

## Module IV: Applications of Multi-level Models to Spatial Epidemiology

Francesca Dominici

&

Scott L Zeger

1

## Outline

- Multi-level models for spatially correlated data
  - Socio-economic and dietary factors of pellagra deaths in southern US
- Multi-level models for geographic correlation studies
  - The Scottish Lip Cancer Data
- Multi-level models for air pollution mortality risks estimates
  - The National Mortality Morbidity Air Pollution Study

2

## Data characteristics

- Data for disease mapping consists of disease counts and exposure levels in small adjacent geographical area
- The analysis of disease rates or counts for small areas often involves a trade-off between statistical stability of the estimates and geographic precision

3

## An example of multi-level data in spatial epidemiology

- We consider approximately 800 counties clustered within 9 states in southern US
- For each county, data consists of observed and expected number of pellagra deaths
- For each county, we also have several county-specific socio-economic characteristics and dietary factors
  - % acres in cotton
  - % farms under 20 acres
  - dairy cows per capita
  - Access to mental hospital
  - % afro-american
  - % single women

4

## Definition of Standardized Mortality Ratio

- $Y_i$  is the observed number of deaths in area  $i$
- $E_i$  is the expected number of deaths in area  $i$
- The “raw” Standardized Mortality Ratio is so defined:

$$SMR_i = (Y_i/E_i) \times 1000$$

5

## Definition of the expected number of deaths

- The expected number of deaths in area  $i$  can be calculated as follows:

$$E_i = \sum_j p_j n_{ij}$$

where

- $j$  is the population stratum generally defined by age  $\times$  gender  $\times$  race
- $p_j$  is observed frequency of death in the reference population
- $n_{ij}$  is the number of people at risk in area  $i$  in stratum  $j$

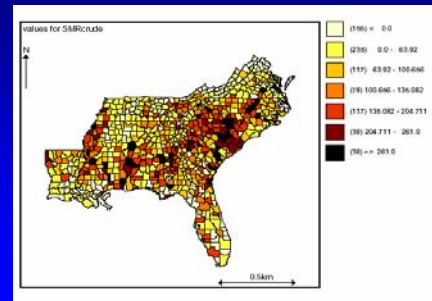
6

## Definition of Pellagra

- Disease caused by a deficient diet or failure of the body to absorb B complex vitamins or an amino acid.
- Common in certain parts of the world (in people consuming large quantities of corn), the disease is characterized by scaly skin sores, [diarrhea](#), mucosal changes, and mental symptoms (especially a [schizophrenia-like dementia](#)). It may develop after gastrointestinal diseases or [alcoholism](#).

7

## Crude Standardized Mortality Ratio (Observed/Expected) of Pellagra Deaths in Southern USA in 1930 (Courtesy of Dr Harry Marks)



8

## Scientific Questions

- Which social, economical, behavioral, or dietary factors best explain spatial distribution of pellagra in southern US?
- Which of the above factors is more important for explaining the history of pellagra incidence in the US?
- To which extent, state-laws have affected pellagra?

9

## Statistical Challenges

- For small areas SMR are very instable and maps of SMR can be misleading
  - Spatial smoothing
- SMR are spatially correlated
  - Spatially correlated random effects
- Covariates available at different level of spatial aggregation (county, State)
  - Multi-level regression structure

10

## Spatial Smoothing

- Spatial smoothing can reduce the random noise in maps of observable data (or disease rates)
- Trade-off between geographic resolution and the variability of the mapped estimates
- Spatial smoothing as method for reducing random noise and highlight meaningful geographic patterns in the underlying risk

11

## Shrinkage Estimation

- Shrinkage methods can be used to take into account instable SMR for the small areas
- Idea is that:
  - smoothed estimate for each area “borrow strength” (precision) from data in other areas, by an amount depending on the precision of the raw estimate of each area

12

## Shrinkage Estimation

- Estimated rate in area  $A$  is adjusted by combining knowledge about:
  - Observed rate in that area;
  - Average rate in surrounding areas
- The two rates are combined by taking a form of weighted average, with weights depending on the population size in area  $A$

13

## Shrinkage Estimation

- When population in area  $A$  is large
  - Statistical error associated with observed rate is small
  - High credibility (weight) is given to observed estimate
  - Smoothed rate is close to observed rate
- When population in area  $A$  is small
  - Statistical error associated with observed rate is large
  - Little credibility (low weight) is given to observed estimate
  - Smoothed rate is “shrunk” towards rate mean in surrounding areas

14

### A Multi-level Model for Spatial Smoothing of SMR

$$Y_i | \mu_i \sim \text{Poisson}(\mu_i)$$

$$\log \mu_i = \log E_i + b_i$$

$$b_i | b_{j \neq i} \sim N\left(\frac{\sum_{j \neq i} w_{ij} b_j}{\sum_{j \neq i} w_{ij}}, \sigma^2 \frac{1}{\sum_{j \neq i} w_{ij}}\right)$$

where:

- $b_i$  are area-specific random effects with a spatially correlated random effect distribution
- $w_{ij}$  are weights defining which regions  $j$  are neighbors to region  $i$  (by convention  $w_{ii} = 0$ , for all  $i$ )
- $\sigma^2$  is the variance controlling how similar the  $b_i$  is to its neighbors

### Raw and Smoothed Standardized Mortality Rates

- $Y_i$  are observed disease counts in area  $i$
- $E_i$  are expected disease counts in area  $i$
- The raw and smoothed standardized mortality ratio ( $SMR_i$  and  $\widehat{SMR}_i$ ) are so defined:

$$SMR_i = \frac{Y_i}{E_i}$$

$$\widehat{SMR}_i = \frac{\hat{\mu}_i}{E_i}$$

- In areas with abundant data:

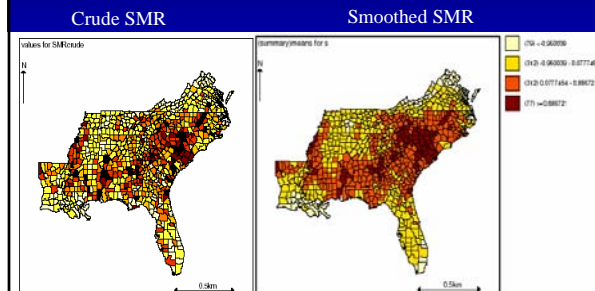
$$\widehat{SMR}_i \approx SMR_i$$

- In areas with sparse data:

$\widehat{SMR}_i \approx$  weighted average of the SMR in the adjacent counties

16

### SMR of pellagra deaths for 800 southern US counties in 1930



17

## Multi-level Models for Geographical Correlation Studies

- Geographical correlation studies seek to describe the relationship between the geographical variation in disease and the variation in exposure

18

**A Multilevel model for disease counts**

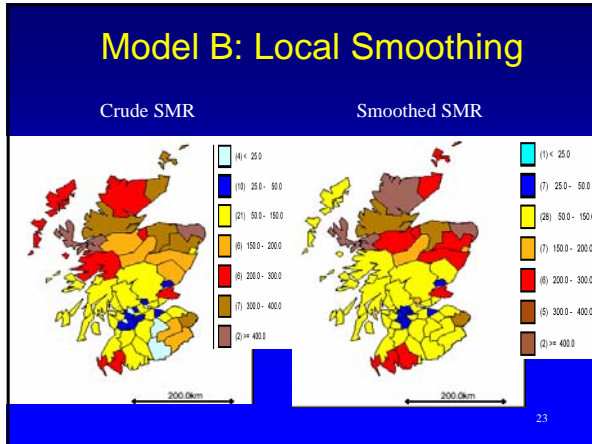
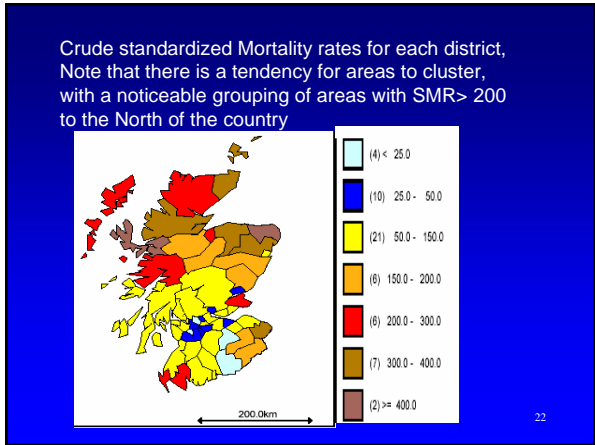
- $Y_{is}$  are observed disease counts in county  $i$  within state  $s$
- $E_{is}$  are expected disease counts in county  $i$  within state  $s$
- **Stage I: County-level, within state model**  
 $Y_{is} | \mu_{is} \sim \text{Poisson}(\mu_{is})$   
 $\log \mu_{is} = \log E_{is} + \beta_{1s}(\text{cot}_{is} - \overline{\text{cot}}) + \beta_{2s}(\text{milk}_{is} - \overline{\text{milk}}) + b_i$   
 $b_i \sim \text{spatially correlated random effects}$
- **Stage II: Between-states model**  
 $\beta_{1s} = \gamma_{11} + \gamma_{12}\text{state-taxes}_s + N(0, \sigma_1^2)$   
 $\beta_{2s} = \gamma_{21} + \gamma_{22}\text{state-taxes}_s + N(0, \sigma_2^2)$   
 where:  
  - $\beta_{1s}$  and  $\beta_{2s}$  are county-specific log-relative rates
  - $\gamma_{11}$  is the overall log-relative rate of pellagra mortality for the counties with average

**Example: Scottish Lip Cancer Data**  
(Clayton and Kaldor 1987 Biometrics)

- Observed and expected cases of lip cancer in 56 local government district in Scotland over the period 1975-1980
- Percentage of the population employed in agriculture, fishing, and forestry as a measure of exposure to sunlight, a potential risk factor for lip cancer

**A Multilevel model for Lip Cancer Study**

- $Y_i$  are observed lip cancer cases in district  $i$
- $E_i$  are expected lip cancer cases in district  $i$
- $Y_i | \mu_i \sim \text{Poisson}(\mu_i)$   
 $\log \mu_i = \log E_i + \beta_0 + \beta_1(\text{agr}_i - \overline{\text{agr}}) + b_i$
- We consider two models for the random effects:  
  - **A: Global Smoothing**  
 $b_i \sim N(0, \sigma^2)$
  - **B: Local Smoothing**  
 $b_i | b_{j \neq i} \sim N\left(\frac{\sum_{j \neq i} w_{ij} b_j}{\sum_{j \neq i} w_{ij}}, \sigma^2 \frac{1}{\sum_{j \neq i} w_{ij}}\right)$



**Parameter estimates**

	A	B
intercept	0.099 (SE = 0.098)	0.091 (SE = 0.051)
slope	<b>0.069</b> (SE = 0.014)	<b>0.045</b> (SE = 0.012)
variance	0.602 (SE = 0.087)	0.667 (SE = 0.119)

### Estimating Relative Risks

---

Relative Risk is defined as

$$RR(\text{agr}_i) = \exp(\beta_0 + \beta_1(\text{agr}_i - \overline{\text{agr}}))$$

We approximated the posterior distributions of:

- RR of lip cancer in the areas with the highest proportion for workers in agriculture ( $\text{agr}_i = 24\%$ ):

$$RR(\text{agr}_i = 24) = \exp(\beta_0 + \beta_1(24 - \overline{\text{agr}}))$$

- RR of lip cancer in the areas with the average proportion for workers in agriculture ( $\text{agr}_i = \overline{\text{agr}}$ ):

$$RR(\text{agr}_i = \overline{\text{agr}}) = \exp(\beta_0)$$

25

### Posterior distribution of Relative Risks for maximum exposure

A: Global smoothing  
(posterior mean = 3.25)

B: Local smoothing  
(posterior mean = 2.18)

### Posterior distribution of Relative Risks for average exposure

A: Global smoothing  
(posterior mean = 1.08)

B: Local smoothing  
(posterior mean = 1.09)

27

### Results

- Under a model for global smoothing, the posterior mean of the relative risk for lip cancer in areas with the highest percentage of outdoor workers is 3.25
- Under model for local smoothing, the posterior mean is lower and equal to 2.18

28

### Discussion

- In multi-level models is important to explore the sensitivity of the results to the assumptions inherent with the distribution of the random effects
- Specially for spatially correlated data the assumption of global smoothing, where the area-specific random effects are shrunk toward and overall mean might not be appropriate
- In the lip cancer study, the sensitivity of the results to global and local smoothing, suggest presence of spatially correlated latent factors

29

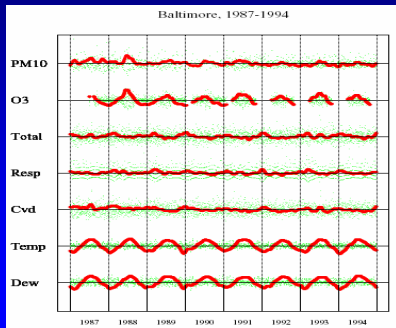
### The National Morbidity Mortality Air Pollution Study

NMMAPS is a multi-site time series study assessing short-term effects of air pollution on mortality/morbidity comprising:

- a national data base of air pollution and mortality;
- statistical methods for estimating associations between air pollution and mortality for the 90 largest US cities, and on average for the entire nation.

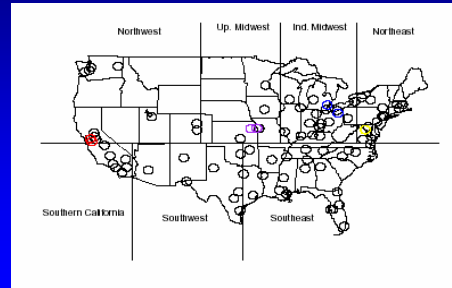
30

## Daily time series of air pollution, mortality and weather in Baltimore 1987-1994



31

## 90 Largest Locations in the USA



32

### A Multilevel Model for NMMAPS

- Let  $\hat{\beta}_{cr}$  and  $v_{cr}$  the relative rate estimate and its statistical variance of the percentage increase in mortality associated with a  $10\mu/m^3$  increase in particulate matter in city  $c$  in region  $r$ .
- These estimates are obtained by fitting time series models with each city (Dominici et al 2000, Royal Statistical Society).
- The NMMAPS multilevel model is so defined:
  - Stage I: county-level, within region

$$\hat{\beta}_{cr} = \beta_{cr} + N(0, v_{cr})$$

$$\beta_{cr} = \alpha_{0r} + \alpha_{1r} \text{income}_{cr} + \alpha_{2r} \text{traffic}_{cr} + N(0, \sigma^2)$$

- Stage II: region-level

$$\alpha_{0r} = \gamma_{00} + \gamma_{01} \text{reg.char}_r + N(0, \tau_1^2)$$

$$\alpha_{1r} = \gamma_{10} + \gamma_{11} \text{reg.char}_r + N(0, \tau_2^2)$$

33

### Multilevel Model for NMMAPS: Analysis of Variance

- We re-write the multilevel model for NMMAPS without covariates:

$$\hat{\beta}_{cr} = \beta_{cr} + N(0, v_{cr})$$

$$\beta_{cr} = \alpha_r + N(0, \sigma^2)$$

$$\alpha_r = \gamma + N(0, \tau^2)$$

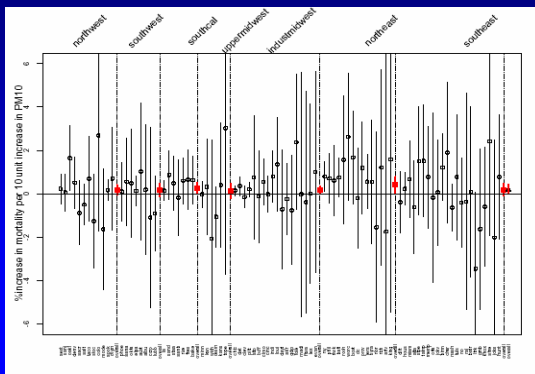
- $\beta_{cr}$  is the true city-specific pollution effect
- $\alpha_r$  is the regional-average air pollution effect
- $\gamma$  is the national average air pollution effect
- $\sigma^2$  heterogeneity of air pollution effects within region
- $\tau^2$  heterogeneity of air pollution effects across regions

We can write the total difference between the city-specific estimate and the national average estimate as follows:

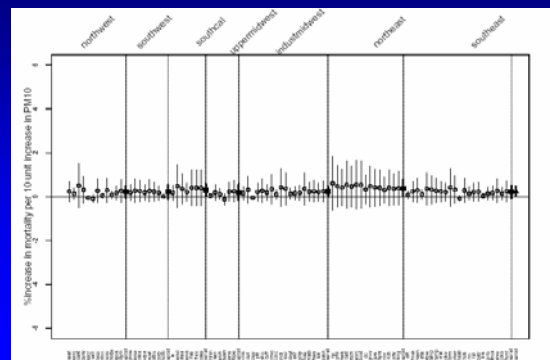
$$\underbrace{(\hat{\beta}_{cr} - \gamma)}_{\text{total diff}} = \underbrace{(\hat{\beta}_{cr} - \beta_{cr})}_{\text{within city}} + \underbrace{(\beta_{cr} - \alpha_r)}_{\text{within region}} + \underbrace{(\alpha_r - \gamma)}_{\text{between regions}}$$

34

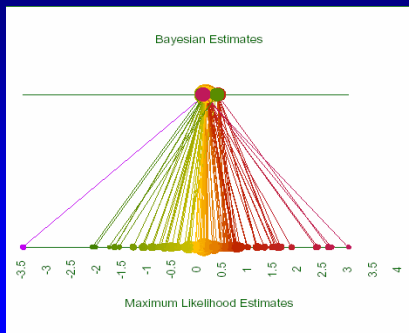
## City-specific MLE



## City-specific Bayesian Estimates

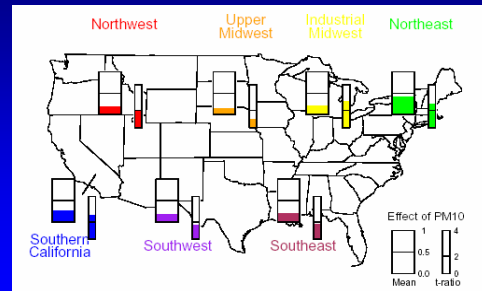


## Shrinkage



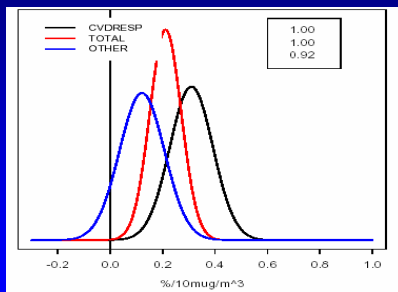
37

## Regional map of air pollution effects



Partition of the United States used in the 1996 Review of the NAAQS

## National-average estimates for CVDRESP, Total and Other causes mortality



Samet, Dominici, Zeger et al. NEJM 2000

39

## Pooling

City-specific relative rates are pooled across cities to:

1. estimate a *national-average* air pollution effect on mortality;
2. explore geographical patterns of variation of air pollution effects across the country

40

## Pooling

- Implement the old idea of *borrowing strength across studies*
- Estimate heterogeneity and its uncertainty
- Estimate a national-average effect which takes into account heterogeneity

41

## Discussion

- Multilevel models are a natural approach to analyze data collected at different level of spatial aggregation
- Provide an easy framework to model sources of variability (within county, across counties, within regions etc..)
- Allow to incorporate covariates at the different levels to explain heterogeneity within clusters
- Allow flexibility in specifying the distribution of the random effects, which for example, can take into account spatially correlated latent variables

42

## Key Words

- Spatial Smoothing
- Disease Mapping
- Geographical Correlation Study
- Hierarchical Poisson Regression Model
- Spatially correlated random effects
- Posterior distributions of relative risks

43