Spatial Aggregation and the Ecological Fallacy

J. Wakefield, H. Lyons

Handbook of Spatial Statistics

Nov 28, 2017

Table of Contents

Introduction

Ecological Studies:

- based on grouped data
- groups usually correspond to geographical areas (like block groups)
- used in political science, geography, sociology, epidemiology, and public health
- prone to drawbacks, including ecological bias

Ecological Bias

Ecological Bias describes the difference between estimated associations based on ecological and individual-level data (a special case of spatial misaligned data)

Introduction

- Focus on spatial regression, in which we are interested in the association between an outcome and covariates (**exposures**)
- Ecological bias is not a problem if we are just looking to summarize an area-level outcome (coefficients aren't of direct interest)

Ecological Fallacy

Problem with ecological analyses:

- loss of information due to aggregation
- mean function is not identifiable from the ecological data alone
- lack of identifiability can lead to the **ecological fallacy**

Ecological Fallacy

Individual and ecological associations between the outcome and an explanatory variable differ, and may even reverse direction

Motivating Example

Relationship between asthma hospitalization and air pollution

- $PM_{2.5}$ which is particulate matter less than 2.5 microns in diameter
- interested California from 1998-2000 time period
- **county-level** hospitalization data (58 counties)
- **point-level** pollution monitor data (86 monitor sites)

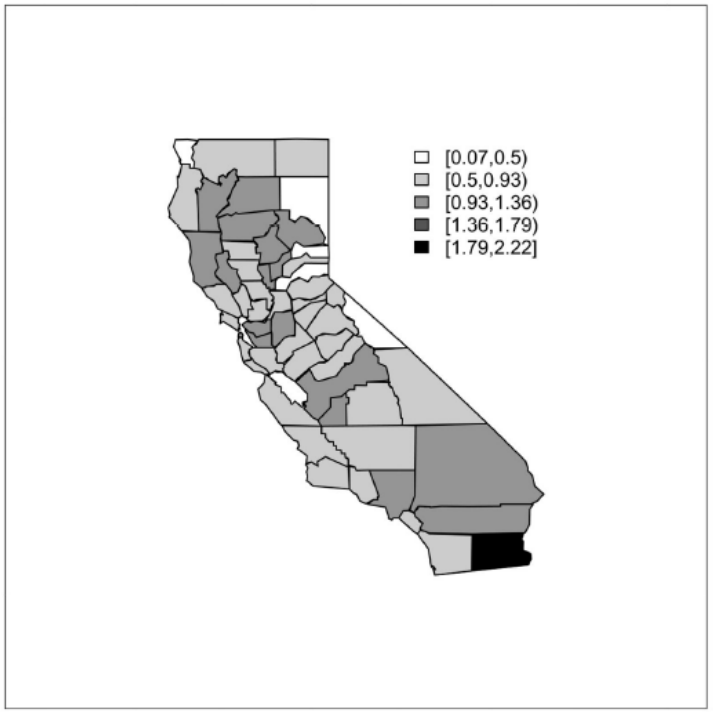

Figure 30.1.
SMRs for asthma hospitalization in Californian counties.

Definition of terms

- Y_i: total disease counts in county i over 3-year period
- x_{ik}: average log exposure in the last year of that period ($PM_{2.5}$ measured at monitor k in county i)
- m_i: range between 0 and 9, with $\sum_{i=1}^{58} m_i = 86$
- C= number of confounder stratum = 8
  - 2 age categories
  - 4 race categories
- z_i = elevation of the centroid of the block group, population-averaged to create as single number per county

SMR = standardized morbidity ratio

- SMR = $\frac{Y_i}{E_i}$ where $E_i = \sum_{c=1}^{C} N_{ic}\hat{q}_c$
- $\hat{q}_j$ are reference risks (calculated from data)
- E_i is meant to control for differences in outcome by stratum
- SMR is a summary of the area level relative risk

Ecological Model

$$E(Y_i|\bar{x}_i, z_i) = \mu_i = E_i exp(\beta_0 + \beta_1 \bar{x}_i + \beta_2 z_i)$$

County-level regression (ecological model)

- $\bar{x}_i$ is the (log) exposure within county (kriged values at the county centroid)
- $exp(\beta_1)$ is the ecological county-level relative risk associated with unit-increase in log exposure
- $exp(\beta_2)$ is the ecological county-level relative risk associate with a unit increase in elevation

Fit a poisson, quasi-poisson, and negative binomial model

- **subject to ecological bias** since they used aggregate risk and exposure measures averaged over monitors within counties
- ecological bias occurs due to *within-area variability* in exposures and confounders

Pure Specification Bias

This arises because a nonlinear risk model changes its form under aggregation.

Individual-level model:

$$E[Y_{ij}|x_{ij}] = \beta_0 + \beta_1 x_{ij}$$

where Y_{ij} and x_{ij} are outcome and exposure for individual j within area i.
Aggregate model:

$$E[\bar{Y}_i|\bar{x}_i] = \beta_0 + \beta_1 \bar{x}_i$$

where there is aggregation over the individuals.

- Under this specific scenario of the **linear model**, we have not lost anything by aggregation.

Non-Linear Model

Individual-level model:

$$E[Y_{ij}|x_{ij}] = e^{\beta_0+\beta_1 x_{ij}}$$

In this model,

- e^{β_0} is the risk associated with x = 0 (baseline risk)
- e^{β_1} is the relative risk corresponding to an increase in x of one unit

Aggregate model:

$$1)\ E[\bar{Y}_i|x_{ij}, j = 1, ..., n_i] = \frac{1}{n_i}\sum_{j=1}^{n_i} e^{\beta_0+\beta_1 x_{ij}}$$

so that ecological risk is the average of the risks of the constituent individuals.

Naive Model

A naive ecological model would assume:

$$2)\ E[\bar{Y}_i|\bar{x}_i] = e^{\beta_0^e + \beta_1^e \bar{x}_i}$$

where the "e" superscript distinguishes them from the individual level parameters.

- This is a **contextual effects model** since risk depends on the proportion of exposed individuals in the area.

Pure Specification Bias

Aggregate model:

$$1)\ E[\bar{Y}_i|x_{ij}, j = 1, ..., n_i] = \frac{1}{n_i}\sum_{j=1}^{n_i} e^{\beta_0+\beta_1 x_{ij}}$$

Contextual Effects model:

$$2)\ E[\bar{Y}_i|\bar{x}_i] = e^{\beta_0^e+\beta_1^e \bar{x}_i}$$

- the first model averages the risk across all exposures and the latter is the risk corresponding to the average exposure.
- $e^{\beta_1} = e^{\beta_1^e}$ only when there is no within-area variability in exposure, so that $x_{ij} = \bar{x}_i$ for all $j = 1, .., n_i$ individuals and for all areas
- data aggregation is usually based on administration groupings, not to obtain areas with constant exposure

Confounding Bias

We consider an example to see why controlling for confounding is, in general, impossible with ecological data:

- binary exposure (unexposed/exposed)
- binary confounder (ex: gender)
- complete within-area distribution of exposure and confounder can be described by three frequencies but the ecologic data usually consist of two quantities only
  - $\bar{x}_i$ (proportion exposed) and $\bar{z}_i$ (proportion male)

Exposure and gender distribution in area i, $\bar{x_i}$ is the proportion exposed and $\bar{z_i}$ is the proportion male; p_{i00}, p_{i01}, p_{i10}, p_{i11} are the within-area cross-classification frequencies.

	Female	**Male**	
Unexposed	p_{i00}	p_{i01}	$1-\bar{x_i}$
Exposed	p_{i10}	p_{i11}	$\bar{x_i}$
	$1-\bar{z_i}$	$\bar{z_i}$	1.0

Confounding Bias

Aggregate Model:

$$E[\bar{Y}_i|p_{i00}, p_{i01}, p_{i10}, p_{i11}] = (1 - \bar{x}_i - \bar{z}_i + p_{i11})e^{\beta_0} + (\bar{x}_i - p_{i11})e^{\beta_0+\beta_1}$$
$$+(\bar{z}_i - p_{i11})e^{\beta_0+\beta_2} + p_{i11}e^{\beta_0+\beta_1+\beta_2}$$

- marginal prevalences, ($\bar{x}_i$ and $\bar{z}_i$) are not sufficient to characterize the joint distribution unless x and z are independent (in which case z is not a within-area confounder)
- if p_{i11} is missing, should be estimated by marginal prevalences ($\bar{x}_i \times \bar{z}_i$)
  - Not possible to do this without individual-level data

Combining Ecological and Individual Data

Only solution to ecologic inference problem that does not require uncheckable assumptions: add individual-level data to the ecological data.

How?

Study designs by level of outcome and exposure data.

		Exposure	
		Individual	**Ecological**
Outcome	Individual	Individual	Semi-Ecological
	Ecological	Aggregate	Ecological

(based on data availability)

Aggregate Data Method

Obvious approach: to add individual-level data, collect a random sample of individuals within areas

Prentice and Sheppard (1995)

- Inference proceeds by constructing an estimating function based on the sample of $m_i \leq n_i$ individuals in each area
- For example, with samples of two variables $x_{ij}, z_{ij}, j = 1, ..., m_i$ we have mean function:

$$E[\bar{Y}_i | x_{ij}, z_{ij}, j = 1, ..., m_i] = \frac{1}{n_i} \frac{1}{m_i} \sum_{j=1}^{m_i} e^{\beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij}}$$

- Still involves bias because the complete set of exposures are not available
- They propose a finite sample correction in the estimating function (omitted from paper)

Semi-Ecological Studies

"semi-individual study", individual data are collected on outcomes and confounders, with exposure information arising from another source.

Naive Semi-Ecologic Model:

$$E[Y_{icj}|\bar{x}_i] = e^{\beta_0^e + \beta_{2c}^e + \beta_1^e \bar{x}_i}$$

- individual j, confounder stratum c, area i
- x_{icj} is the exposures of the individuals within stratum c and area i
- β_{2c} is the baseline risk in stratum c
- $\bar{x}_i$ is an exposure summary in area i

Still two sources of bias: pure specification bias (have not acknowledged within-area variability in exposure) and also have not allowed to exposure to vary by confounder stratum (not controlled for within-area confounding)

So, only provide approximately unbiased estimate of β_1 in select cases with low variability.

Conclusion

Suggested routine:

1. Write individual-level model for outcome-exposure association of interest (including confounders)
2. Assess for within-area variability in exposures and confounders
3. Proceed with one of the models given in the previous table, as appropriate

References

[1] Jonathan Wakefield and H Lyons. Spatial aggregation and the ecological fallacy. *Handbook of spatial statistics*, pages 541–558, 2010.

Example of Bias

We assume the following, based on a log-linear model and a normal within-area exposure distribution ($N(x|\bar{x}_i, s_i^2)$), with large n_i

$$E[\bar{Y}_i|\bar{x}_i] = e^{\beta_0+\beta_1\bar{x}_i+\beta_1^2 s_i^2/2}$$

which we compare to the naive ecological model ($e^{\beta_0^e+\beta_1^e\bar{x}_i}$)

- the within-area variance (s_i^2) is acting like a confounder (no pure specification bias if the exposure is constant within each area)
- allows us to characterize the direction of the bias
  - suppose $s_i^2 = a + b\bar{x}_i$ with b>0
  - in the ecological model, we are estimating $\beta_1^e = \beta_1 + \beta_1^2 b/2$
  - if $\beta_1 > 0$, then overestimation will occur using ecological model
  - if $\beta_1 < 0$, the ecological association (β_1^e) may reverse the sign when compared to β_1

Hierarchical Modeling

Unless the mean model is correct, adjustment for spatial dependence is a "pointless exercise"

- *A Solution to the Ecological Inference Problem* (1997) proposed a hierarchical model as "solution to the ecological inference problem"
- identifiability in this model is imposed through random effects prior (not possible to check appropriateness of this prior through ecologic data alone)