%	4.705	100%
Model 2	3.066 (0.604)	55.0%	-	-	2.510 (0.010)	45.0%	5.577	100%
Model 3	2.419 (0.494)	48.8%	1.814 (0.047)	36.6%	0.727 (0.003)	14.7%	4.961	100%
Female								
Model 1	-	-	2.290 (0.058)	78.6%	0.625 (0.003)	21.4%	2.915	100%
Model 2	1.524 (0.301)	47.0%	-	-	1.717 (0.007)	53.0%	3.241	100%
Model 3	1.226 (0.252)	41.4%	1.112 (0.029)	37.5%	0.625 (0.003)	21.1%	2.962	100%

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)

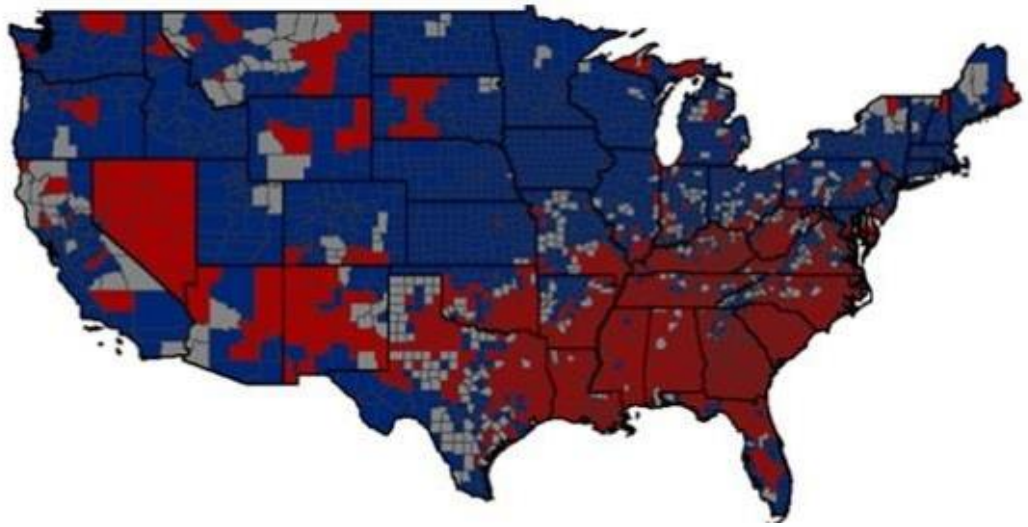
Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

1. Importance of considering multiple (nested) geographies

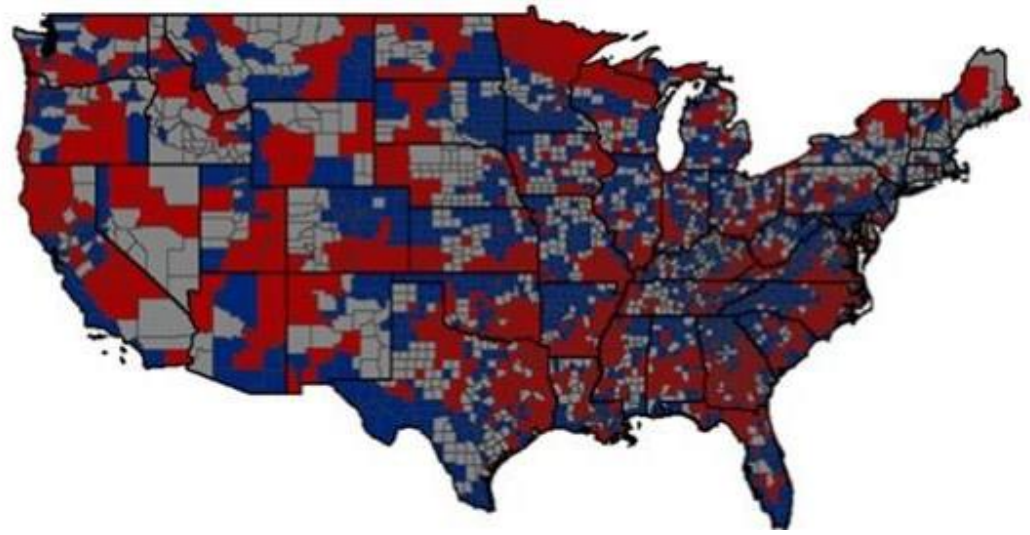
Example: Life Expectancy Patterns in the United States

County Effects on Male Life Expectancy

(ignoring State)



(accounting for State)



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

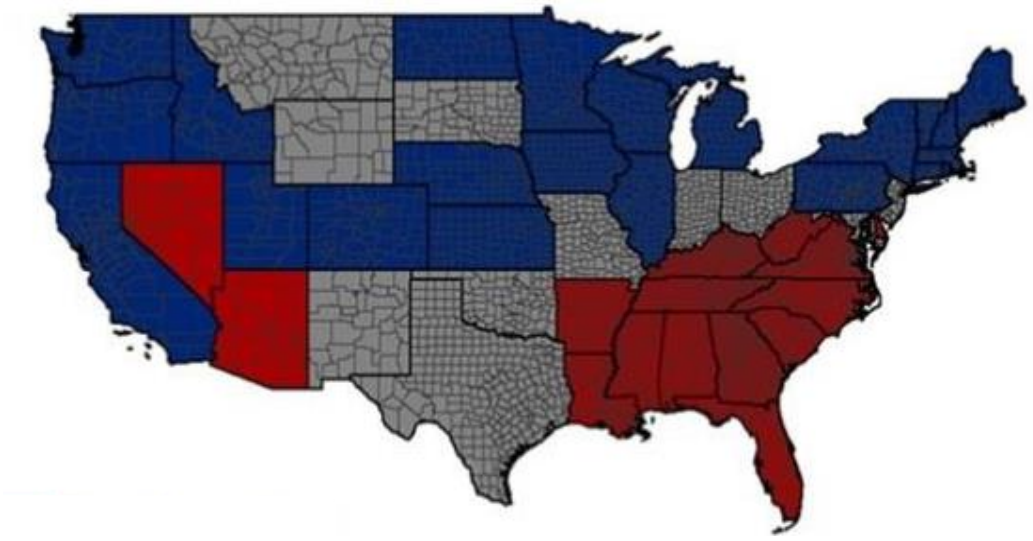
Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

1. Importance of considering multiple (nested) geographies

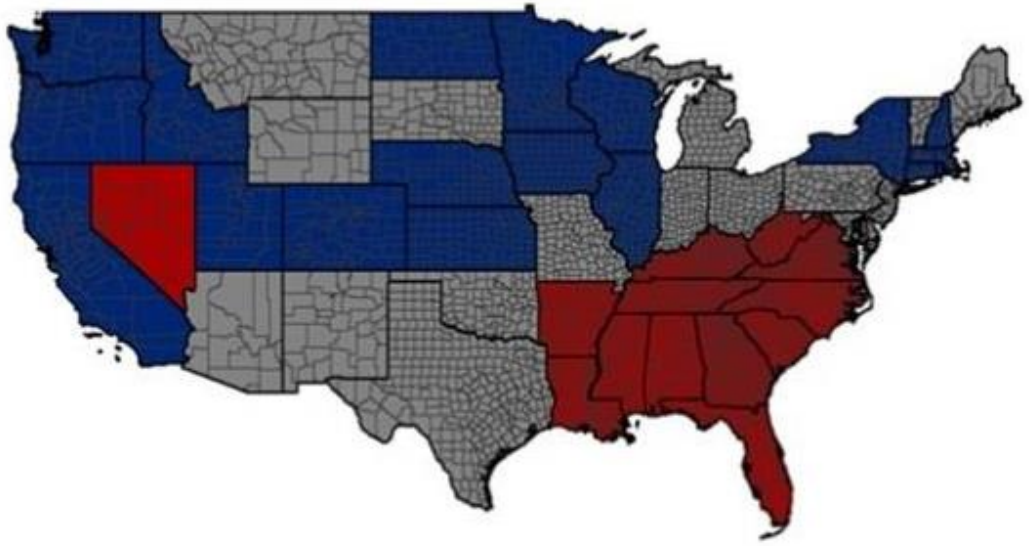
Example: Life Expectancy Patterns in the United States

State Effects on Male Life Expectancy

(ignoring County)



(accounting for County)



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

Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

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.

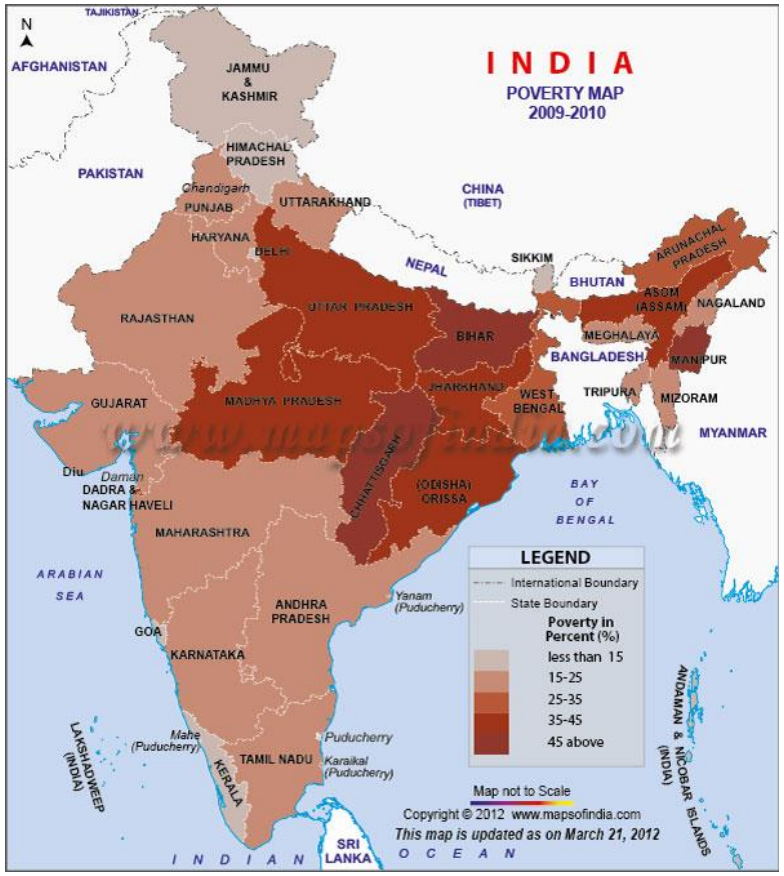
- When geographic processes are likely to occur at multiple scales, empirical assessments should expand the units of analysis to accurately understand the scale at which action lies.

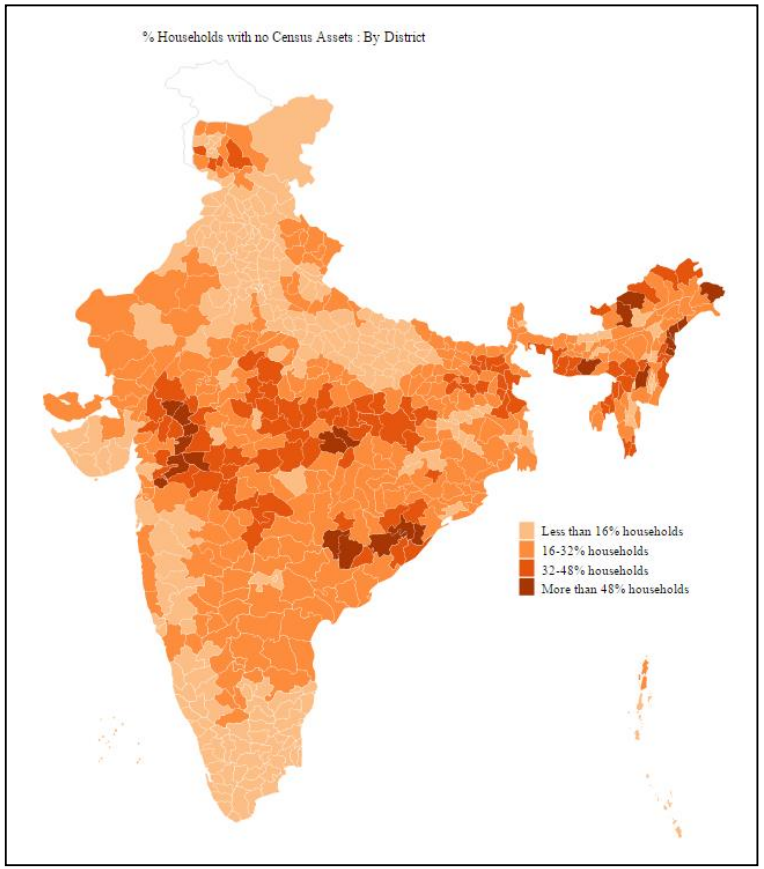
Kim, R. & Subramanian, S.V. (2016). What's Wrong with Understanding Variation Using a Single-Geographic Scale? A Multilevel Geographic Assessment of Life Expectancy in the United States. Procedia Environmental Sciences. 36. p 4-11

1. Importance of considering multiple (nested) geographies

1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India

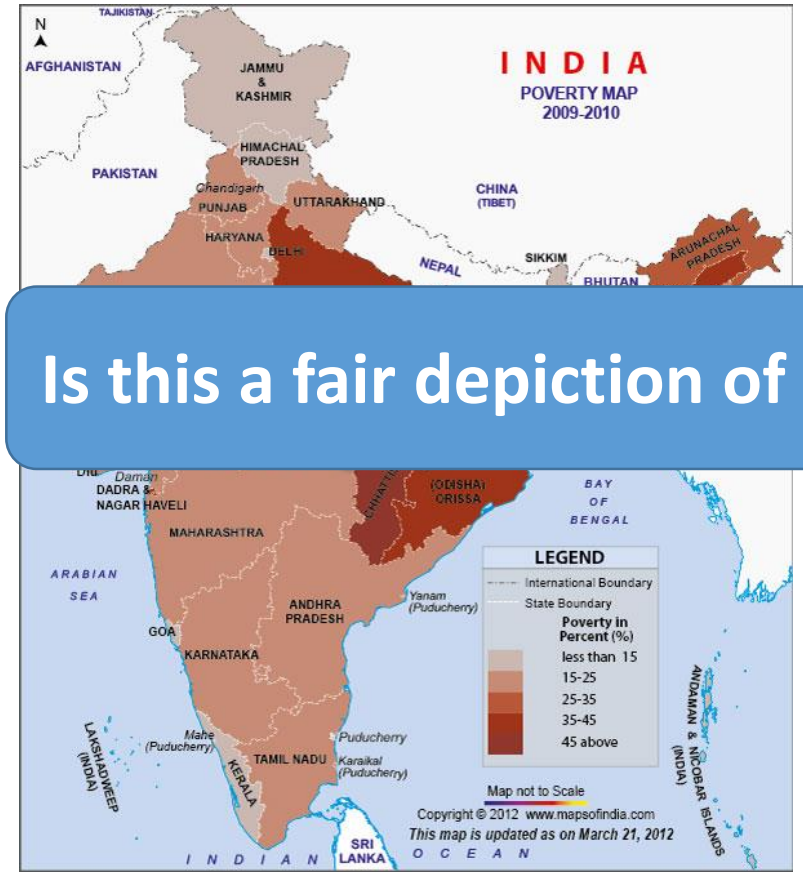


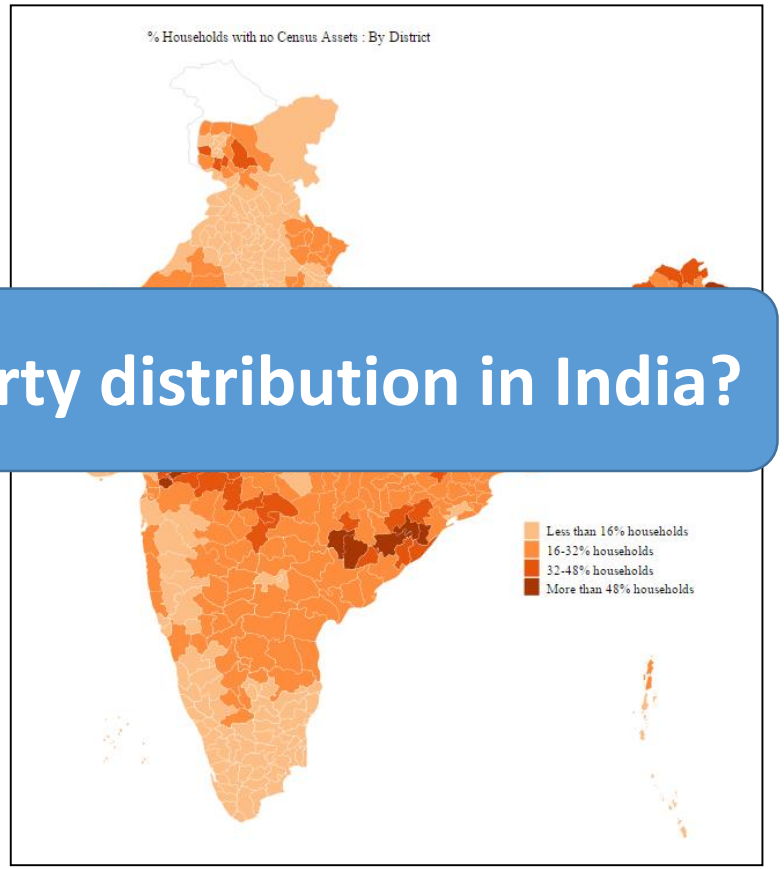


HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India

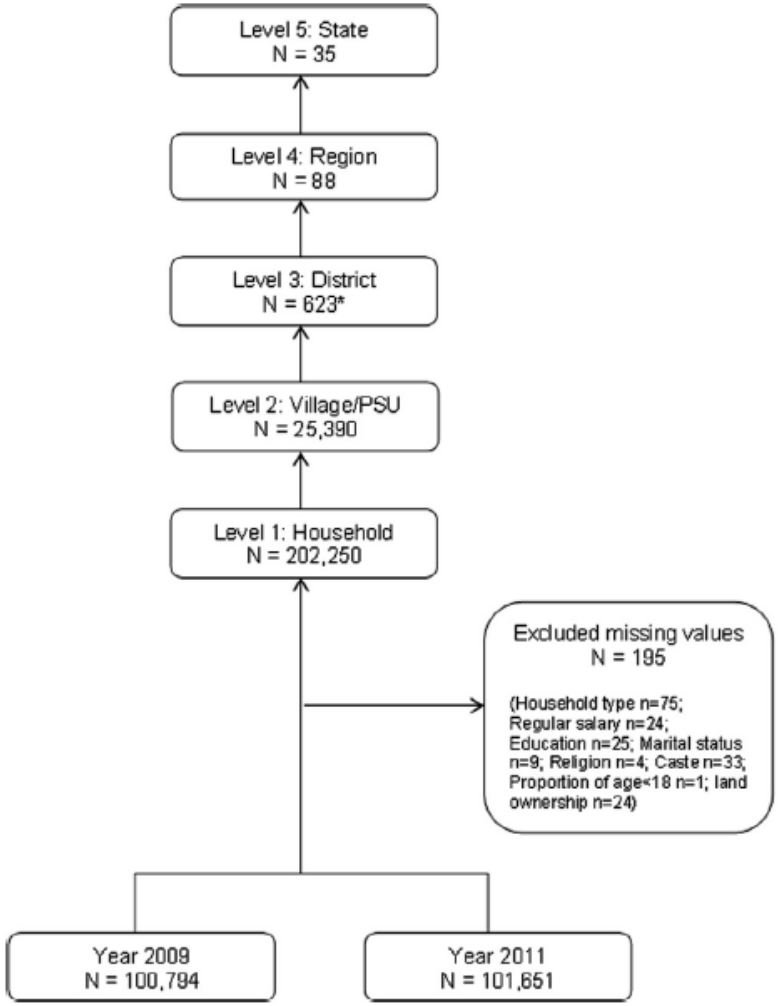




Is this a fair depiction of poverty distribution in India?

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
  - $logit(\pi_{ijklm}) = \beta_0 + \beta X'_{ijklm} + (g_{0m} + f_{0lm} + v_{0klm} + u_{0jklm})$
  - $g_{0m} \sim N(0, \sigma^2_{g0})$, $f_{0lm} \sim N(0, \sigma^2_{f0})$, $v_{0klm} \sim N(0, \sigma^2_{v0})$, $u_{0jklm} \sim N(0, \sigma^2_{u0})$
  - *Level 1 variation approximated as* $\pi^2/3$



Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

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

	Five level model		Two level models	
	Variance estimates (95% CI)	% Variance attributable	Variance estimates (95% CI)	% Variance attributable
Village	0.590 (0.522, 0.659)	12.10		
District	0.190 (0.128, 0.252)	3.90		
Region	0.159 (0.092, 0.225)	3.25		
State	0.647 (0.271, 1.023)	13.27		

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

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

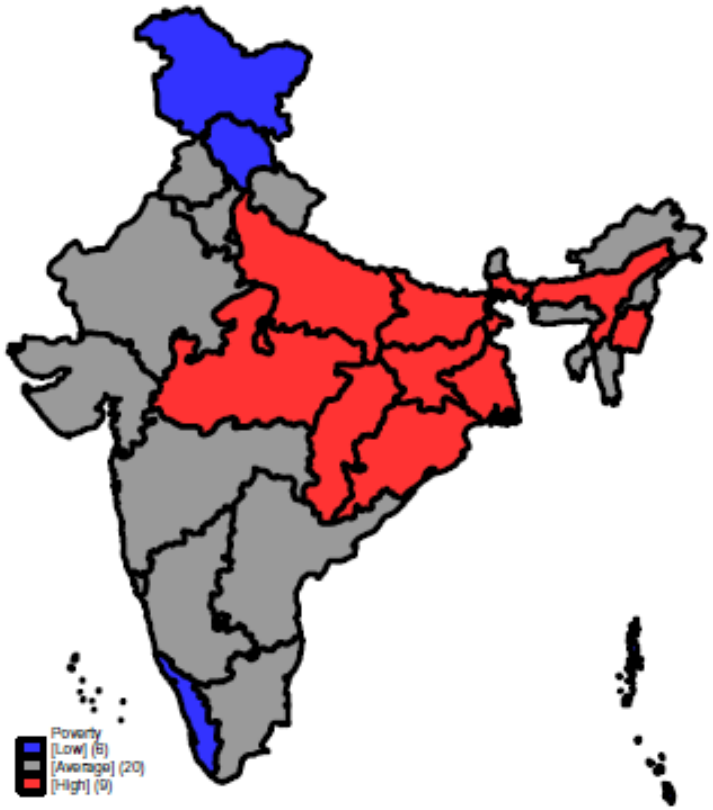
	Five level model		Two level models	
	Variance estimates (95% CI)	% Variance attributable	Variance estimates (95% CI)	% Variance attributable
Village	0.590 (0.522, 0.659)	12.10	2.097 (2.038, 2.157)	38.93
District	0.190 (0.128, 0.252)	3.90	0.769 (0.680, 0.858)	18.94
Region	0.159 (0.092, 0.225)	3.25	0.709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

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

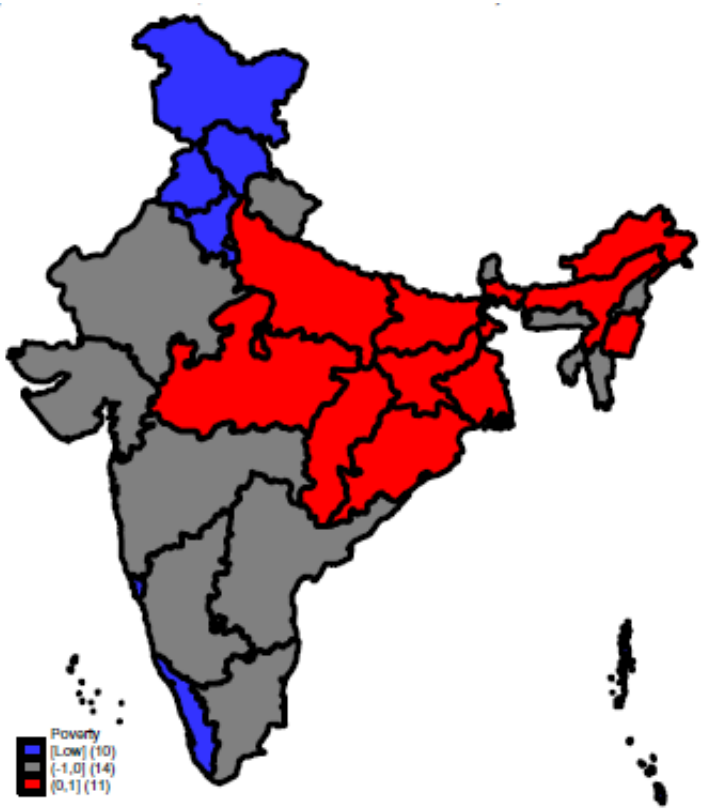
1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India

State effects (five level)



State effects (two level)



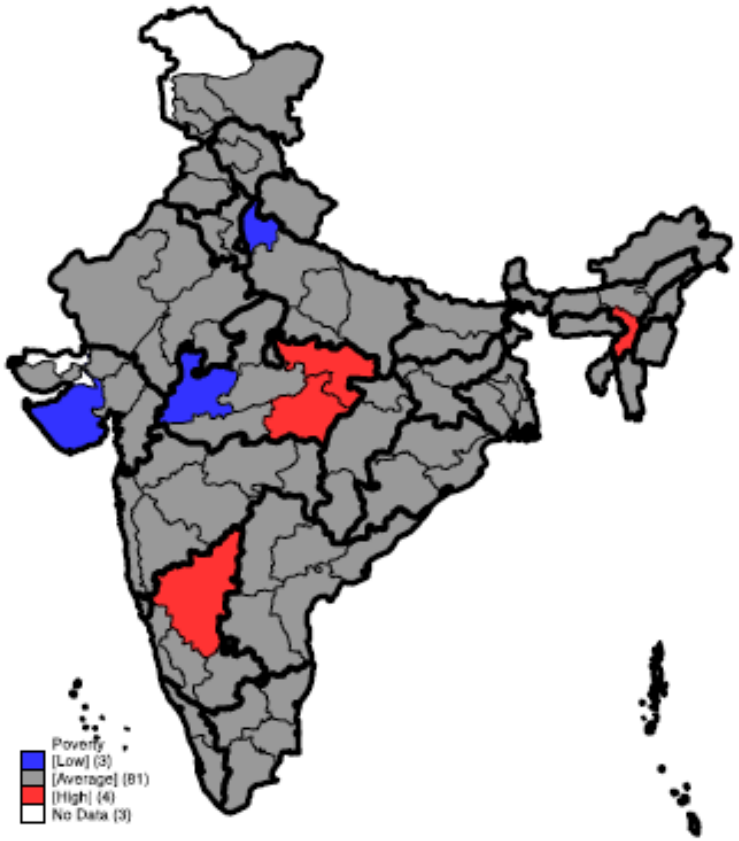
Red: Significantly high poverty
Blue: Significantly low poverty
Gray: No different from average

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

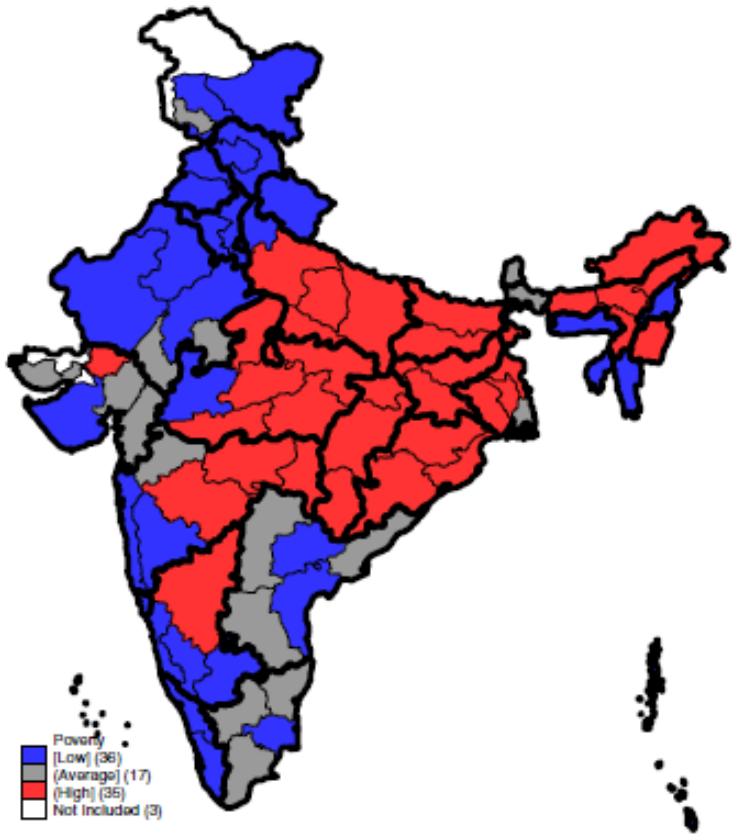
1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India

Region effects (five level)



Region effects (two level)



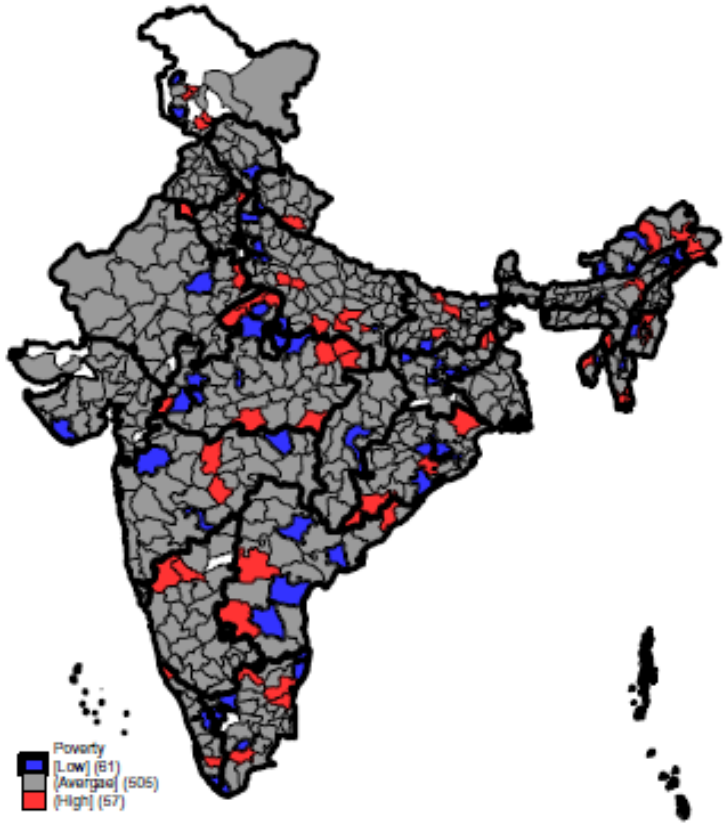
Red: Significantly high poverty
Blue: Significantly low poverty
Gray: No different from average

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

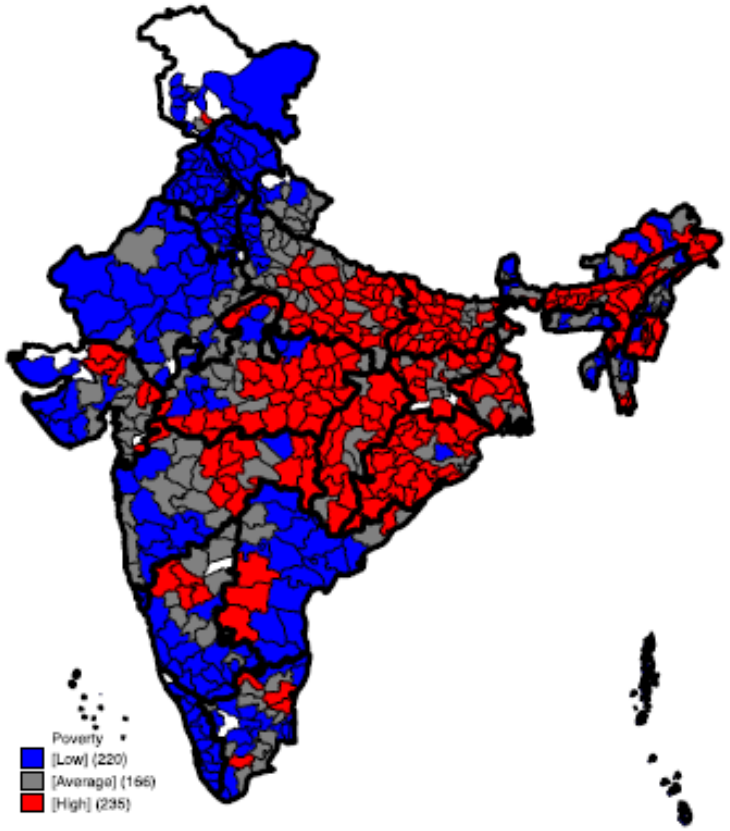
1. Importance of considering multiple (nested) geographies

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

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India

States

- 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**

Household poverty

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

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

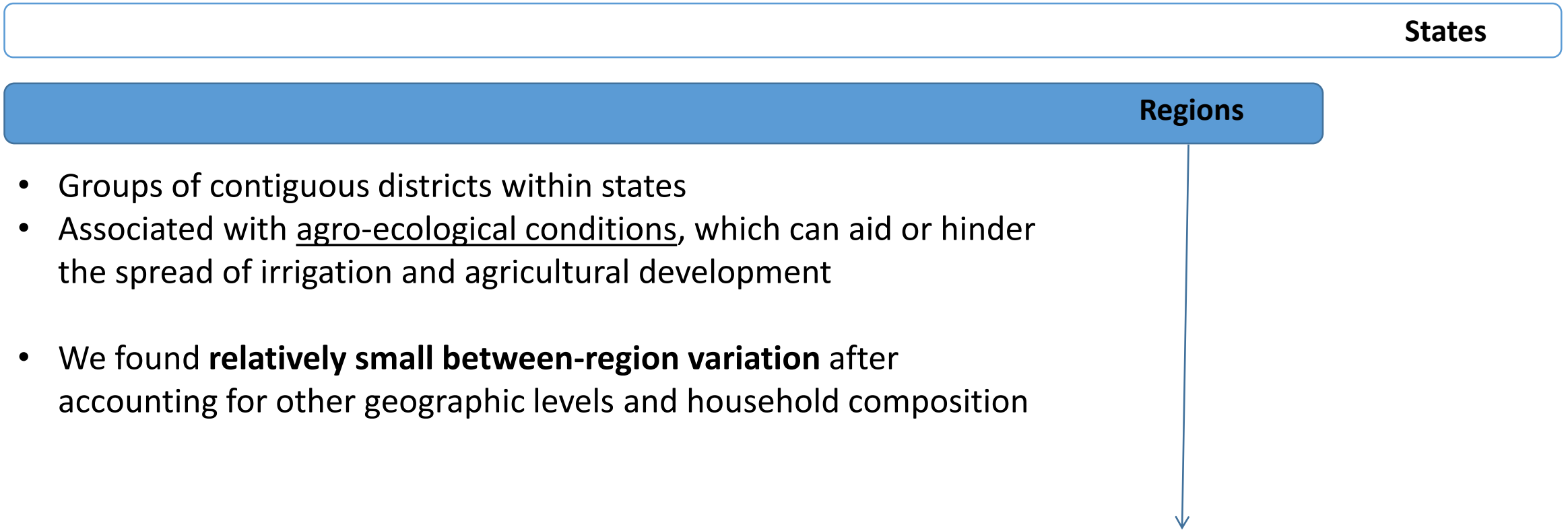
		State-level poverty *N* (% of total district)		
		Low	Average	High
District-level poverty *N* (% of total district)	Low	5 (0.80%)	24 (3.83%)	33 (5.27%)
	Average	37 (5.91%)	246 (39.30%)	224 (35.78%)
	High	6 (0.96%)	23 (3.67%)	28 (4.47%)

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

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India



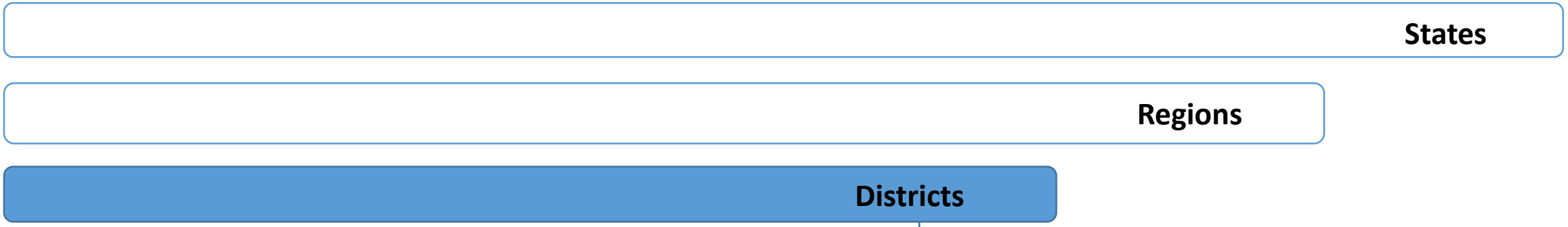
- Groups of contiguous districts within states
- Associated with agro-ecological conditions, which can aid or hinder the spread of irrigation and agricultural development

- We found **relatively small between-region variation** after accounting for other geographic levels and household composition

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

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

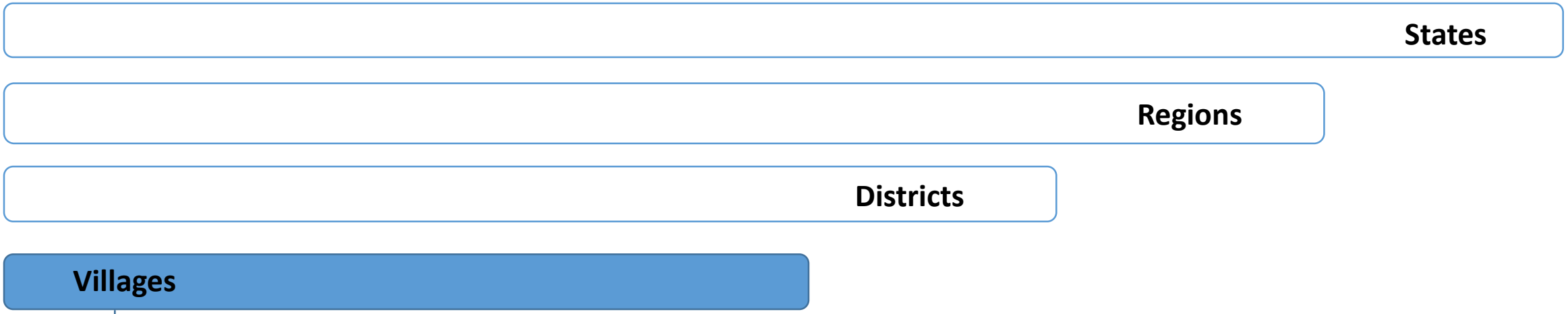
- **Prior district-wise analyses of poverty have over-estimated** the importance of district level

Household poverty

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

1. Importance of considering multiple (nested) geographies

Example: Geographies of Poverty in India



- The primary sampling units representing relatively homogenous local environment
- Heterogeneous service delivery and quality of implementation of social programs may contribute to the large between-village variation

Household poverty

- **First** analysis to assess the importance of this micro level

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

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.

Kim, R., Mohanty, S.K. and Subramanian, S.V. (2016). Multilevel geographies of poverty in India. World Development, 87, pp.349-359.

2. Importance of considering spatial and areal geographies

Example: Spatial Multilevel Modeling Approach to Mortality Trends in the US (1968-2013)

Everything is related to everything else, but near things are more related than distant things
(Tobler, 1970)

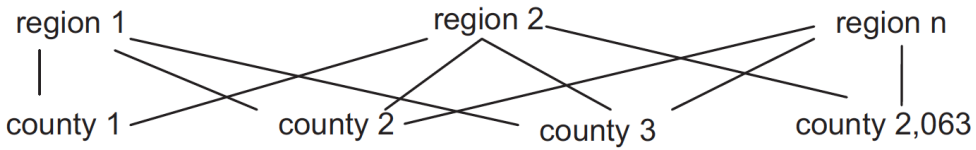


Fig. 2. Model 3 cross-classified structure.

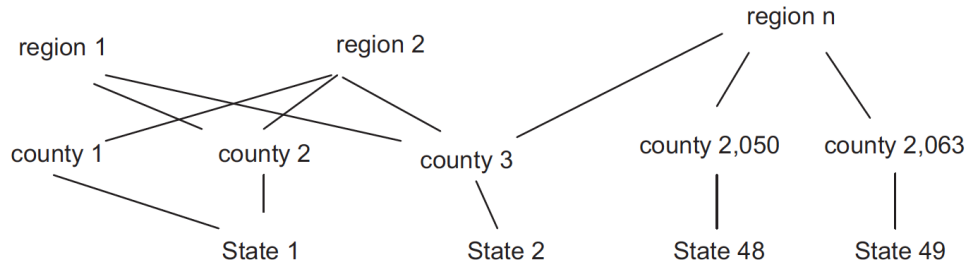


Fig. 3. Model 4 cross-classified structure.

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

2. Importance of considering spatial and areal geographies

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

Model 1. Single-level

$(i = county, 1, \ldots, 2063)$

$LE_i = \beta + (e_i), e_i \sim N(0, \sigma_e^2)$

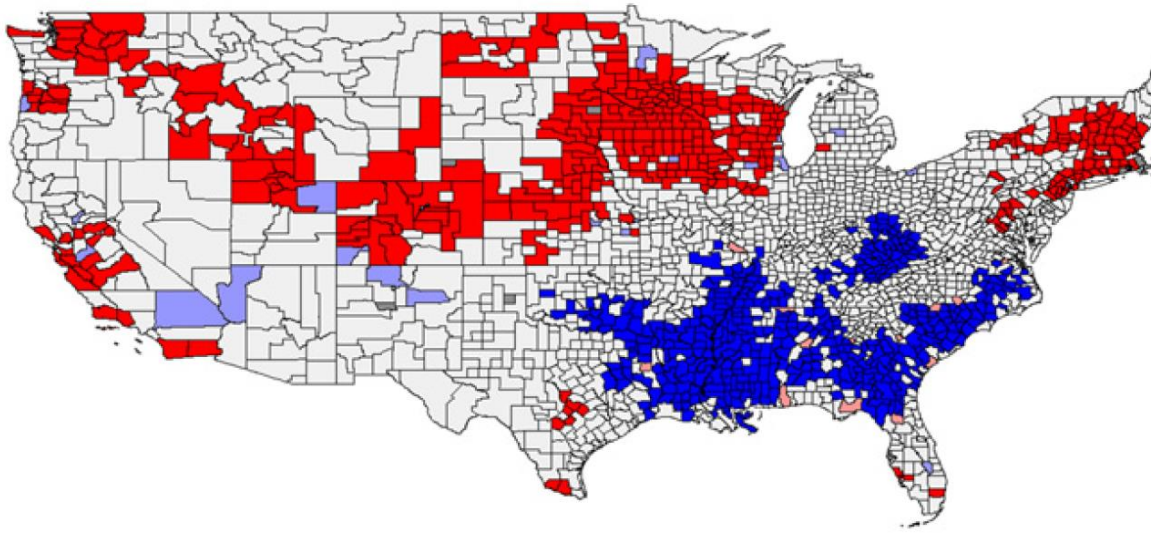


Fig. 4. Model 1 residuals exhibit spatial clustering (Moran's I=.638).

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

$$(i = county, 1, \dots, 2063)$$

$$(j = state, 1, \dots, 49)$$

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

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

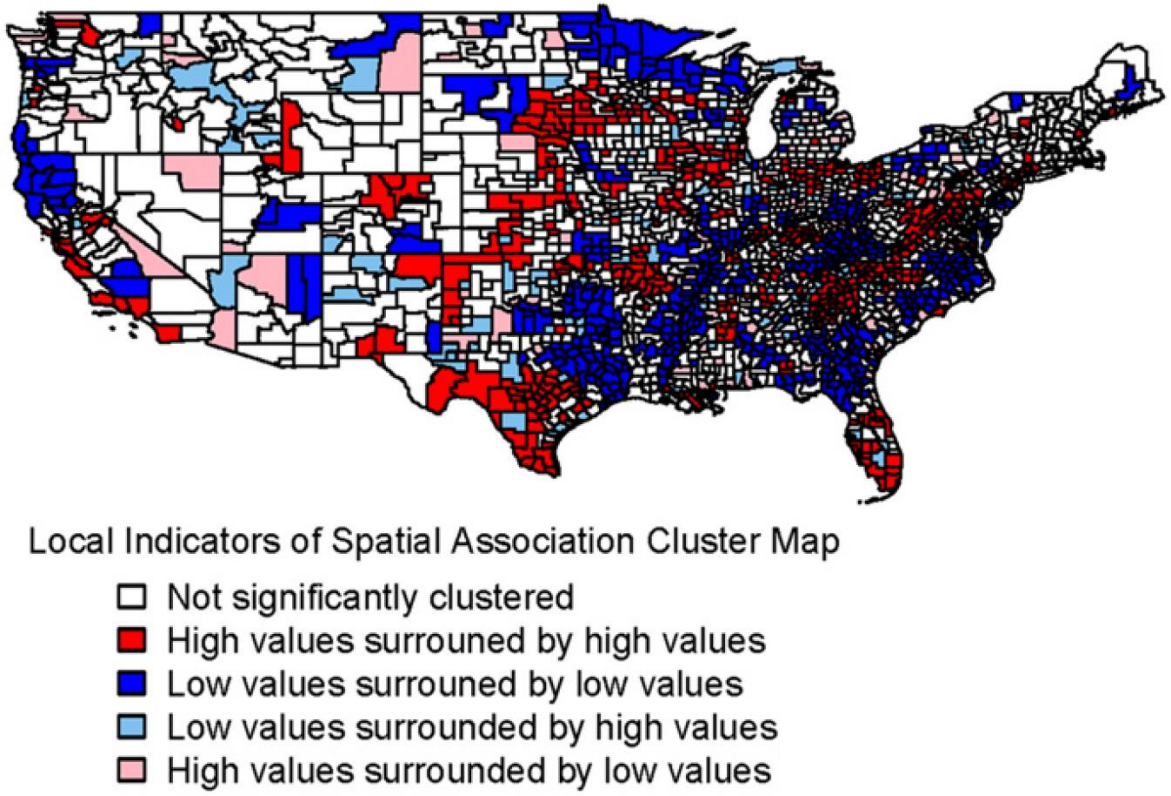


Fig. 5. Model 2 county-level residuals exhibit spatial clustering after accounting for membership in states (Moran's I=.289, 95% CI:.281 – .298).

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

$$(i = county, 1, \dots, 2063)$$

$$(j = state, 1, \dots, 49)$$

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

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

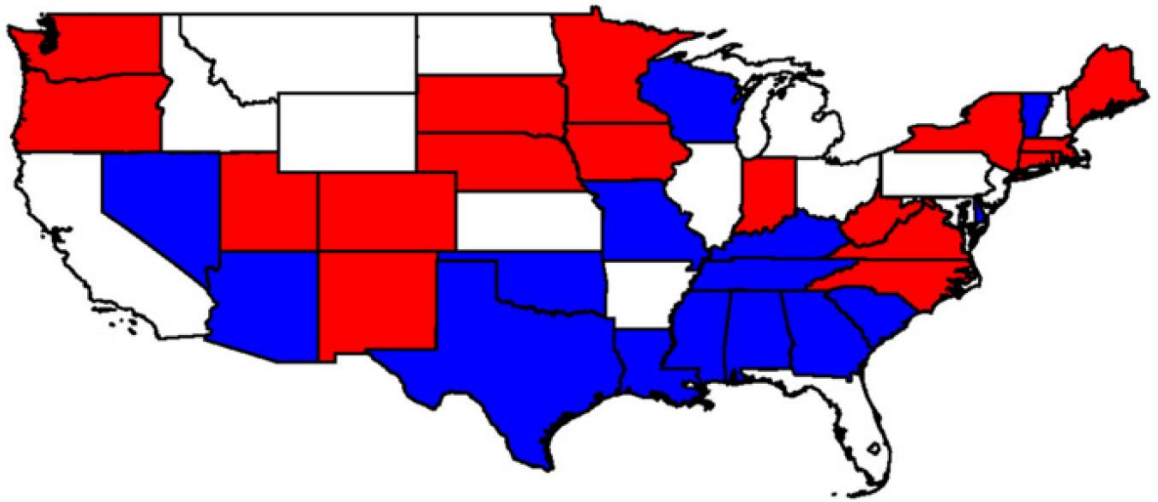


Fig. 6. Model 2 state-level residuals exhibit spatial clustering (Moran's I=.641, 95% CI:.561 – .707).

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

$$LE_i = \beta + (e_{neighbors(i)} + e_i),$$

$$e_{neighbors(i)} \sim N(\bar{e}_{neighbors(i)}, \sigma^2_{(e,neighbors)}/n_i),$$

$$e_i \sim N(0, \sigma_e^2)$$

- 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"

a

a	b	c
d	e	f
g	h	i

b

	a	b	c	d	e	f	g	h	i
A	0	1	0	1	1	0	0	0	0
B	1	0	1	1	1	0	0	0	0
C	0	1	0	0	1	1	0	0	0
D	1	1	0	0	1	0	1	1	0
E	1	1	1	1	0	1	1	1	1
F	0	0	1	0	1	0	0	1	1
G	0	0	0	1	1	0	0	1	0
H	0	0	0	1	1	1	1	0	1
I	0	0	0	0	1	1	0	1	0

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

$$LE_i = \beta + (e_{neighbors(i)} + e_i),$$

$$e_{neighbors(i)} \sim N(\bar{e}_{neighbors(i)}, \sigma^2_{(e,neighbors)}/n_i),$$

$$e_i \sim N(0, \sigma_e^2)$$

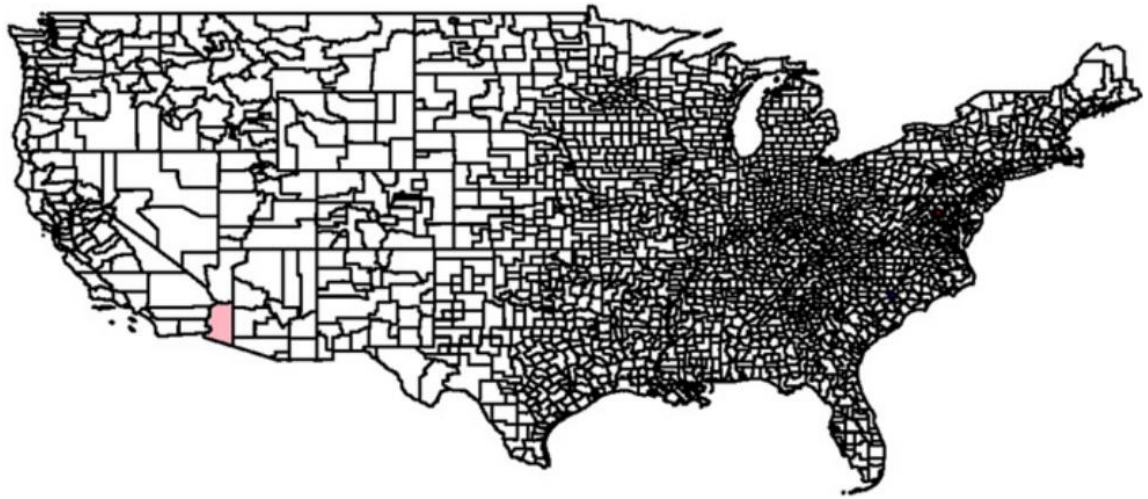


Fig. 7. Model 3 county-level residuals are spatially independent (Moran's I=.008 95% CI: −.021 – .038).

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

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

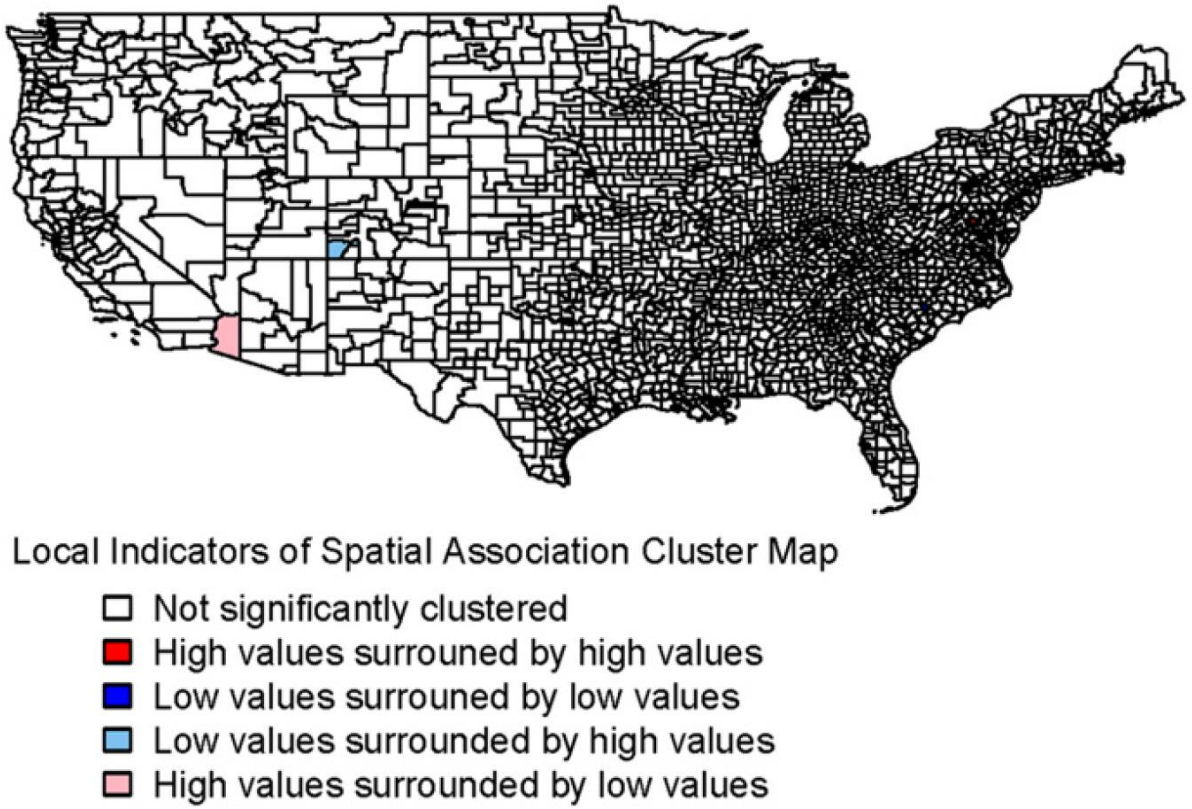


Fig. 8. Model 4 county-level residuals are spatially independent (Moran's $I = .006$, 95% CI: $-.022 - .036$).

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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

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

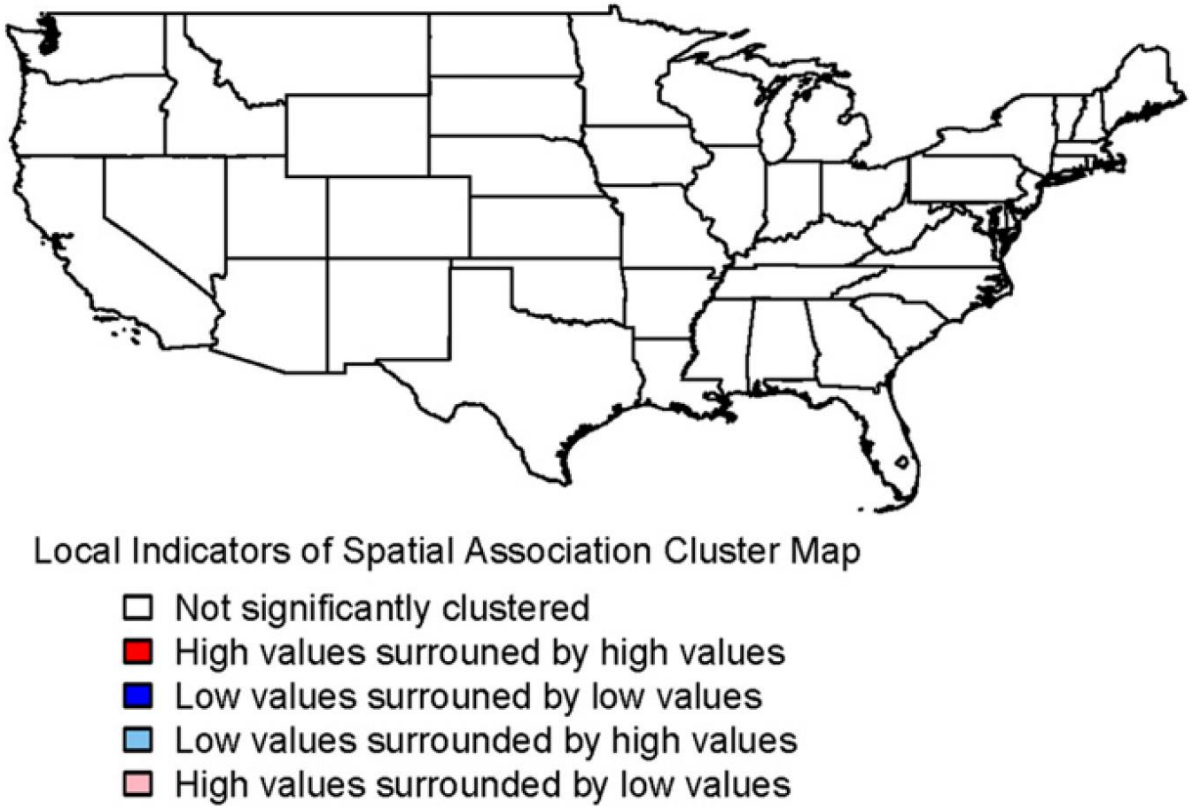


Fig. 9. Model 4 state-level residuals are spatially independent (Moran's $I = -.028$, 95% CI: $-.180 - .152$).

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

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.

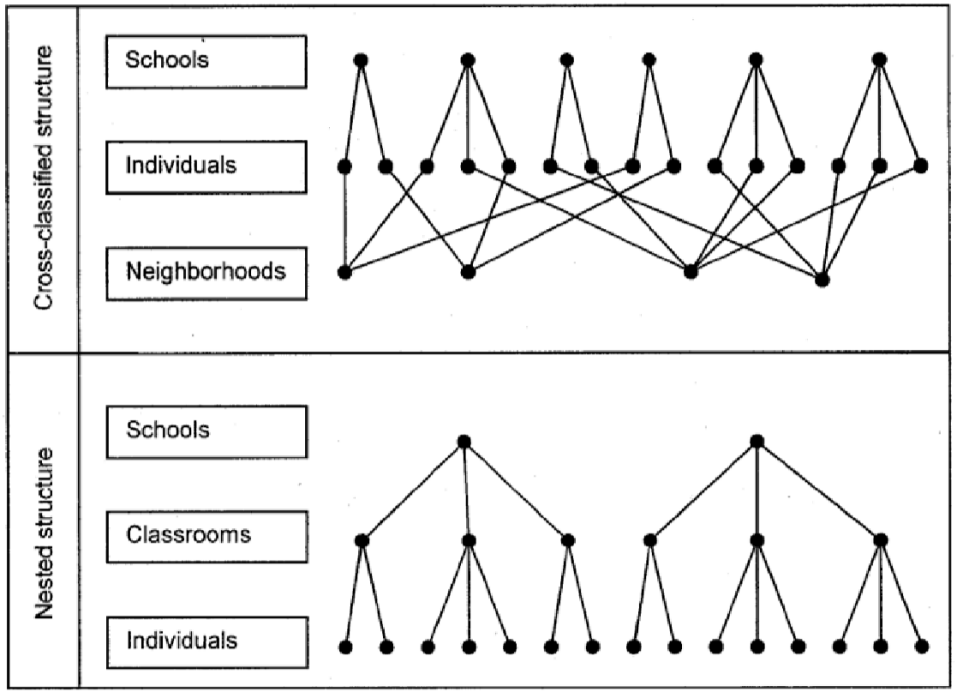
	DIC
Model 1	8696.5
Model 2	7367.0
Model 3	6126.6
Model 4	6106.2

- 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

Arcaya, M., Brewster, M., Zigler, C.M. and Subramanian, S.V. (2012). Area variations in health: A spatial multilevel modeling approach. Health & place, 18(4), pp.824-831.

3. Importance of considering cross-classified levels

Example: Schools and Neighborhoods Effect on Adolescents' Smoking Behaviors



Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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?

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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
  - $y_{i\,(j\,k)} = \beta_0 + \beta_{1\,0\,i\,(j\,k)} + \beta_{2\,0\,j} + \beta_{3\,0\,k} + \left(u_{0j} + u_{0k} + e_{0\,i\,(j\,k)}\right)$

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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 $\chi 2$)
  - 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

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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

	Null Models			Adjusted Model
Random effect estimates	**Model 1. School only**	**Model 2. Neighborhood only**	Model 3. Cross-classified	Model 4. Cross-classified
Neighborhood	-	**4.58 (3.66, 5.65)**		
School	**5.44 (4.04, 7.19)**	-		
Individual	**83.1 (81.3, 85.0)**	**84.0 (82.1, 85.9)**		

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

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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

	Null Models			Adjusted Model
Random effect estimates	Model 1. School only	Model 2. Neighborhood only	**Model 3. Cross-classified**	Model 4. Cross-classified
Neighborhood	-	4.58 (3.66, 5.65)	**0.46 (0.13, 0.88)**	
School	5.44 (4.04, 7.19)	-	**5.36 (3.95, 7.07)**	
Individual	83.1 (81.3, 85.0)	84.0 (82.1, 85.9)	**82.7 (80.9, 84.6)**	

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

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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

	Null Models			Adjusted Model
Random effect estimates	Model 1. School only	Model 2. Neighborhood only	Model 3. Cross-classified	**Model 4. Cross-classified**
Neighborhood	-	4.58 (3.66, 5.65)	0.46 (0.13, 0.88)	**0.24 (0.06, 0.61)**
School	5.44 (4.04, 7.19)	-	5.36 (3.95, 7.07)	**2.08 (1.42, 2.92)**
Individual	83.1 (81.3, 85.0)	84.0 (82.1, 85.9)	82.7 (80.9, 84.6)	**80.5 (78.8, 82.3)**

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

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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.

Dunn, E.C., Richmond, T.K., Milliren, C.E. and Subramanian, S.V. (2015). Using cross-classified multilevel models to disentangle school and neighborhood effects: an example focusing on smoking behaviors among adolescents in the United States. Health & place, 31, pp.224-232.

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