Spatial statistics in public health research: methodological opportunities and computational challenges

Chris Paciorek and Louise Ryan

January 14, 2004

Department of Biostatistics

Harvard School of Public Health

www.biostat.harvard.edu/~paciorek

Outline

- explosion of spatial data in health research
- examples of spatial health data
- modelling spatial risk in a case-control study
  - focus on computational efficiency
- methodological and research challenges

Increased attention to spatial analysis in public health

- areal data:
  - public databases and geocoding of individuals to areas
  - interest in health disparities and social science questions
  - focus is on covariates, not spatial structure

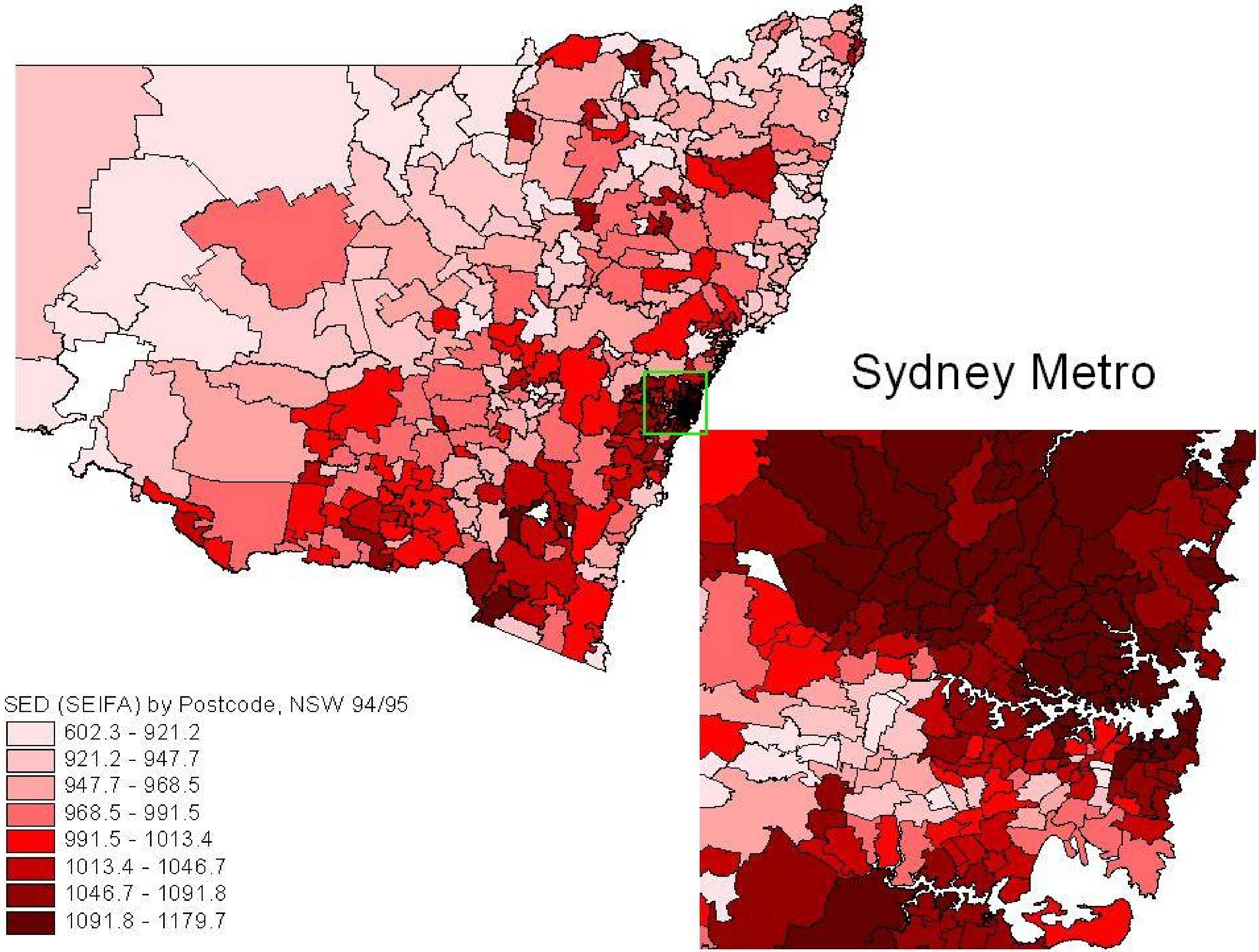
- point data
  - geocoding and GPS are mainstream
    - health outcomes can be assigned point locations
  - GIS software
    - easy data management and manipulation
    - graphical presentation
    - spatially-varying covariate generation
  - strong applied interest in kriging and related smoothing methods
  - opportunities for more sophisticated spatio-temporal modelling, particularly Bayesian models

- environmental exposure modelling
  - spatial smoothing and additive modelling of monitoring data

- mixed point and area data
  - individual locations plus area-level covariates

- multivariate responses
  - multiple pollutants, multiple health endpoints
  - latent variable modelling, causal relationships

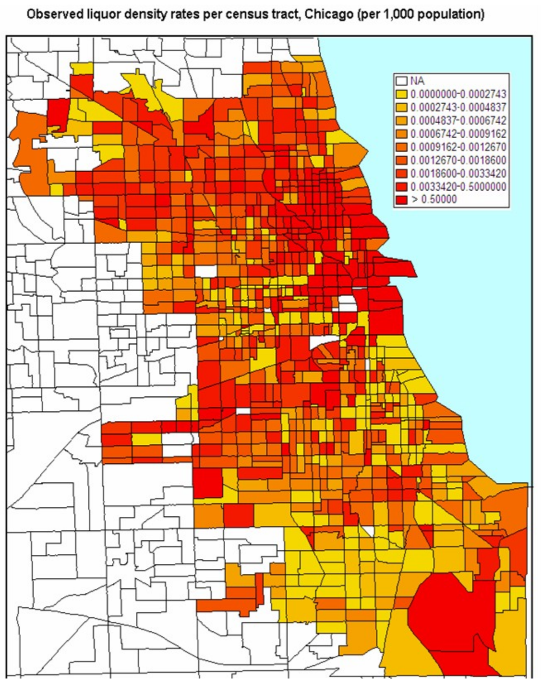
Socioeconomic factors in health outcomes in NSW, Australia



- challenges
  - areal (postcode) units vary drastically in size
  - computational challenge
    - 650 units, 5 years daily data, 2 sexes, 9 age groups
  - spatial effect and spatially-varying covariates hard to tease apart
  - data misalignment
    - outcome at postcode, covariate at census analogue

- relate areal data to a latent smooth process (Kelsall & Wakefield, Rathouz)

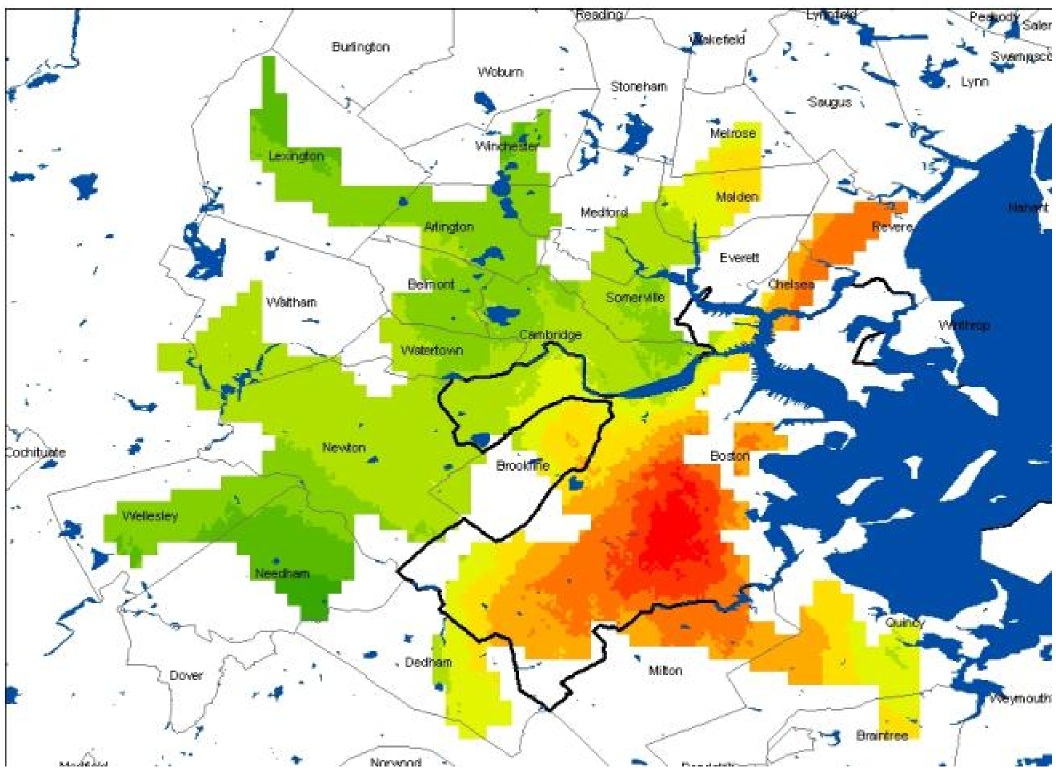
Combining area and individual-level information



- area-level covariates based on point process data
  - access to contraception at health clinics in Malawi
  - accessibility of liquor retail outlets in Chicago
- spatial scale of interest is based on outcome
- consider two-stage Bayesian model so smoothing is informed by the health outcome

Spatial variation in allergenic response

- geocoding of new mothers' residences
- measurement of blood serum IgE immune response
- interest in variance partitioning

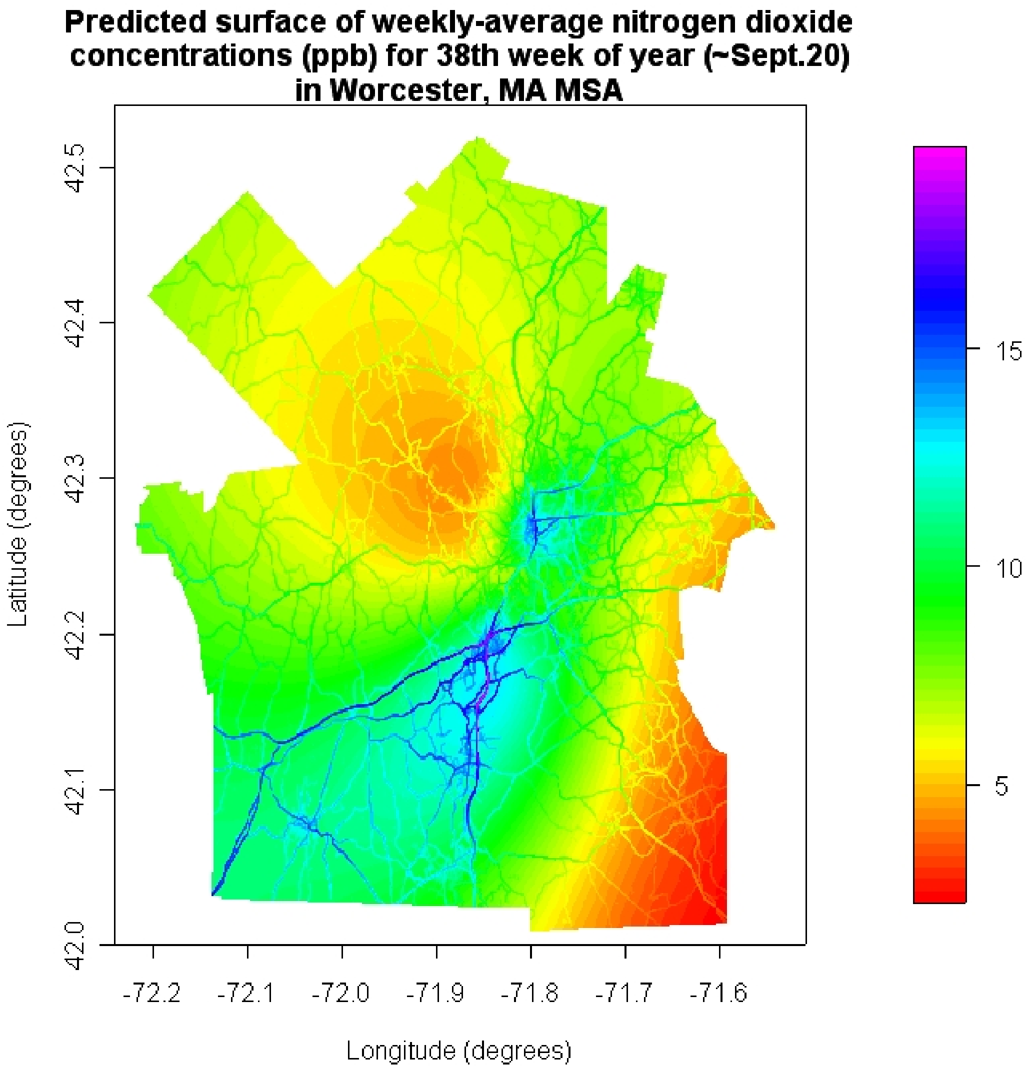


Exposure estimation in the Nurses' Health Study

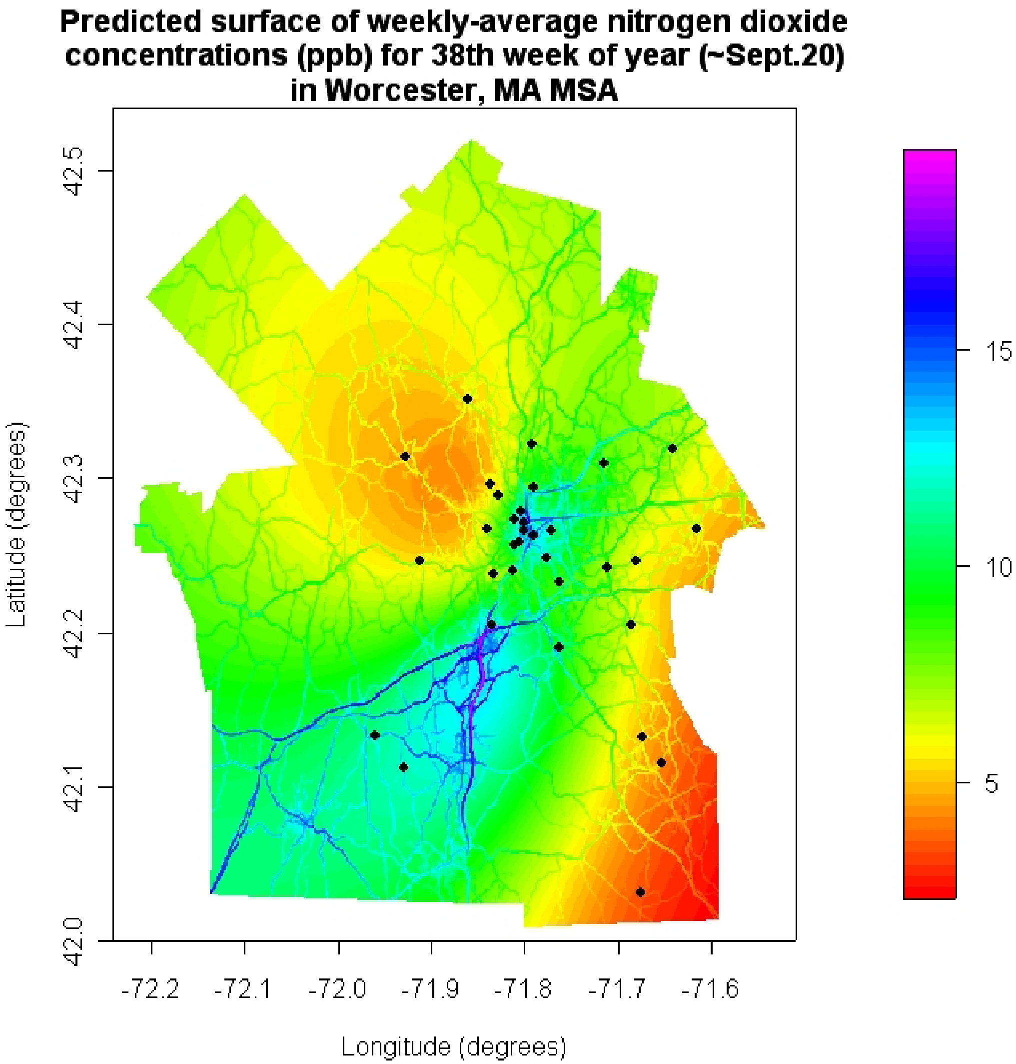
- spatial estimation of individual environmental exposures
  - often air pollution

- particulate matter (PM) exposure in large cohort of nurses
  - estimate individual exposure, 1985-2003
  - EPA monitoring for large-scale spatio-temporal heterogeneity
  - spatially-varying covariates for local heterogenity
    - distance to roads, climate variables, local land use, ...
    - generated using GIS
  - geocoding of individual residences every two years
    - relate estimated exposure to health outcomes (chronic heart disease)

- geocoding and GIS make this possible; spatial statistics provides a rigorous framework



- geocoding and GIS make this possible; spatial statistics provides a rigorous framework for estimation

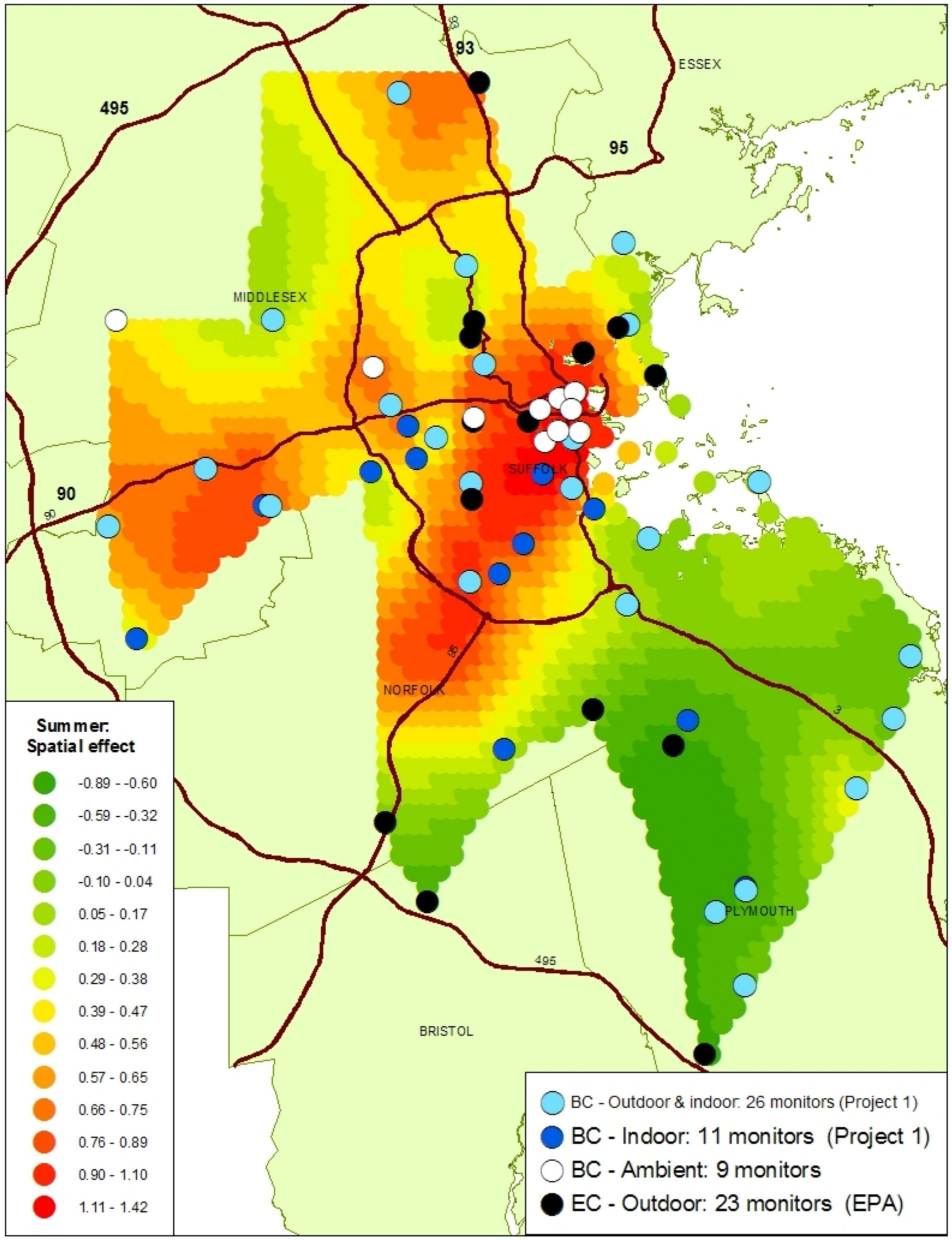


Challenges for spatio-temporal exposure estimation

- computations: 50,000 monthly pollution measurements over 20 years at 500 monitoring sites
  - kriging is difficult, particularly Bayesian implementations
  - efficient, user-friendly computation is critical (gam() in R)
  - more complicated spatio-temporal structures for better prediction, but ...
    * Bayesian implementation would require a statistician
    * more computationally efficient methods needed
- non-standard measurement error results from smoothing
- multivariate, non-Gaussian modelling
  - modelling PM2.5 based on PM10 and on airport visibilility
  - simple multivariate normality not reasonable

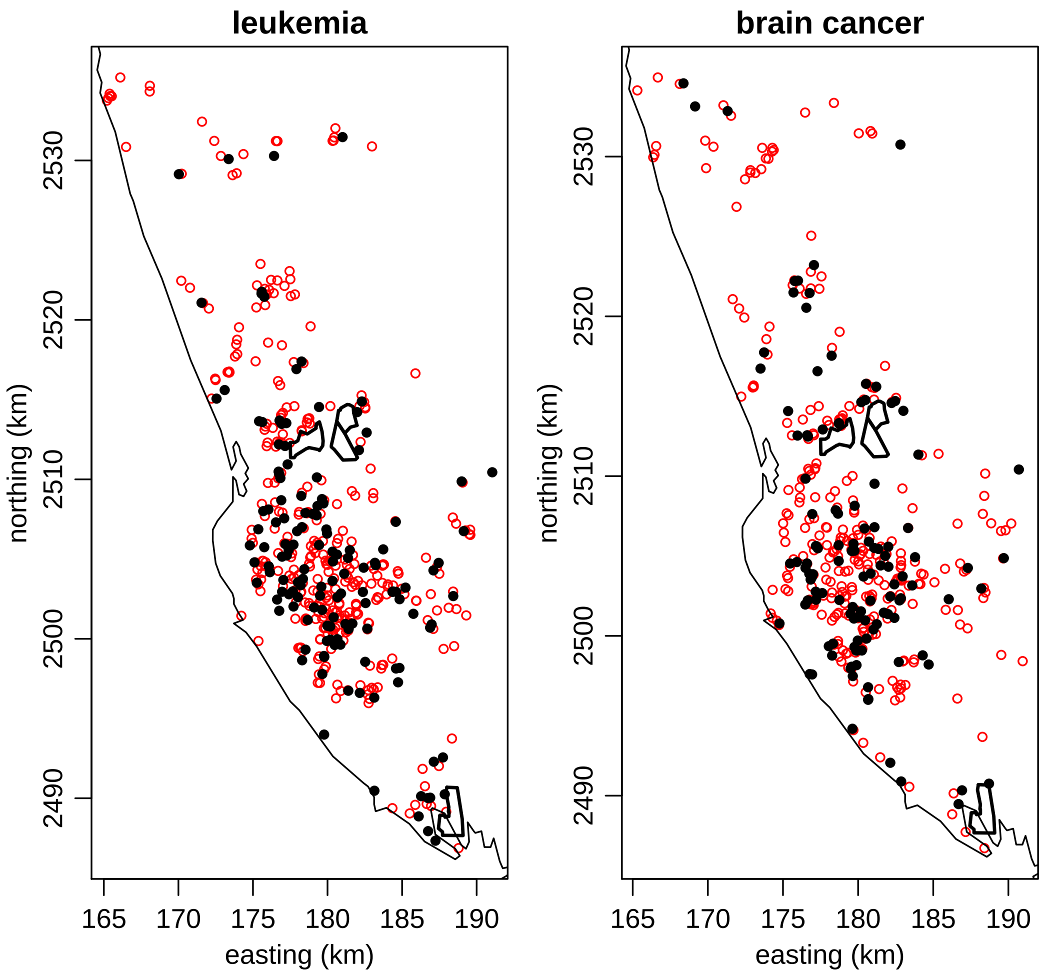
Latent variable modelling

- exposure estimation for PM in the Boston area
- which pollutant sources are responsible for health outcomes?
  - traffic is locally heterogeneous, power plant pollutants (e.g., sulfates) are not
- estimate latent traffic exposure and relate to health outcomes
- two surrogates for traffic, elemental carbon and black carbon
- hierarchical Bayesian model with multiple data sources



93
495
95
ESSEX
MIDDLESEX
90
SUFFOLK
NORFOLK
PLYMOUTH
BRISTOL
495
Summer:
Spatial effect
-0.89 - -0.60
-0.59 - -0.32
-0.31 - -0.11
-0.10 - 0.04
0.05 - 0.17
0.18 - 0.28
0.29 - 0.38
0.39 - 0.47
0.48 - 0.56
0.57 - 0.65
0.66 - 0.75
0.76 - 0.89
0.90 - 1.10
1.11 - 1.42
BC - Outdoor & indoor: 26 monitors (Project 1)
BC - Indoor: 11 monitors (Project 1)
BC - Ambient: 9 monitors
EC - Outdoor: 23 monitors (EPA)

Petrochemical exposure in Kaohsiung, Taiwan



$n = 495$ $\quad$ $n = 433$

$n_1 = 141$ $\quad$ $n_1 = 121$

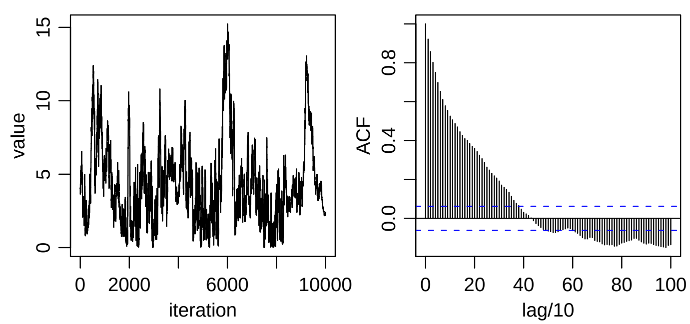
Possible approaches for health analysis

- Explicitly estimate pollutant exposure - difficult retrospectively
- Use distance to exposure source as covariate
- Use a moving window/multiple testing to detect clusters of cases
  - default approach - software available
- **Include space as a covariate to provide a map of risk**

$$
\begin{aligned}
Y_i &\sim \text{Ber}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) \\
\text{logit}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) &= \boldsymbol{x_i}^T \boldsymbol{\beta} + g_\theta(\boldsymbol{s_i})
\end{aligned}
$$

Modelling challenges from a Bayesian perspective

- thousands of case-control observations - difficult for Bayesian kriging
- non-Gaussian spatial models particularly difficult
  - spatial process cannot be analytically integrated out of the likelihood/posterior
  - MCMC mixing is very slow because of high-level structure
    - ∗ correlation amongst process values and between process values and process hyperparameters



Modelling Framework

$$\begin{aligned} Y_i &\sim \text{Ber}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) \\ \text{logit}(p(\boldsymbol{x_i}, \boldsymbol{s_i})) &= \boldsymbol{x_i}^T \boldsymbol{\beta} + g_\theta(\boldsymbol{s_i}) \end{aligned}$$

- basic spatial model for $\boldsymbol{g}_\theta^s = (g_\theta(\boldsymbol{s_1}), \ldots, g_\theta(\boldsymbol{s_n}))$
  - GAM: $g_\theta(\cdot)$ is a two-dimensional smooth term
    - basis representation

      $$\boldsymbol{g}_\theta^s = Z\boldsymbol{u}$$

    - Gaussian process representation:

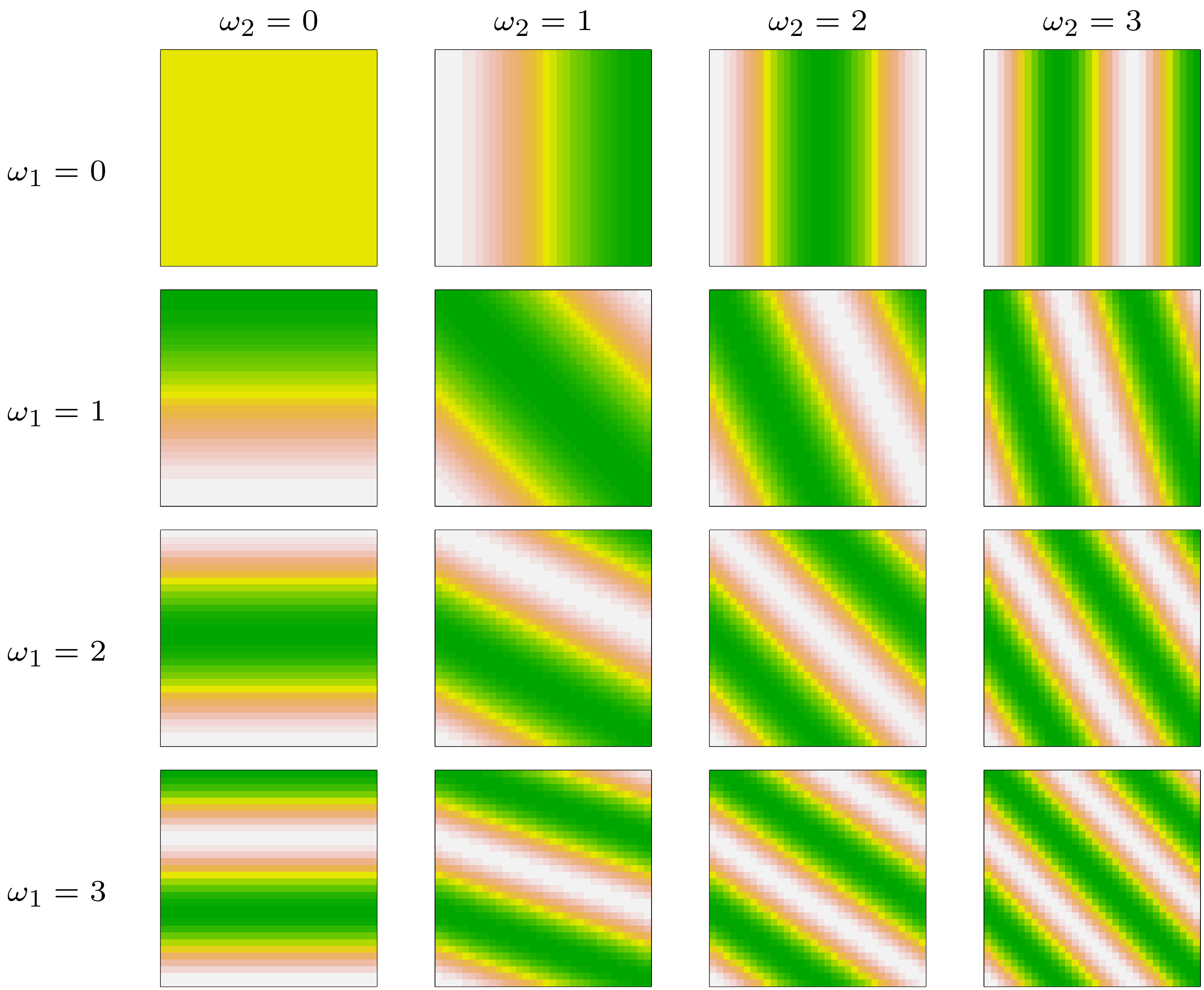
      $$g(\cdot) \sim \text{GP}(\mu(\cdot), C_\theta(\cdot, \cdot)) \Rightarrow \boldsymbol{g}_\theta^s \sim N(\boldsymbol{\mu}, C_\theta)$$

  - GLMM: $\boldsymbol{g}_\theta^s = Z\boldsymbol{u}$
    - correlated random effects, $\boldsymbol{u} \sim N(0, \Sigma)$

Bayesian spectral basis function model

- computationally efficient basis function construction (Wikle 2002)
- $\boldsymbol{g}^{\#} = Z\boldsymbol{u}$ and $\boldsymbol{g}^{s} = \sigma P \boldsymbol{g}^{\#}$
  - piecewise constant gridded surface on k by k grid
  - P maps observation locations to nearest grid point
- Z is the Fourier (spectral) basis and $Z\boldsymbol{u}$ is the inverse FFT
- $Z\boldsymbol{u}$ is approximately a Gaussian process (GP) when...
  - $\boldsymbol{u} \sim N(0, \text{diag}(\pi_\theta(\boldsymbol{\omega})))$ for Fourier frequencies, $\boldsymbol{\omega}$
  - spectral density, $\pi_\theta(\cdot)$, of GP covariance function defines $\text{V}(\boldsymbol{u})$

Bayesian spectral basis functions



Comparison with usual GP specification

- usual GP model: $\boldsymbol{g}^s \sim N(\boldsymbol{\mu}, C_\theta)$
  - $O(n^3)$ fitting: $|C_\theta|$ and $C_\theta^{-1}\boldsymbol{g}$
- spectral basis uses FFT
  - $O\left((k^2)\log(k^2)\right)$
  - additional observations are essentially free for fixed grid
  - fast computation and prediction of surface given coefficients
  - a priori independent coefficients give fast computation of prior and help with mixing

Other approaches

- penalized likelihood based on mixed model (radial basis functions) with REML smoothing
  (Kammann and Wand, 2003; Ngo and Wand, 2004) [PL-PQL]

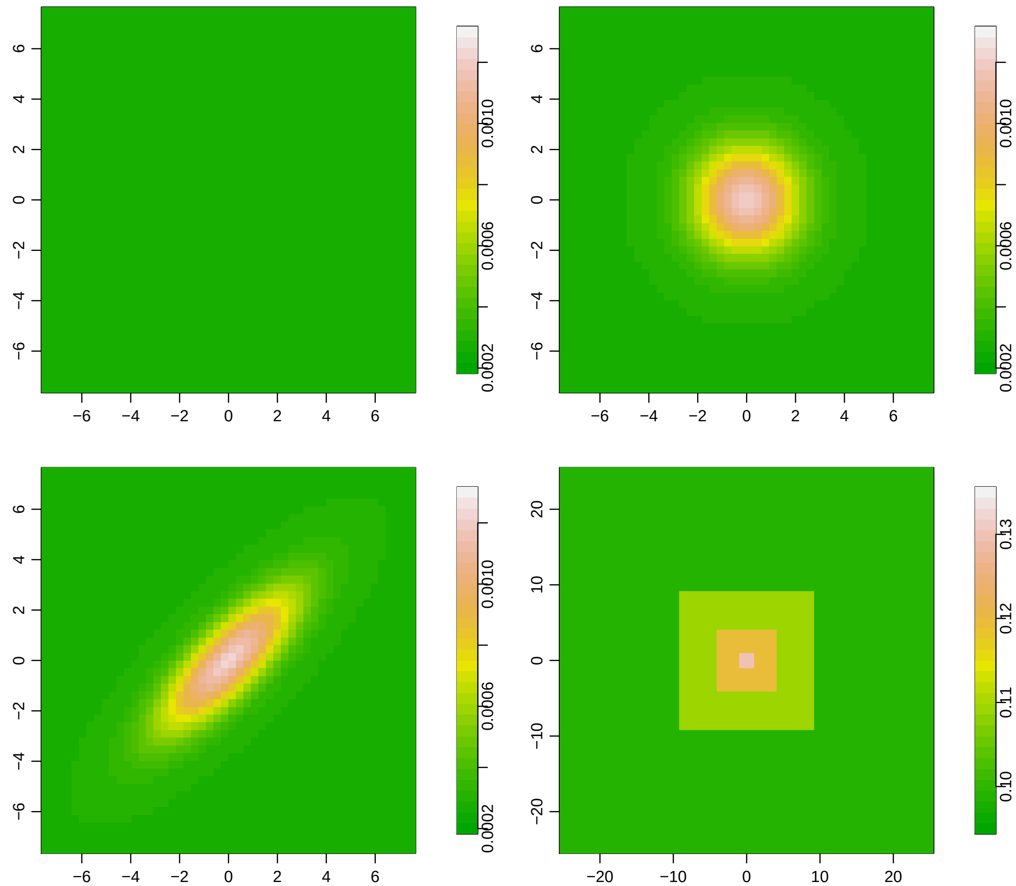
- penalized likelihood with GCV smoothing
  (Wood, 2001, 2003, 2004) [PL-GCV]

- Bayesian mixed model/radial basis functions fit by MCMC
  (Zhao and Wand 2004) [B-Geo]

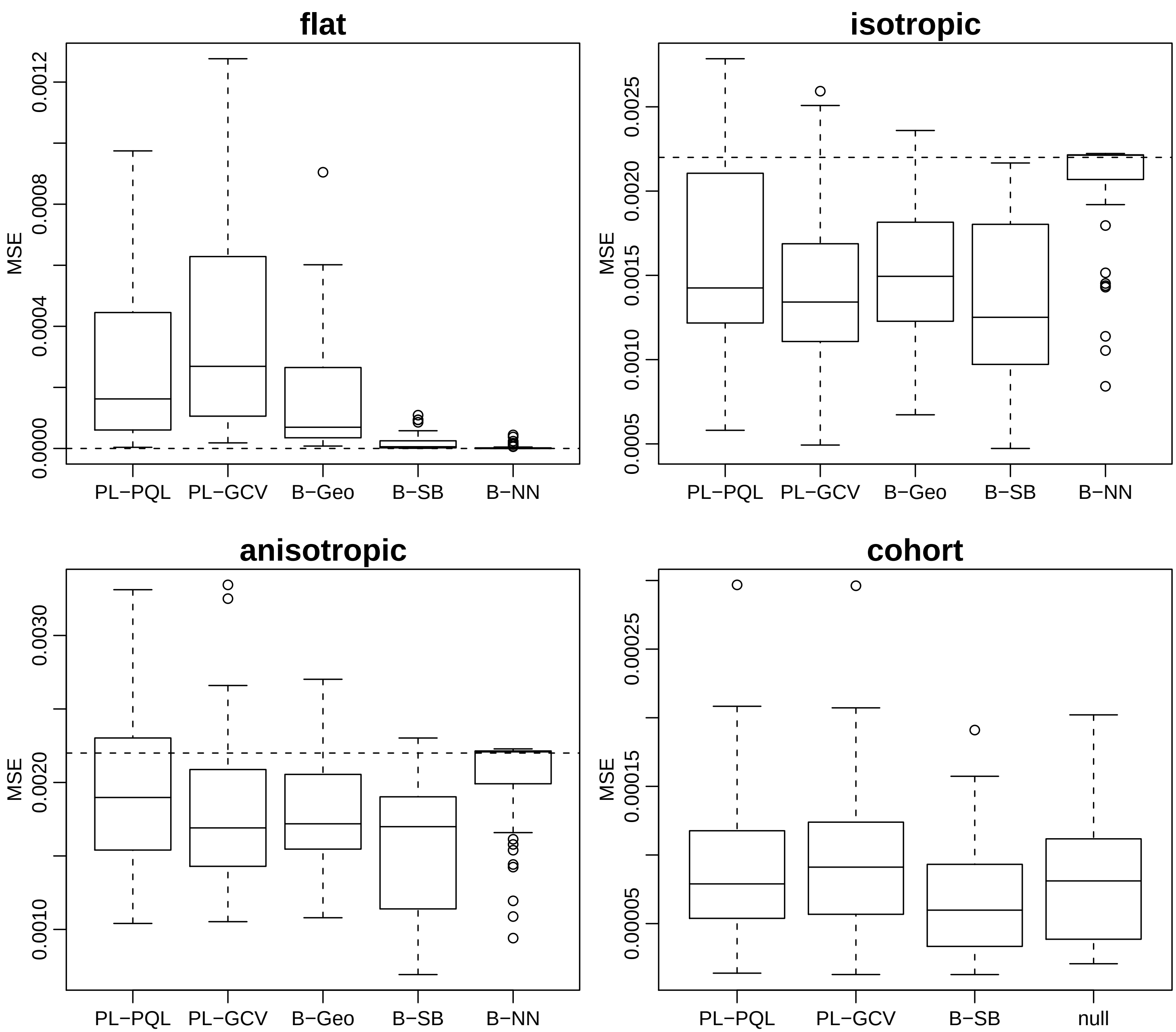
- Bayesian neural network model fit by MCMC
  (R. Neal) [B-NN]

Simulated datasets

- 3 case-control scenarios: $n_0 = 1,000$; $n_1 = 200$; $n_{\text{test}} = 2500$ on 50 by 50 grid
- 1 cohort scenario: $n = 10,000$; $n_{\text{test}} = 2500$ on 50 by 50 grid



Assessment on 50 simulated datasets

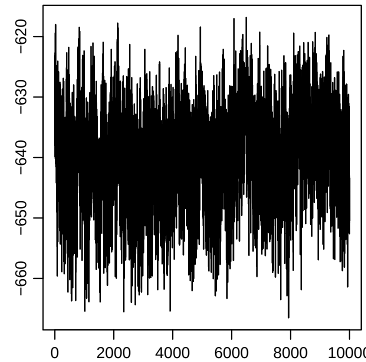


Mixing and speed of Bayesian methods

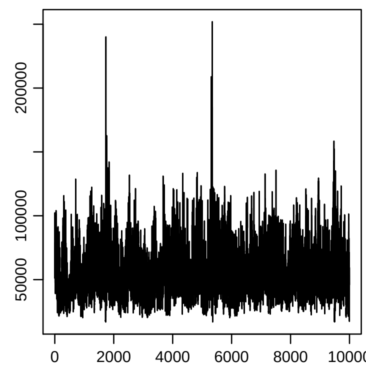
	speed (1000 its)	speed (1000 eff. its)
B-Geo	15 min.	104 hr.
B-SB	1.3 min.	6 hr.

log posterior trace

B-Geo:

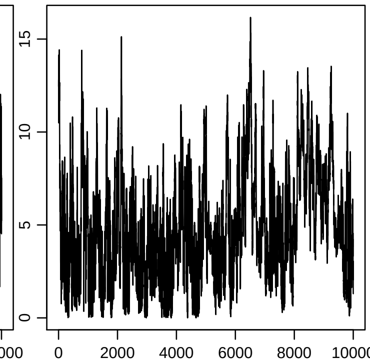


B-SB:

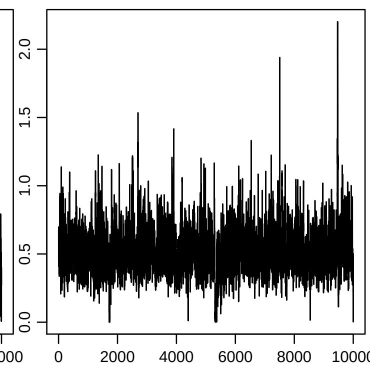


σ trace

B-Geo:

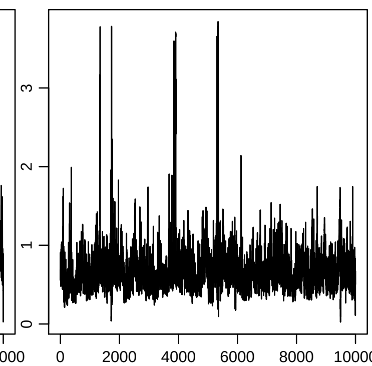


B-SB:



ρ trace

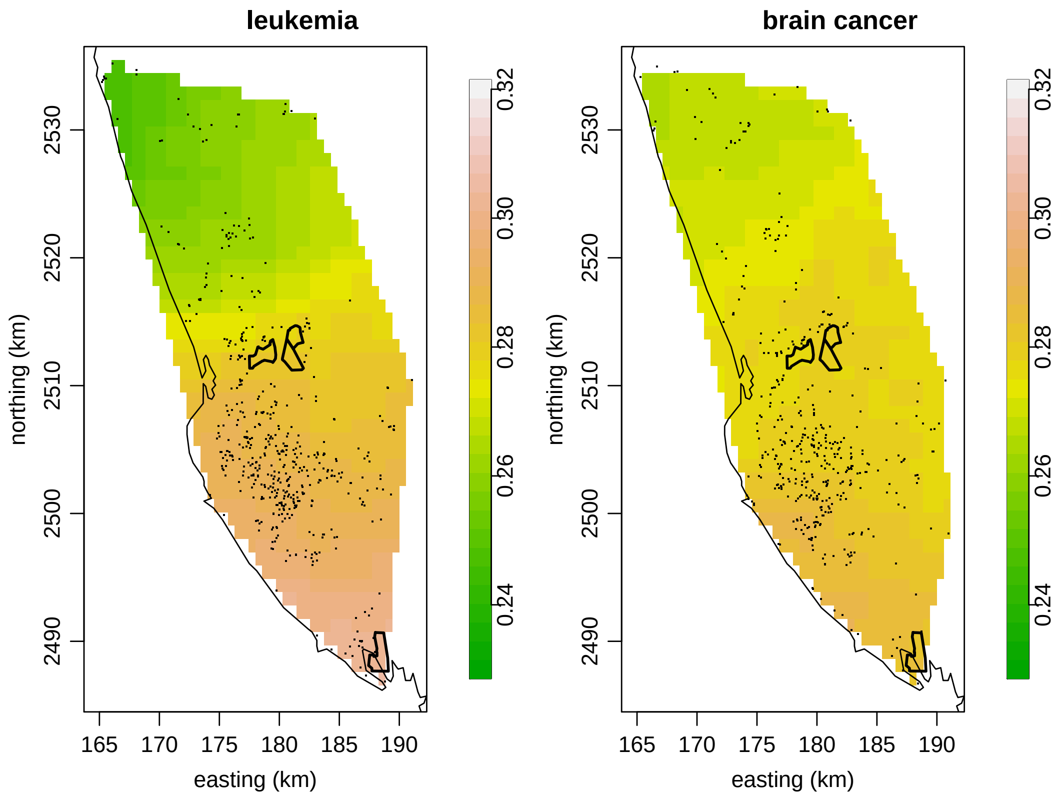
B-SB:



Taiwan revisited - assessment

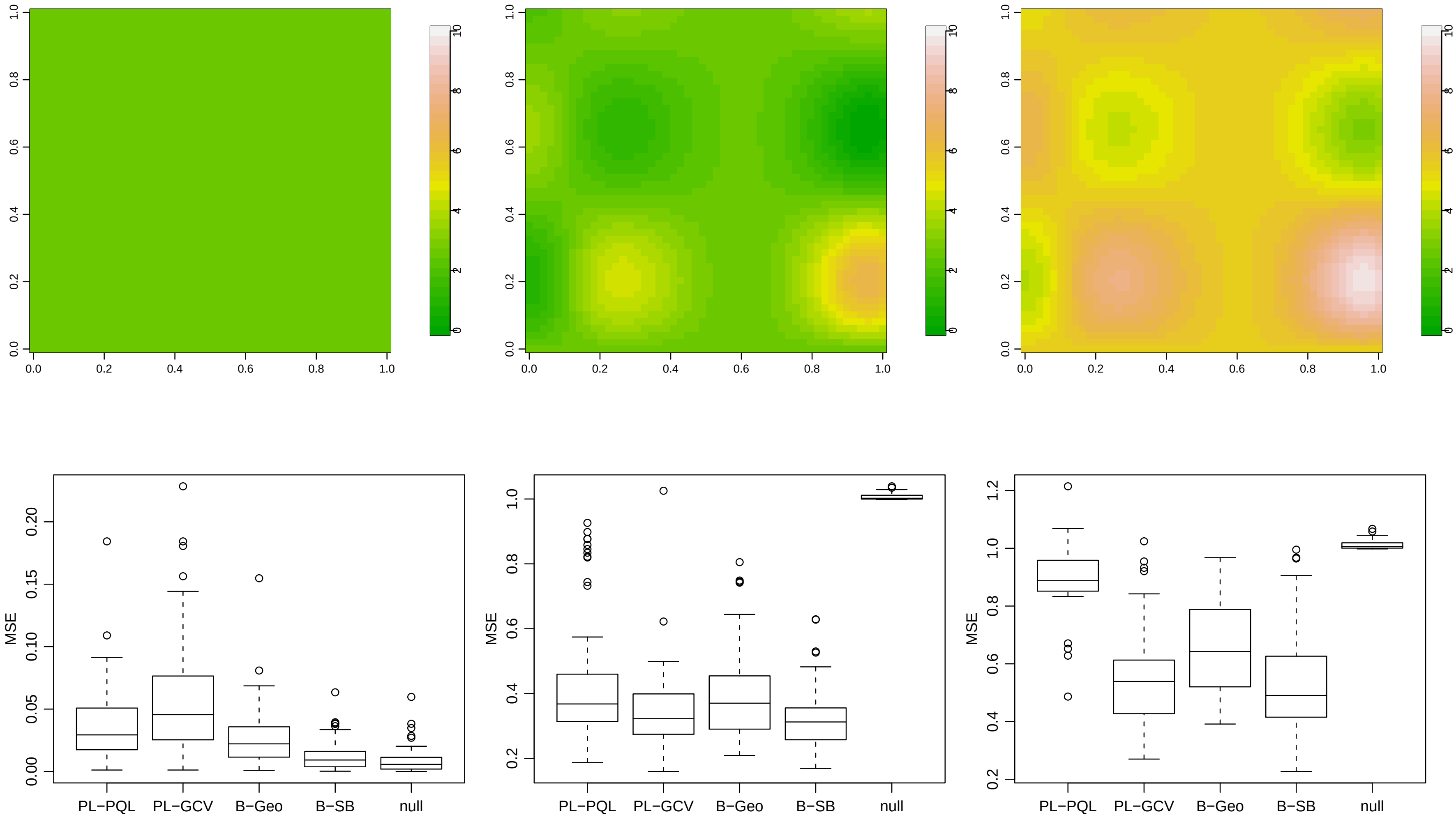
Summed test deviance
over 10-fold C-V sets

	leukemia	brain cancer
PL-GCV	590.1	529.8
PL-PQL	585.6	529.5
B-Geo	583.3	525.7
B-SB	582.1	525.1
null	581.6	525.5



Assessment on count simulations

$n = 225$, $n_{\text{test}} = 2500$ on 50 by 50 grid



Evaluation of methods

- Effective process parameterization = effective Bayesian estimation
  - feasible for spatial models with thousands of observations
- Natural Bayesian complexity penalty works well
  - GP representation zeroes out high-frequency coefficients as appropriate
- Implementation requires MCMC, not very accessible to practicioners
- Power is a real issue with spatial data in general, but particularly with binary observations
- Focused cluster-hunting or distance-based assessment of health risk may provide more power, but without full spatial assessment

Methodology challenges in spatial statistics related to public health

- design and power
  - how do we choose monitoring sites?
  - when we have enough power to estimate spatial features?
  - how do we model spatial processes when monitoring data is at lower resolution than the true surface?
- surveillance and hotspot detection
  - do Bayesian methods have a place in biosurveillance and cluster detection?
    - current applied work focuses on testing not modelling
  - surveillance likely to benefit from a decision theoretic approach that carefully considers both false positives and false negatives
- assigning one location to an individual is problematic
- variance partitioning between spatial terms and spatially-varying covariates
- confidentiality restrictions with respect to point locations and individual privacy

General challenges for spatial statistics in public health research

- computational: big datasets and fitting of complicated models
- collaborative: developing expertise among applied researchers
- leadership
  - statisticians should be at the forefront of analyzing geographically-indexed health data
  - we shouldn't leave this area to GIS analysts/geographers
  - necessity of providing and publicizing software for rigorous statistical methods
    - e.g., success of mixed model software – PROC MIXED, lme()
    - evidence of mgcv: public health researchers will learn R if useful model-building tools exist

- reproducibility: difficult to replicate analyses with complicated models, particularly MCMC implementations
  - posting code and releasing software with papers
  - standardized MCMC in R
    - ∗ many models, particularly new methods, can't be implemented in BUGS
      - · e.g., complicated spatio-temporal models
    - ∗ library of MCMC sampling functions with random variable classes
      - · Jouni Kerman (Columbia) has an initial implementation for Gibbs and Metropolis sampling (umacs)
      - · contributed sampling functions (e.g., slice sampling, Langevin sampling) would make this very powerful
    - ∗ reduce bugs, increase portability and reproducibility, optimize mixing