Space, Time and Gradients: Why we need them in statistical modeling for public health data

Sudipto Banerjee

Based upon joint work with Harrison Quick and Bradley P. Carlin

Department of Biostatistics
Fielding School of Public Health
University of California, Los Angeles

The Information Age

- We live in an Information Age.
- Computers collect and store information in quantities that were earlier unimaginable.
- What is this information?
  - Measurements, counts, costs, sales revenue...
  - arising in sciences, public health, business...
- Raw, "undigested" data stored on computer disks is useless unless we make sense of it.
- Statistics: the art and science of extracting meaning from seemingly incomprehensible data.
- Make good use of *information* to make sound *decisions*.

Space debris

- Researchers in diverse areas such as climatology, ecology, environmental health, and real estate marketing are increasingly faced with the task of analyzing data that:
  - have many important predictors and response variables,
  - are often presented as maps,
  - and/or as data streaming in over time

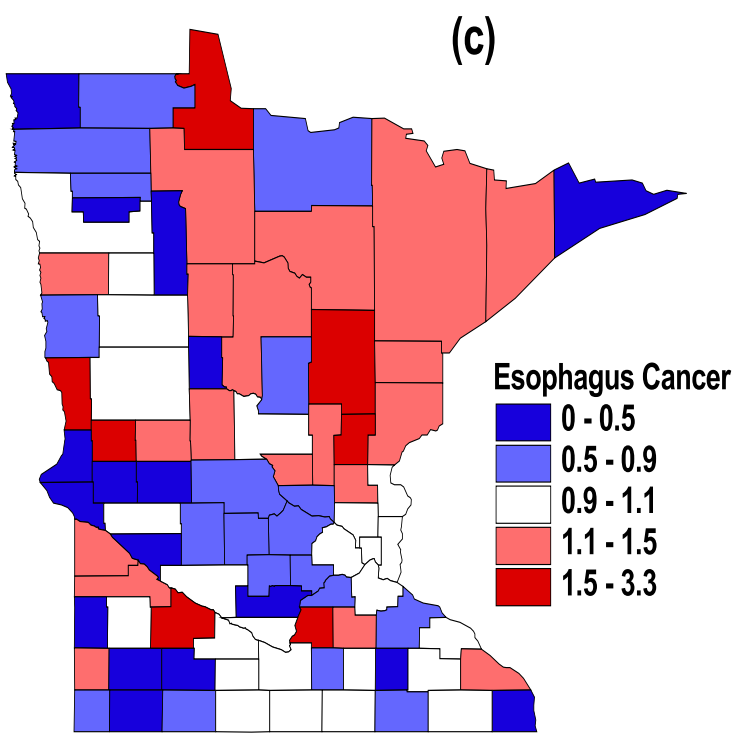
Example:

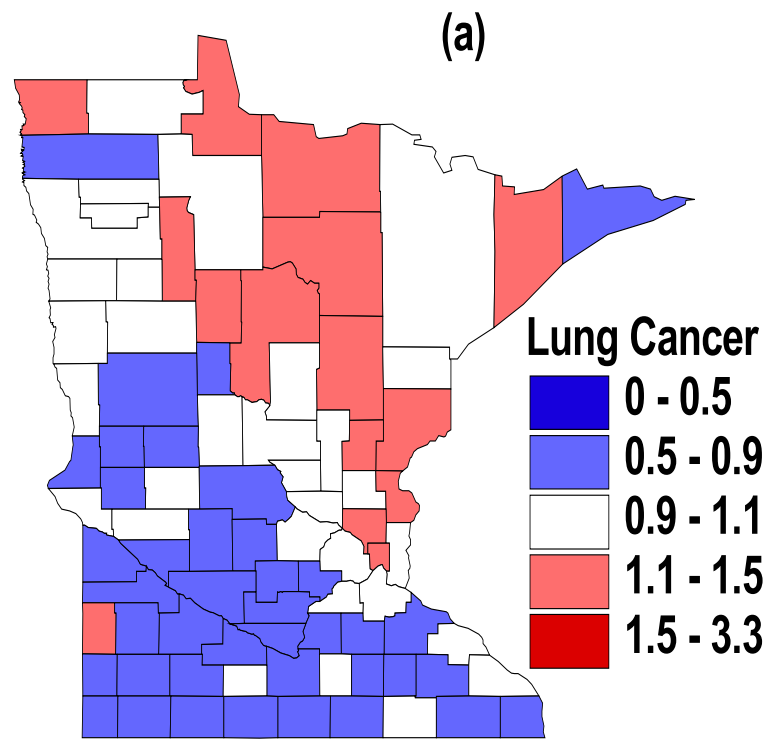
In an epidemiological investigation, we might wish to analyze lung, breast, colorectal, and cervical cancer rates

- by county and year in a particular state
- with smoking, mammography, and other important screening and staging information also available at some level.

Areal unit data

Maps of raw standard mortality ratios (SMR) of lung and esophagus cancer between 1991 and 1998 in Minnesota counties





Public health professionals who collect such data are charged not only with surveillance, but also statistical *inference* tasks, such as

- *modeling* of trends and correlation structures
- *estimation* of underlying model parameters
- *hypothesis testing* (or comparison of competing models)
- *prediction* of observations at unobserved times or locations.

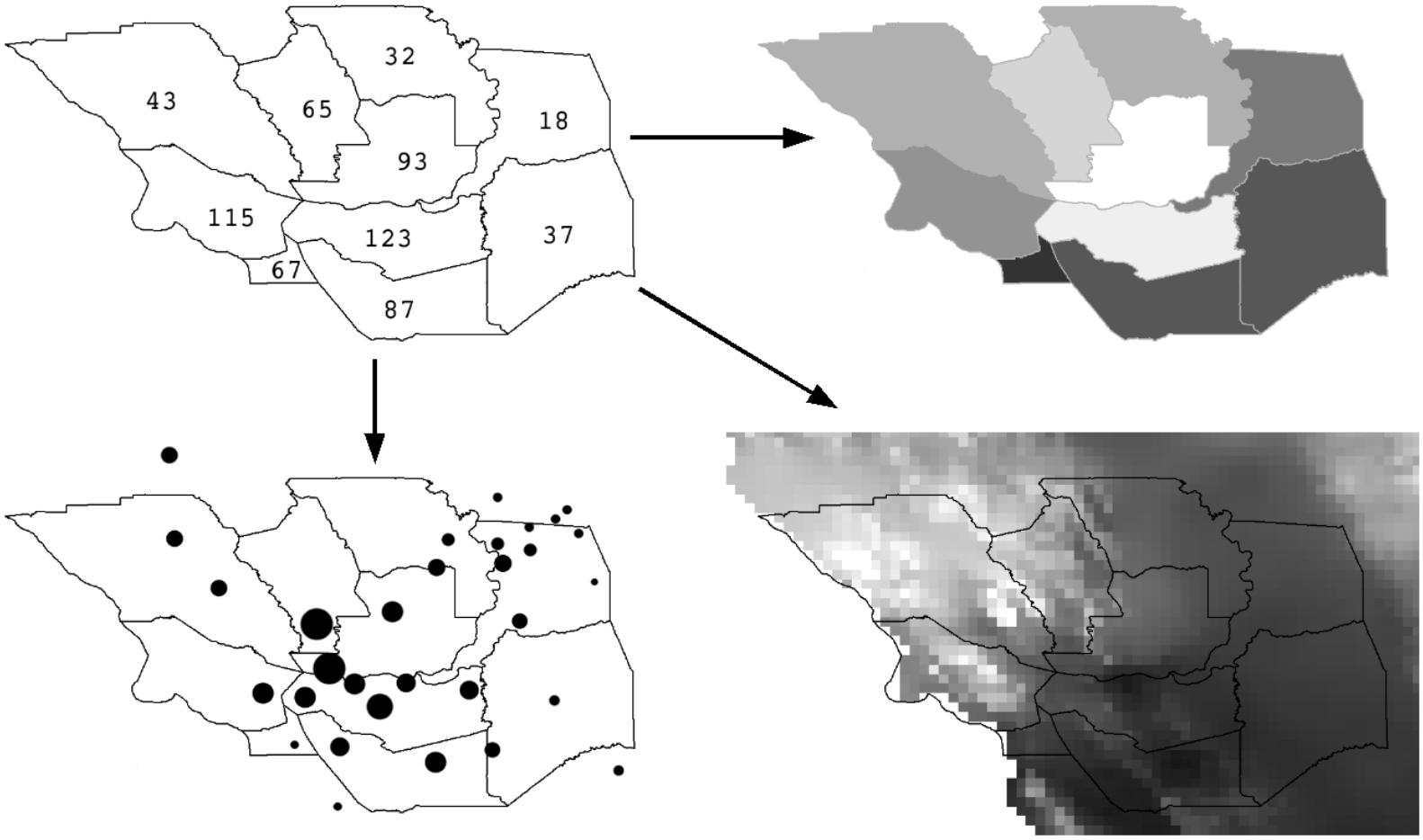
Datasets we are exploring

- **Asthma:** Asthma has been commonly linked to air pollutants, especially ozone and PM_{10}. The heath outcome of interest is county level Emergency Room (ER) visit counts for asthma within high risk regions in California (CA), New York (NY), and Michigan (MI) observed between 2000-2008 (or sub-intervals depending on data availability). For California, the Office of Statewide Health Planning and Development (OSHPD) provides county specific information on all emergency room visits in licensed hospitals `http://www.oshpd.state.ca.us`. The Statewide Planning and Research Cooperative System (SPARCS) `www.health.state.ny.us/statistics/sparcs/index.htm` offers asthma related hospital visits for New York State. Similar data are maintained by the Michigan Inpatient Database.

- **Skin-cancer:** Non-melanoma skin cancer (NMSC), comprised primarily of basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) in a ratio of $4:1$, is linked to exposure to ultraviolet radiation – perhaps one of the major contributors to the development of non-melanoma skin cancer. Data from NCI's SEER database.

- **Salmonellosis:** The first detectable changes in human health from the impact of climatic factors may well be alterations in the geographic range and seasonality of certain infectious diseases – including food-borne infections (e.g. salmonellosis) which peak in the warmer months. Data for salmonellosis will come from the Center for Disease Control's (CDC) Foodborne Diseases Active Surveillance Network (FoodNet) surveillance system.

- Predictors come from National Air Monitoring Stations and State and Local Air Monitoring Stations (`http://www.epa.gov/cludygxb/programs/namslam.html`); U.S. Census Bureau (`www.census.gov`) along with available health risk or exposure data provided by states' bureau of health; complete coverage climate (precipitation, temperature, temperature extremes, etc.) raster data generated by The National Centers for Environmental Prediction's (NCEP's) North American Regional Reanalysis (NARR) (`http://www.emc.ncep.noaa.gov/mmb/rreanl`).


43
32
65
18
93
115
123
37
67
87

Introduction – Space-Time Data Analysis

- The importance of "Where" and "When" in statistics.
- Space-time modeling falls under one of four settings:
  1. Continuous space, discrete time
     - Example: Monthly temperature data
  2. Continuous space, continuous time
     - Example: Directional wind data
  3. Discrete space, discrete time
     - Example: Yearly asthma rates across counties

Introduction – Space-Time Data Analysis

- The importance of "Where" and "When" in statistics.
- Space-time modeling falls under one of four settings:
  1. Continuous space, discrete time
     - Example: Monthly temperature data
  2. Continuous space, continuous time
     - Example: Directional wind data
  3. Discrete space, discrete time
     - Example: Yearly asthma rates across counties
  4. Discrete space, continuous time
     - Example: Daily asthma rates across counties?
- "Discrete" usually refers to some level of aggregation.
- The last category has, arguably, garnered scant attention.

Introduction

- Asthma Hospitalization Rates in California
- County level data ($N_s = 58$ spatial regions)
- Aggregated monthly from 1991–2008 ($N_t = 216$ time points)
- Often converted to rates — say per 1,000 residents

Introduction

- Asthma Hospitalization Rates in California
- County level data ($N_s = 58$ spatial regions)
- Aggregated monthly from 1991–2008 ($N_t = 216$ time points)
- Often converted to rates — say per 1,000 residents

Can we reconstruct hospitalization rates at a daily level? Can we estimate rate of change in hospitalization rates?

Introduction

- Asthma Hospitalization Rates in California
- County level data ($N_s = 58$ spatial regions)
- Aggregated monthly from 1991–2008 ($N_t = 216$ time points)
- Often converted to rates — say per 1,000 residents

Can we reconstruct hospitalization rates at a daily level? Can we estimate rate of change in hospitalization rates?

- We need to operate at a resolution much *finer* than that of the observations.
- Cannot treat this problem as one of discrete space - discrete time.

Introduction – The Data

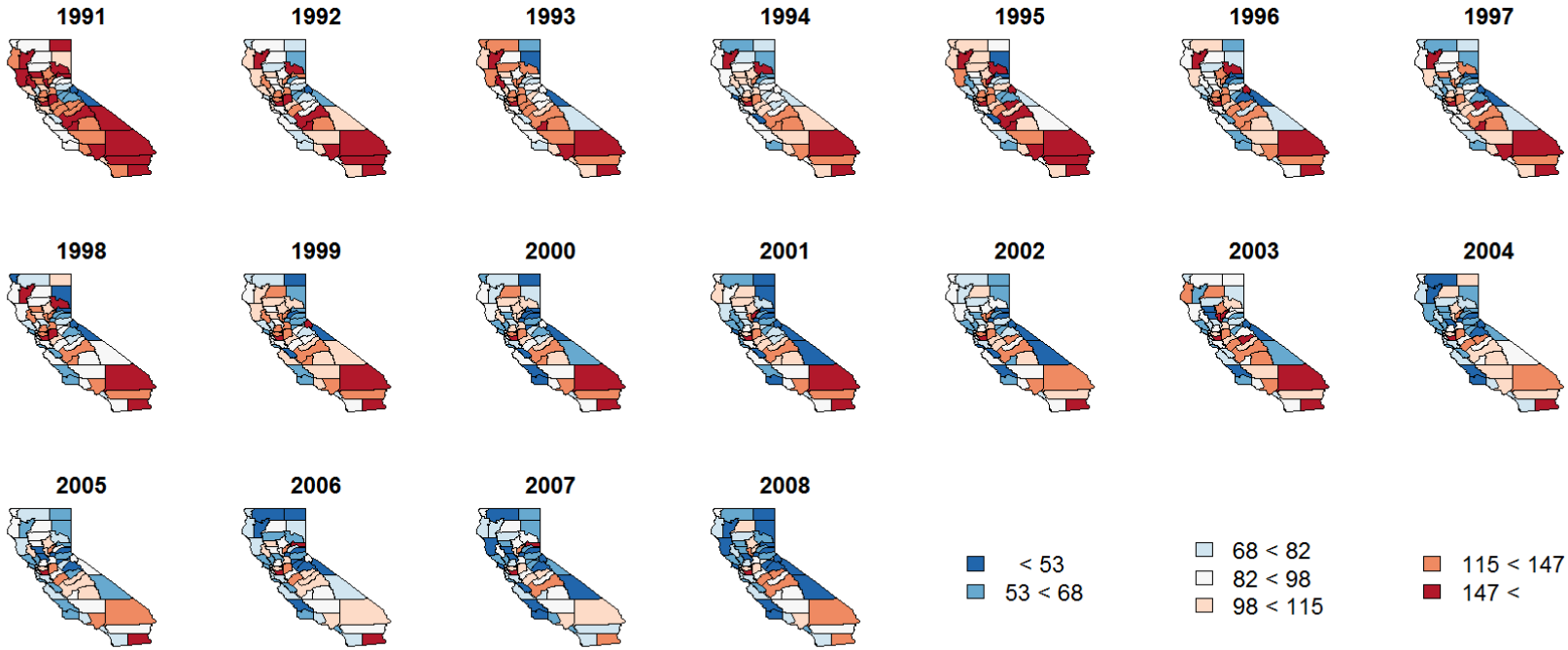


Figure: Raw asthma hospitalization rates, aggregated over year

Introduction – The Data

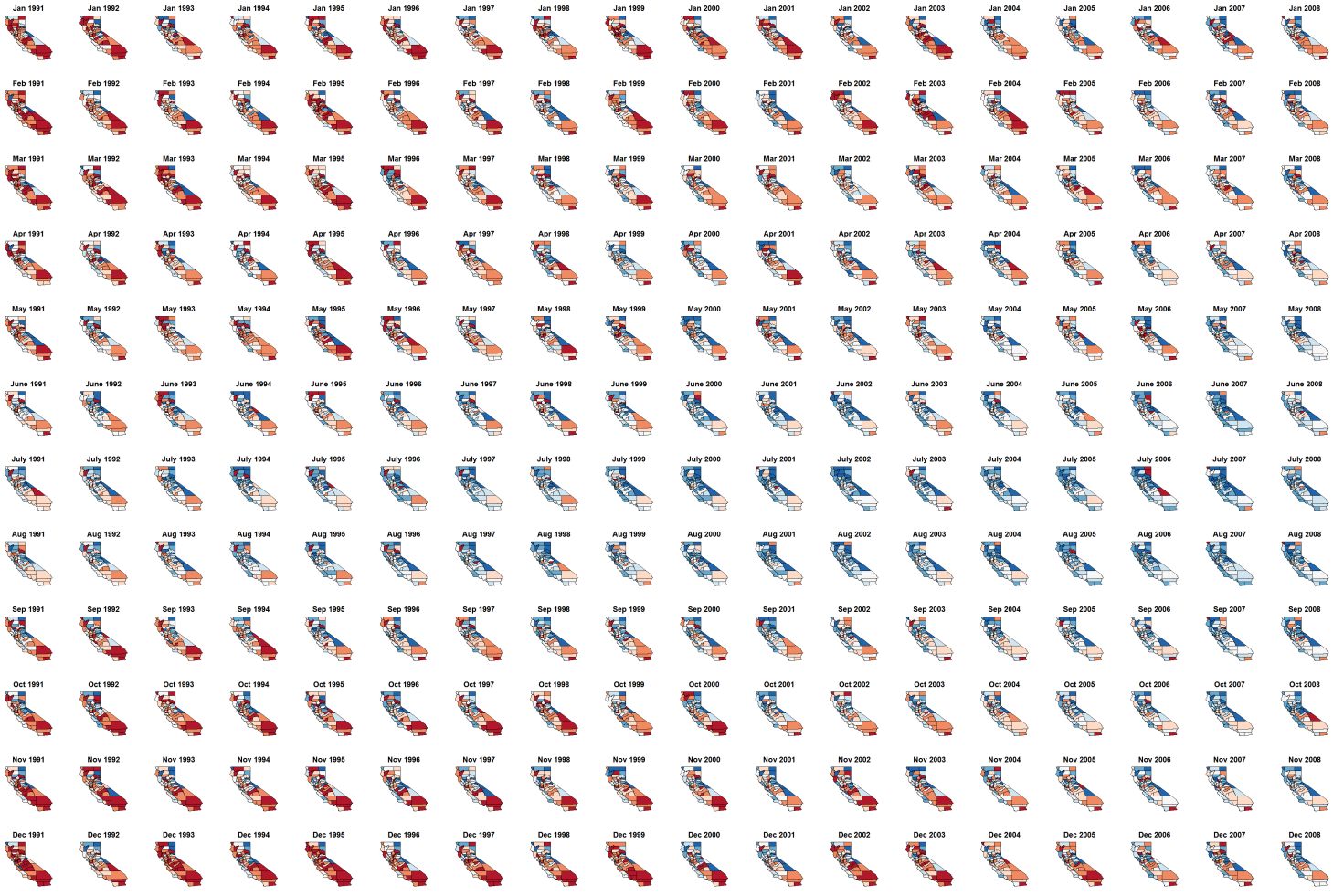


Figure: Raw asthma hospitalization rates

Introduction—Covariates

Based on the asthma literature, we consider the following covariates:

- To account for seasonality, we include monthly fixed effects (using January as a baseline)
- Population density, % Black, % Under 18
  - Using 2000 U.S. Census
  - Vary spatially, but not temporally
- Ozone level (spatiotemporally varying)
  - From California Environmental Protection Agency
  - Number of days per month exceeding the state 8 hour standard for acceptable ozone levels
  - Compiled at the *air basin* level (regions with similar meteorological and geographic conditions throughout)

Introduction – Covariates

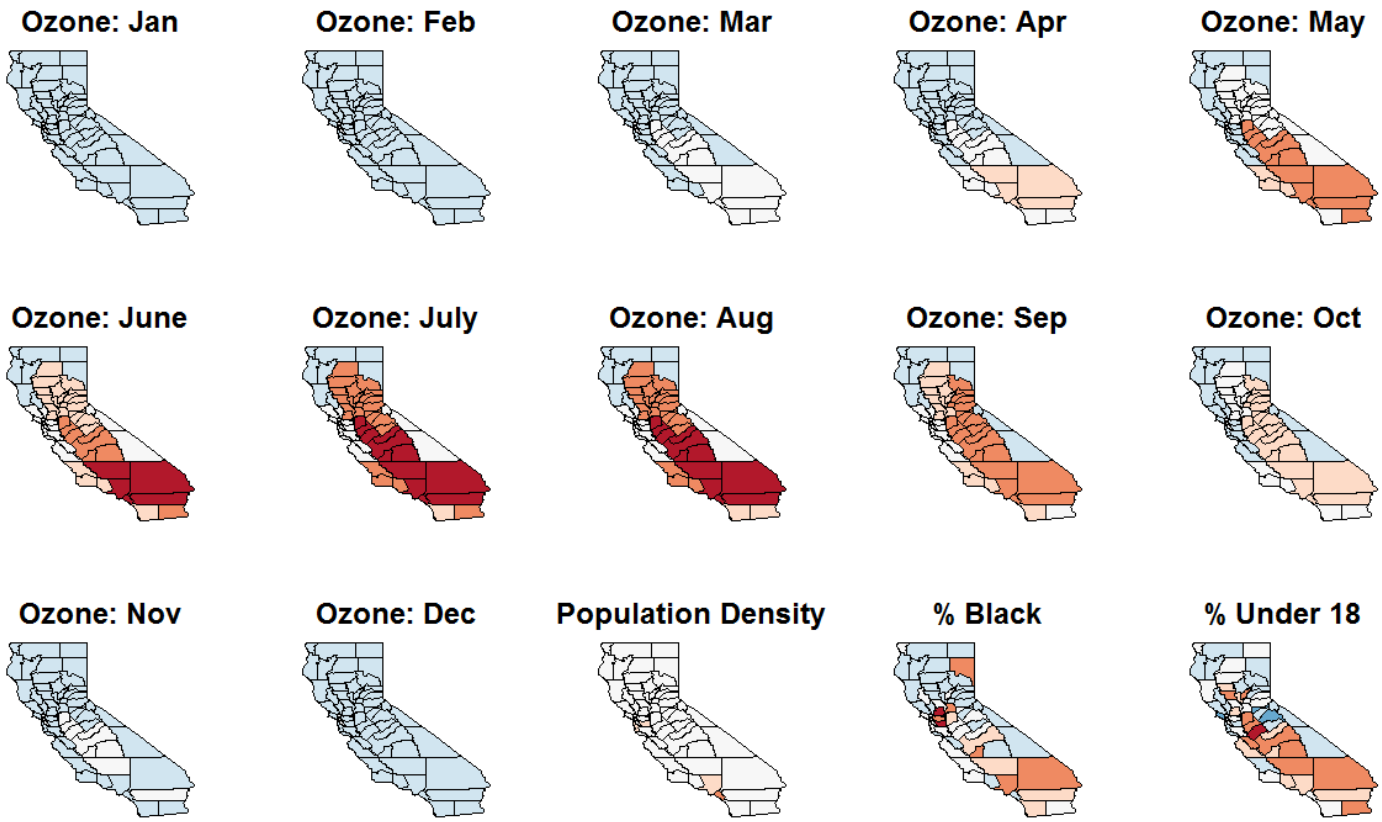


Figure: Asthma hospitalization covariates, standardized with mean 0 and variance 1. Colors range from dark blue to dark red, with cutoffs at $(-2, -1.2, -0.4, 0.4, 1.2, 2)$. Note: San Francisco County is significantly more densely populated than any other county, but is too small to be visible.

Areally referenced space-time model

$$Y_i(t) = \mu_i(t) + Z_i(t) + \epsilon_i(t), \quad \epsilon_i(t) \overset{ind}{\sim} N(0, \tau_i^2) \text{ for } i = 1, 2, \ldots, N_s$$

- $\mu_i(t) = \beta_{i0} + \beta_{i1} x_1(t) + \cdots + \beta_{ip} x_p(t)$
- $Z_i(t)$ is a region-specific stochastic process (an unknown function) over time
- $\tau_i^2 > 0$ implies discontinuity in outcome.

Or model the mean in a GLM framework

$$g(E[Y_i(t)]) = \mu_i(t) + Z_i(t), \quad \text{for } i = 1, 2, \ldots, N_s$$

We posit that $Z_i(t)$'s for neighboring i's will be similar.

Temporal gradients

- Temporal finite differences:

$$Z_i(t_0) = \frac{Z_i(t_0 + h) - Z_i(t_0)}{h}\ ,\ i = 1, 2, \ldots, N_s$$

- Temporal gradient:

$$\frac{d}{dt} Z_i(t) = Z_i'(t_0) = \lim_{h \to 0} \frac{Z_i(t_0 + h) - Z_i(t_0)}{h}$$

- Why are we interested in these quantities?
  - High gradients = "outliers"
  - Can identify lurking covariates that affect response through local change
  - Helps in policy formulation and hospital administration

Example: Demonstrate ability to capture gradients

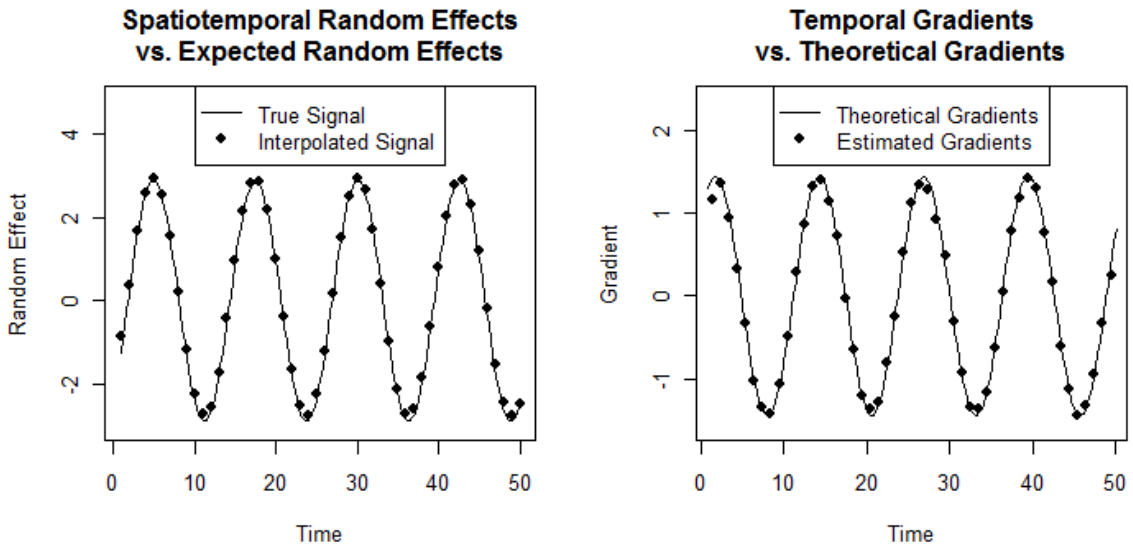


Figure: Simulation Results: plots of the spatiotemporal random effects and temporal gradients.

Political maps as a graph

- A (political) map (i.e. one showing territorial borders) can be fully described as an algebraic graph $G = (V, E)$
- V is the set of nodes = regions/territories
- E is the set of edges = "is a neighbor of" (symmetric relation)
- There are no self-edges: no region is a neighbor of itself
- Let $W = \{w_{ij}\}$ be the *adjacency matrix* of G: $w_{ij} = 0$ if regions i and j are not neighbors and $\neq 0$ (usually > 0) when i and j are neighbors, denoted $i \sim j$.
- Construct $D = \mathrm{diag}(w_{1+}, w_{2+}, \ldots . w_{N_s+})$; $w_{i+} = \sum_{j=1}^{N_s} w_{ij}$
- Diagonal elements in D = number of neighbors of that region.
- If a region is an "island" the corresponding element in D is zero.

The Laplacian of a graph

- The Laplacian of a connected graph (map) is:

$$D - \alpha W = D^{1/2}(I - \alpha D^{-1/2} W D^{-1/2}) D^{1/2}$$

- If λ_{min} and λ_{max} are the minimum and maximum eigenvalues of $D^{-1/2} W D^{-1/2}$, then

$$\lambda_{\text{min}} < 0 \quad \text{and} \quad \lambda_{\text{max}} = 1 \, .$$

- The Laplacian is positive definite if and only if:

$$\alpha \in \left(\frac{1}{\lambda_{\text{min}}}, 1 \right) \, . \quad \text{It is singular if} \quad \alpha = 1 \, .$$

- Laplacian is p.d. if and only if $(D - \alpha W)^{-1}$ is p.d.
- Two potential candidates for modeling variance-covariances.
- For a map with islands: build p.d. Laplacians from connected components.

Markov random field for each time point

- A geographical map is an algebraic graph with nodes = regions and edges = "is a neighbor of"
- Adjacency matrix: $W = \{w_{ij}\}$; $w_{ij} = 0$ if regions i and j are not neighbors and 1 when regions i and j are neighbors, denoted $i \sim j$.
- Conditional distribution for each $Z_i(t)$:

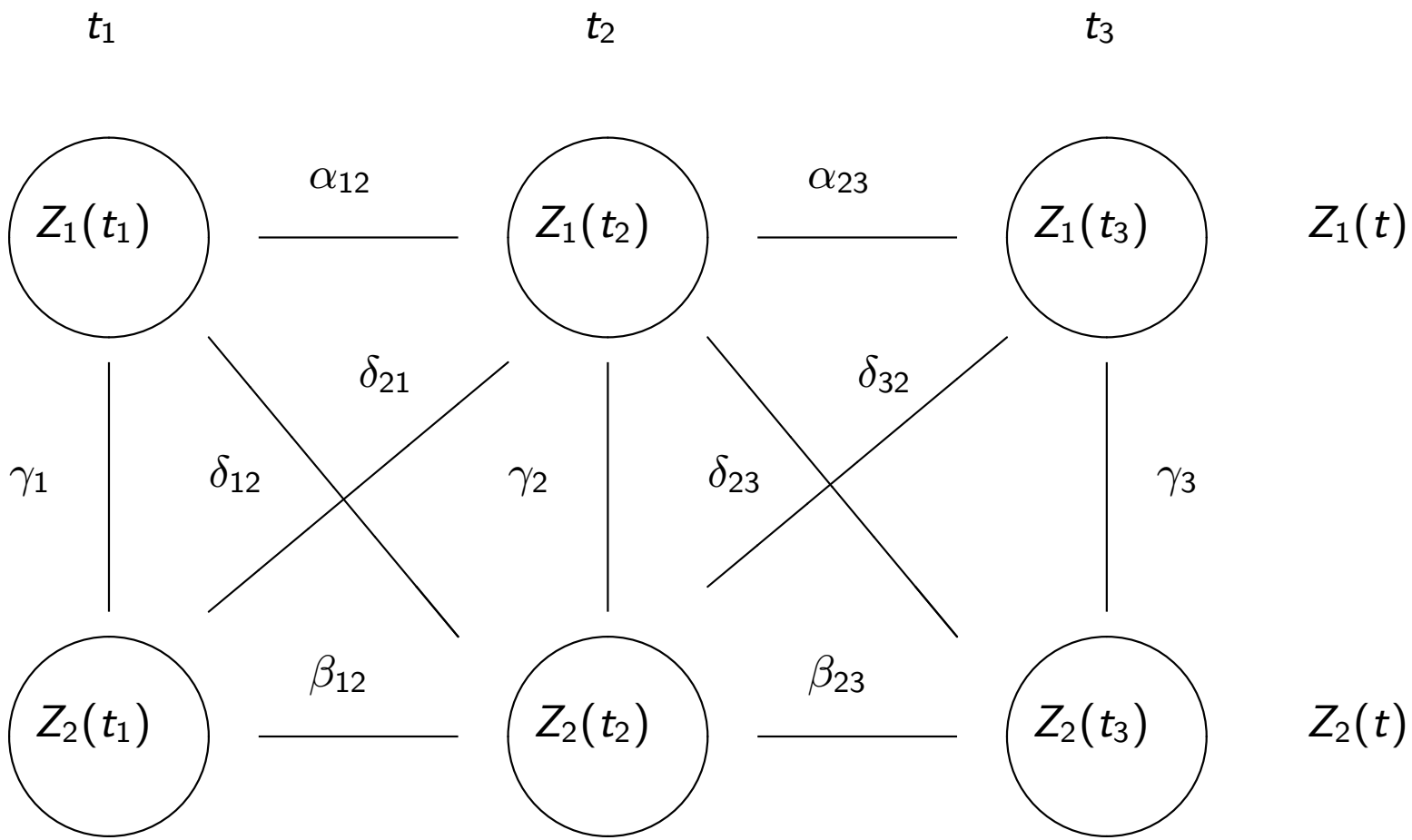
$$p(Z_i(t) \,|\, \{Z_{j\neq i}(t)\}) \sim N\left(\sum_{j\sim i} \alpha \frac{w_{ij}}{w_{i+}} Z_j(t), \frac{\sigma^2}{w_{i+}}\right) ,$$

  where $w_{i+} = \sum_{j\sim i} w_{ij}$, $\sigma^2 > 0$, and $\alpha \in (0, 1)$ ensures "propriety".
- If $\mathbf{Z}(t) = (Z_1(t), Z_2(t), \ldots, Z_{N_s}(t))^\top$, then

$$\mathbf{Z}(t) \sim N\left(\mathbf{0}, \sigma^2(D - \alpha W)^{-1}\right) \; ; \; D = \text{diag}\{w_{i+}\} .$$

Note: the MRF above does not model temporal dependence yet.

Complex Spatial Dependencies



Data Analysis – Model Comparisons

	p_D	DIC*	G-R*
Simple Linear Regression	79	9,894	-16,166
Random Intercept and Slope	165	4,347	-10,403
CAR Model	117	7,302	-13,436
Areally Referenced Gaussian Process	5,256	0	0

Table: Comparisons between our areally referenced Gaussian process model and the three alternatives. Smaller DIC values indicates a better trade-off between model fit and model complexity, as do larger (less negative) Gneiting-Raftery (G-R) scores. Both DIC and G-R shown are standardized relative to our areally referenced Gaussian Process model.

Data Analysis—Parameter Estimates

Using the covariates listed earlier, our $x_i(t_j)$ is a 16×1 vector

Parameter	Median (95% CI)	Parameter	Median (95% CI)
β_0 (Intercept)	9.17 (8.93, 9.42)	β_{10} (July)	-3.78 (-4.21, -3.37)
β_1 (Pop Den)	0.60 (0.49, 0.70)	β_{11} (August)	-3.58 (-4.02, -3.13)
β_2 (Ozone)	-0.18 (-0.28, -0.08)	β_{12} (September)	-1.96 (-2.37, -1.54)
β_3 (% Under 18)	1.24 (1.15, 1.34)	β_{13} (October)	-1.36 (-1.73, -1.00)
β_4 (% Black)	1.12 (1.01, 1.24)	β_{14} (November)	-0.71 (-1.02, -0.42)
β_5 (February)	-0.25 (-0.46, -0.04)	β_{15} (December)	0.63 (0.41, 0.86)
β_6 (March)	-0.21 (-0.48, 0.07)	ϕ	0.90 (0.84, 0.97)
β_7 (April)	-1.47 (-1.81, -1.12)	α	0.77 (0.71, 0.80)
β_8 (May)	-1.17 (-1.53, -0.8)	σ^2	21.52 (20.18, 23.06)
β_9 (June)	-2.79 (-3.21, -2.4)	$\bar{\tau}_\cdot^2$	3.32 (0.18, 213.16)

Table: Parameter estimates for asthma hospitalization data, where estimates for $\bar{\tau}_\cdot^2$ represent the median (95% CI) for all of the τ_i^2

Data Analysis—Parameter Estimates

Parameter	Median (95% CI)	Parameter	Median (95% CI)
β_0 (Intercept)	9.51 (8.46, 10.51)	$\beta_{15} - \beta_{26}$ (Ozone)	
β_1 (Pop Den)	0.63 (0.55, 0.71)	— January	0.51 (-0.95, 1.94)
β_2 (% Black)	1.23 (1.13, 1.33)	— February	0.39 (-0.61, 1.53)
β_3 (% $<$ 18)	1.24 (1.13, 1.34)	— March	0.42 (-0.05, 0.89)
β_4 (Feb)	-0.36 (-1.49, 0.85)	— April	0.21 (-0.05, 0.49)
β_5 (Mar)	-0.24 (-1.32, 0.83)	— May	-0.17 (-0.33, 0.00)
β_6 (Apr)	-1.60 (-2.66, -0.51)	— June	-0.36 (-0.53, -0.20)
β_7 (May)	-1.39 (-2.46, -0.30)	— July	-0.22 (-0.35, -0.09)
β_8 (June)	-2.46 (-3.59, -1.37)	— August	-0.20 (-0.33, -0.07)
β_9 (July)	-3.29 (-4.47, -2.19)	— September	-0.28 (-0.42, -0.12)
β_{10} (Aug)	-3.16 (-4.33, -2.08)	— October	0.06 (-0.13, 0.25)
β_{11} (Sep)	-1.94 (-3.03, -0.88)	— November	0.52 (0.03, 1.05)
β_{12} (Oct)	-1.78 (-2.82, -0.70)	— December	3.15 (1.43, 5.08)
β_{13} (Nov)	-0.87 (-1.94, 0.24)	α	0.88 (0.85, 0.90)
β_{14} (Dec)	2.42 (1.12, 3.64)	ϕ	1.24 (1.18, 1.30)

Table: Posterior medians and 95% credible intervals (CI) for $\boldsymbol{\beta}$ and ϕ from our asthma hospitalization rate data.

Data Analysis – Spatiotemporal Random Effects

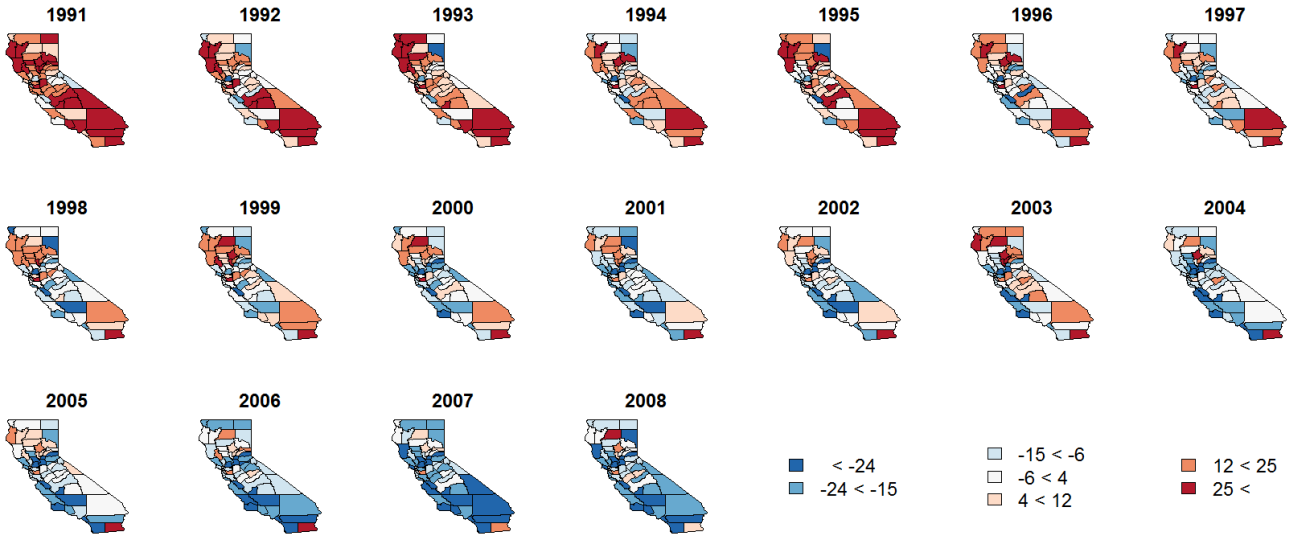


Figure: Spatiotemporal random effects for asthma hospitalization data, by year

Data Analysis – Temporal Gradients

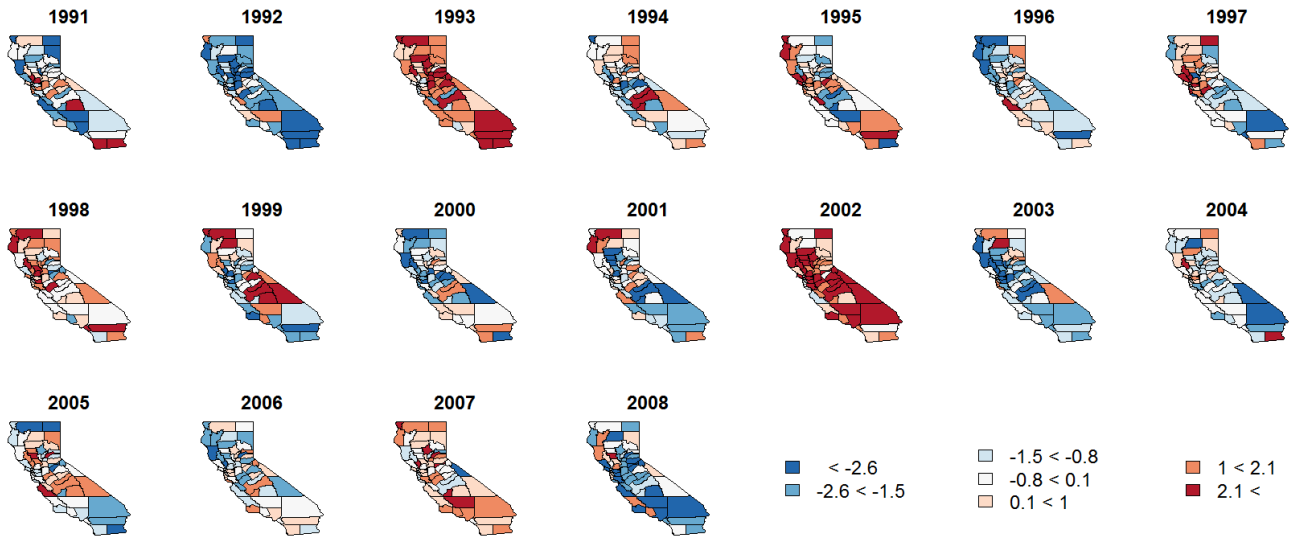


Figure: Temporal gradients for asthma hospitalization data, by year

Data Analysis – Los Angeles County vs. San Francisco County

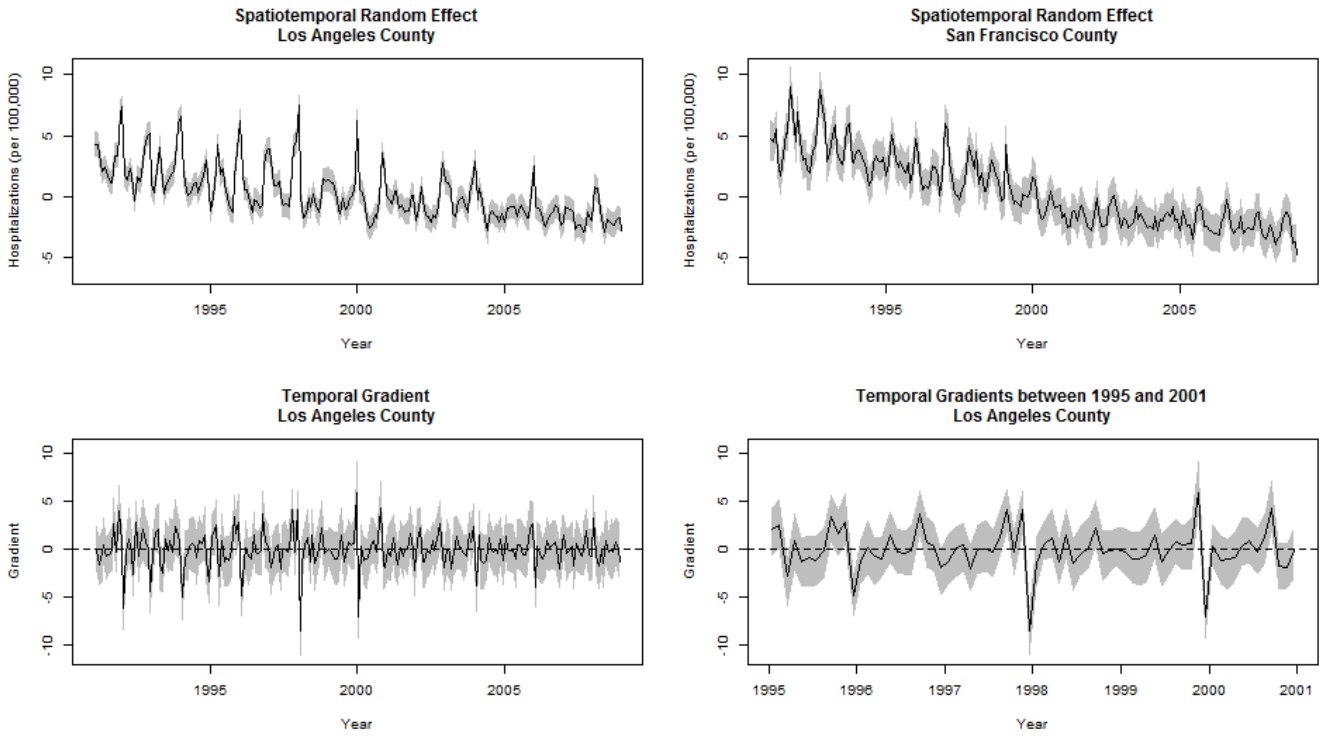


Figure: Comparison between the spatiotemporal random effects in Los Angeles and San Francisco Counties, and an investigation of temporal gradients in Los Angeles County.

Data Analysis – Los Angeles County

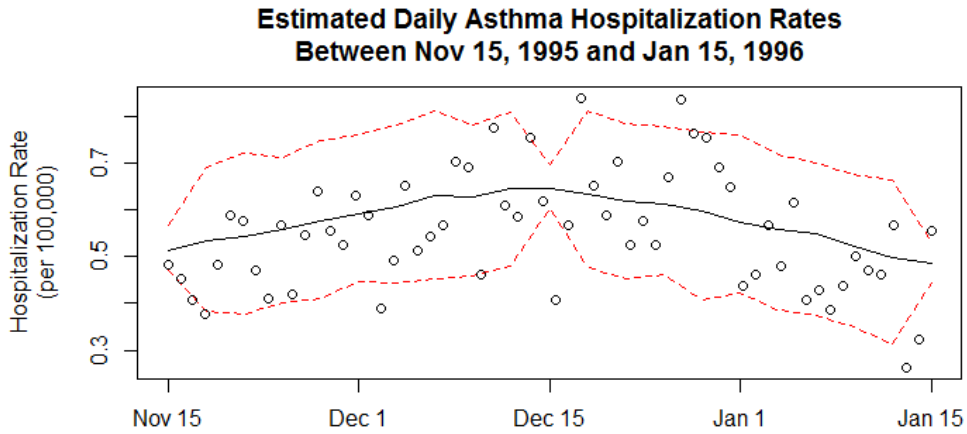
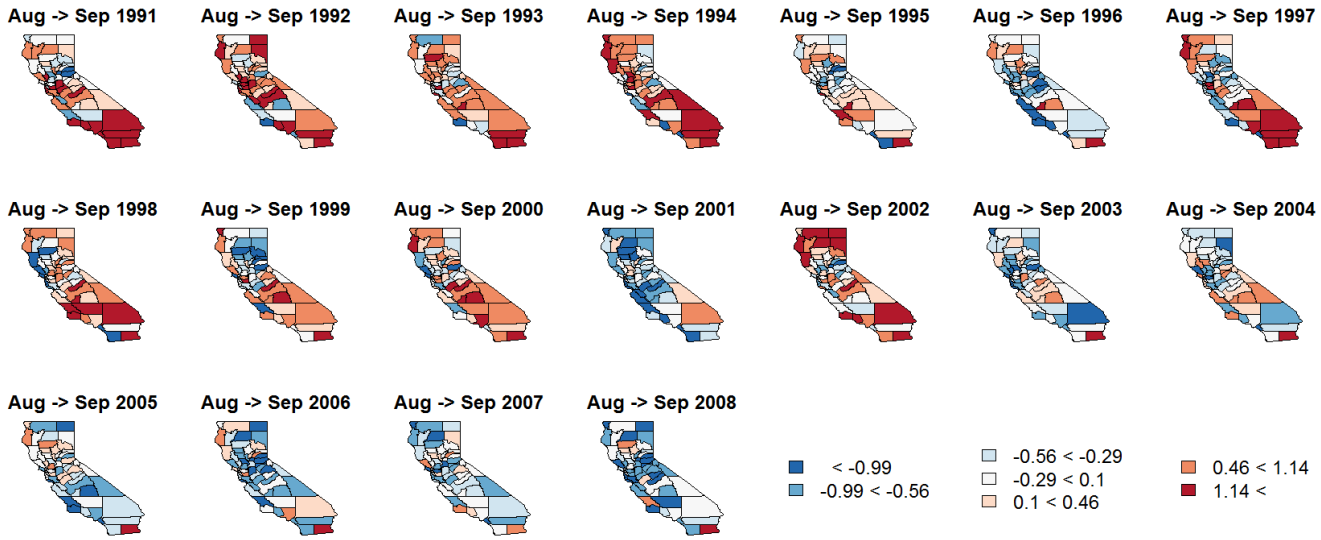


Figure: Posterior predicted curves (and 95% credible bounds) for the daily asthma hospitalization rates in Los Angeles County between November 15, 1995 to January 15, 1996. This county and interval was selected due the presence of a significantly positive gradient between November and December and a significantly negative gradient between December and January. The true hospitalization are also shown for comparison purposes, though the model was fit using only the monthly aggregates.

Data Analysis – August → September

Note that, on average, September has 1.62 more hospitalizations per 1,000 people than August



This figure indicates this difference is decreasing over time.
A similar phenomenon is occurring between March and April, as well.

- The seasonality in the data appears to diminish over time
- Winter is becoming more like summer, or at least its effect on asthma hospitalizations is

Conclusion

- Using a continuous-time model permits inference at a resolution finer than that of the observed data.
- Insight can be gained from an assessment of temporal gradients in the residual process.
  - could be used to motivate search for temporally interesting covariates not included in our model.
- Statistical Significance
  - While possible to identify statistically significant gradients, context is important.
  - Here, gradients are computed between time points which we believe are "different" (due to seasonality)
- Models for regionally-referenced functional data.

Thank you!