Spatial Statistics in a Bayesian Framework

Eugenio Paglino & Treva Tam
June 2022

Spatial Statistics

Spatial Thinking

Spatial data analysis focuses on modifications, extensions & additions to standard to standard statistical data analytical methods

These modifications consider explicitly the **spatial arrangement of the objects** being analyzed.

Spatially referenced data bring special problems to an analysis

The assumption of iid errors in a standard OLS regression specification is violated, and statistical inference is not valid.

Modifiable Areal Unit Problem (MAUP)

The aggregation of units (both the shape and scale) used in analysis are **arbitrary with respect to the phenomena of interest**, but also affect the analytical results.

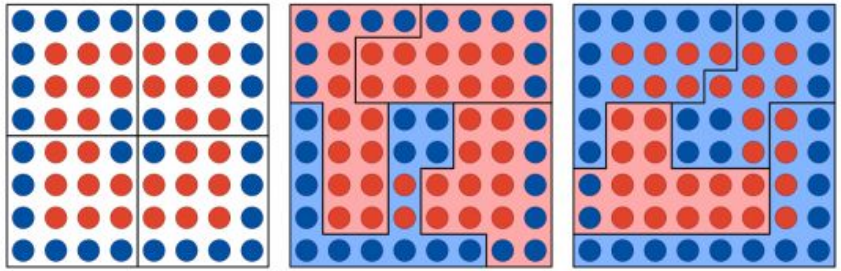
If the spatial units were specified differently, we might observe different patterns and relationships.

Essentially, a different data set is generated when the same data is gathered on different boundary definitions

MAUP

Scale: The larger the spatial unit the stronger the relationship

Shape: Units are demarcated artificially and can change/be redrawn



GISGeography. (2022). *MAUP-Modifiable Areal Problem*.
https://gisgeography.com/maup-modifiable-areal-unit-problem/

Openshaw and Taylor (1979): Using the same underlying data, showed it was possible to aggregate units to produce correlations anywhere between -1.0 to +1.0.

Spatial Autocorrelation

Describes the presence of systematic spatial variation in a variable

- *“everything is related to everything else, but near things are more related than distant things”* - Tobler’s Law of Geography

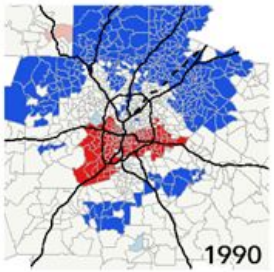
Residual dependence

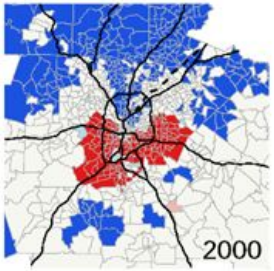
Seen as both a nuisance AND an important feature

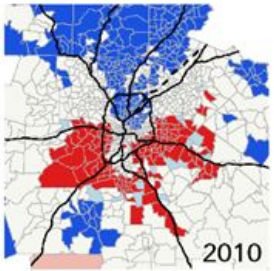
- 2 classes of tests:
  - Global Statistics (e.g. Moran’s I, Geary’s C, Geris-Ord G): one value for the entire dataset
  - Local Statistics (e.g. Local Moran’s I): each feature has a measure of autocorrelation

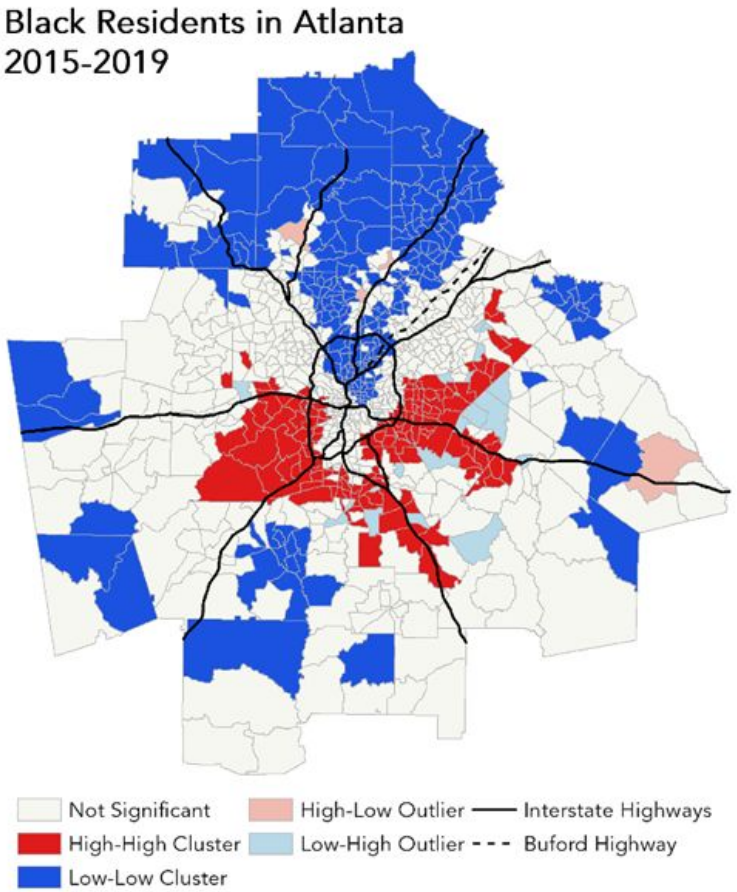
LISA statistic

- Anselin (1995)
- local indicator of spatial association
- Descriptively identifies “hot” and “cold” spots in your data





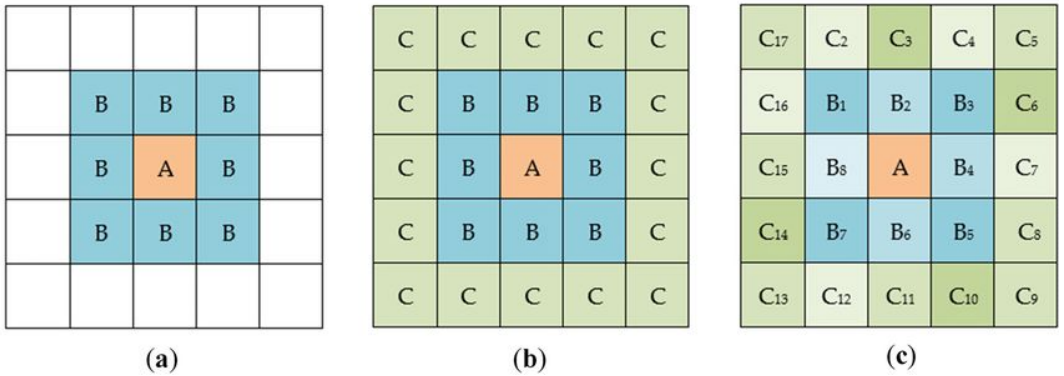




Spatial Weights

Defines the spatial relationship that exists among the features in your dataset

- Queen's Contiguity
- K nearest neighbors
- Fixed Distance
- Inverse Distance Weighting
- Spatial interaction



(a) queen contiguity weight **(b)** geographic distance weight **(c)** economic distance weight

Li, Kai & Ma, & Liu,. (2019). A New Trend in the Space–Time Distribution of Cultivated Land Occupation for Construction in China and the Impact of Population Urbanization. Sustainability. 11. 5089. 10.3390/su11185089.

Components of Spatial Data Analysis

Visualization → **showing** interesting patterns

Exploratory Data Analysis → **finding** interesting patterns

Spatial Modeling → **explaining** interesting patterns

- Spatial Heterogeneity
- Spatial Dependence

Bayesian Statistics and INLA

Frequentist versus Bayesian Approach

Both approaches are interested in modeling a variable **y** as a function of a set of parameters $\boldsymbol{\theta}$ through a generic model:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

The key difference is that frequentists consider the unknown parameters in $\boldsymbol{\theta}$ as fixed constants while Bayesians consider them as random variables with their own distributions.

The Bayesian Approach in one Slide

In its essence, the Bayesian approach consists of three steps:

1. Choose a model for $\mathbf{y}$ given $\boldsymbol{\theta}$
2. Choose a prior for $\boldsymbol{\theta}$
3. Derive the posterior of $\boldsymbol{\theta}$ given $\mathbf{y}$

While theoretically straightforward, the last step can be analytically impossible outside conjugate models (models in which the prior and and the posterior for $\boldsymbol{\theta}$ “turn out” to be of the same type).

This is why, despite its early origins, Bayesian statistics was not widely used until the 1990s.

Markov Chain Monte Carlo

The introduction of Markov Chain Monte Carlo (MCMC) methods allowed researchers to move beyond conjugated models by making it possible to sample from non-standard probability distributions.

While incredibly flexible, MCMC methods can quickly become computationally intensive when the number of parameters in the vector $\boldsymbol{\theta}$ or the size of the data $\mathbf{y}$ increase.

Furthermore, faster MCMC algorithms (such as Hamiltonian Monte Carlo) have led researchers to use more complicated methods.

Latent Gaussian Models

Consider a model where each $\mathbf{y}_i$ belong to the exponential family (Normal, Poisson, Beta, Gamma) and its mean $\boldsymbol{\mu}_i$ is modeled as a function of a set of linear predictors $\boldsymbol{\eta}_i$ ($g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i$ where:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i .$$

Latent Gaussian models are a subset of all models that can be represented in this way for which we are willing to assign a Gaussian prior to $\boldsymbol{\alpha}$, $f^{(j)}(.)$, $\boldsymbol{\beta}_k$, and $\boldsymbol{\varepsilon}_i$.

Latent Gaussian Models

While this can seem intimidating, there is a vast number of commonly used models that can be represented in this way:

1. Linear and generalized linear regressions models are obtained by setting all the $f^{(j)}(.)$ to 0.
2. Hierarchical linear models (random effects) are obtained by setting $f^{(j)}(u_i) = f_i$ and assume they are i.i.d. Normal.
3. Commonly used spatial models such as the Besag-York-Mollié model are obtained by setting $f^{(j)}(u_s) = f_s$ where *s* indicates a spatial location and the distribution of f_s smooths the area specific parameters across neighbors.

Integrated Nested Laplace Approximation

The parameters of our latent Gaussian model can be separated into two groups:

1. Non-Gaussian hyperparameters (variances, precisions) that regulate the priors of the other parameters (we group these in the vector $\boldsymbol{\psi}$).
2. Gaussian parameters ($\boldsymbol{\alpha}$, $f^{(j)}(.)$, $\boldsymbol{\beta}_k$, and $\boldsymbol{\varepsilon}_i$) which we group in the vector $\boldsymbol{\theta}$.

The distributions we are interested in are then $p(\boldsymbol{\Theta}_i \mid \mathbf{y})$, the marginal posterior distributions of the parameters of interest.

We can think of these as the distributions of the $\boldsymbol{\beta}$s in a linear regression together with the random intercepts and/or slopes in a hierarchical linear model.

Integrated Nested Laplace Approximation

The key insight here is that we can write:

$$p(\boldsymbol{\theta}_i \mid \mathbf{y}) = \int p(\boldsymbol{\theta}_i \mid \psi, \mathbf{y})\, p(\psi \mid \mathbf{y})\, d\psi$$

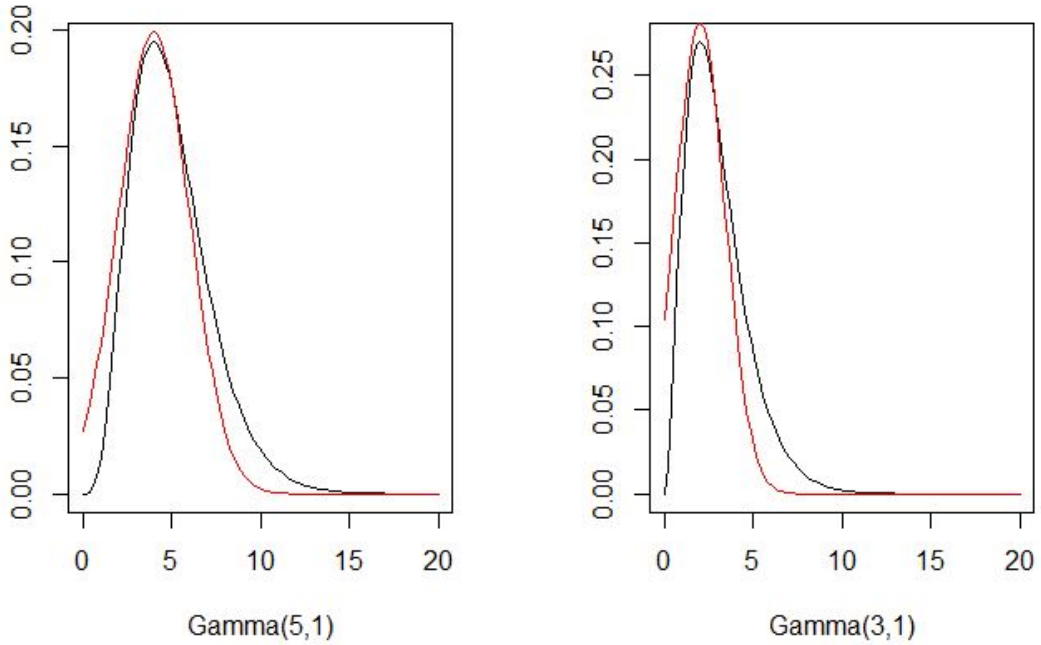
and then approximate $p(\boldsymbol{\theta}_i \mid \mathbf{y})$ by first approximating $p(\boldsymbol{\theta}_i \mid \psi, \mathbf{y})$ and $p(\psi \mid \mathbf{y})$ and then apply numerical integration (which is possible because the dimension of ψ is small).

The approximations to $p(\psi \mid \mathbf{y})$ and $p(\boldsymbol{\theta}_i \mid \psi, \mathbf{y})$ are based on the Laplace approximation and the simplified Laplace approximation respectively.

The Laplace Approximation

The Laplace approximation is based on using a Gaussian distribution with the same mode as the distribution *f(x)* being approximated (which thus needs to be unimodal) and a variance such that the approximated distribution's spread resembles that of the original distribution.

Laplace Approximation to Gamma distributions



Assumptions (not strict)

This approximation works well and is fast when:

1. The number of hyperparameters in ψ is small
2. The covariance matrix of the parameters in $\boldsymbol{\Theta}$ is sparse (dependence, if present, is only local).

Both assumptions are generally satisfied.

Advantages and Limitations

Advantages:

- ➔ INLA is extremely fast compared to a full Bayesian analysis but is not necessarily less accurate.
- ➔ It can be used to estimate a wide range of models

Limitations:

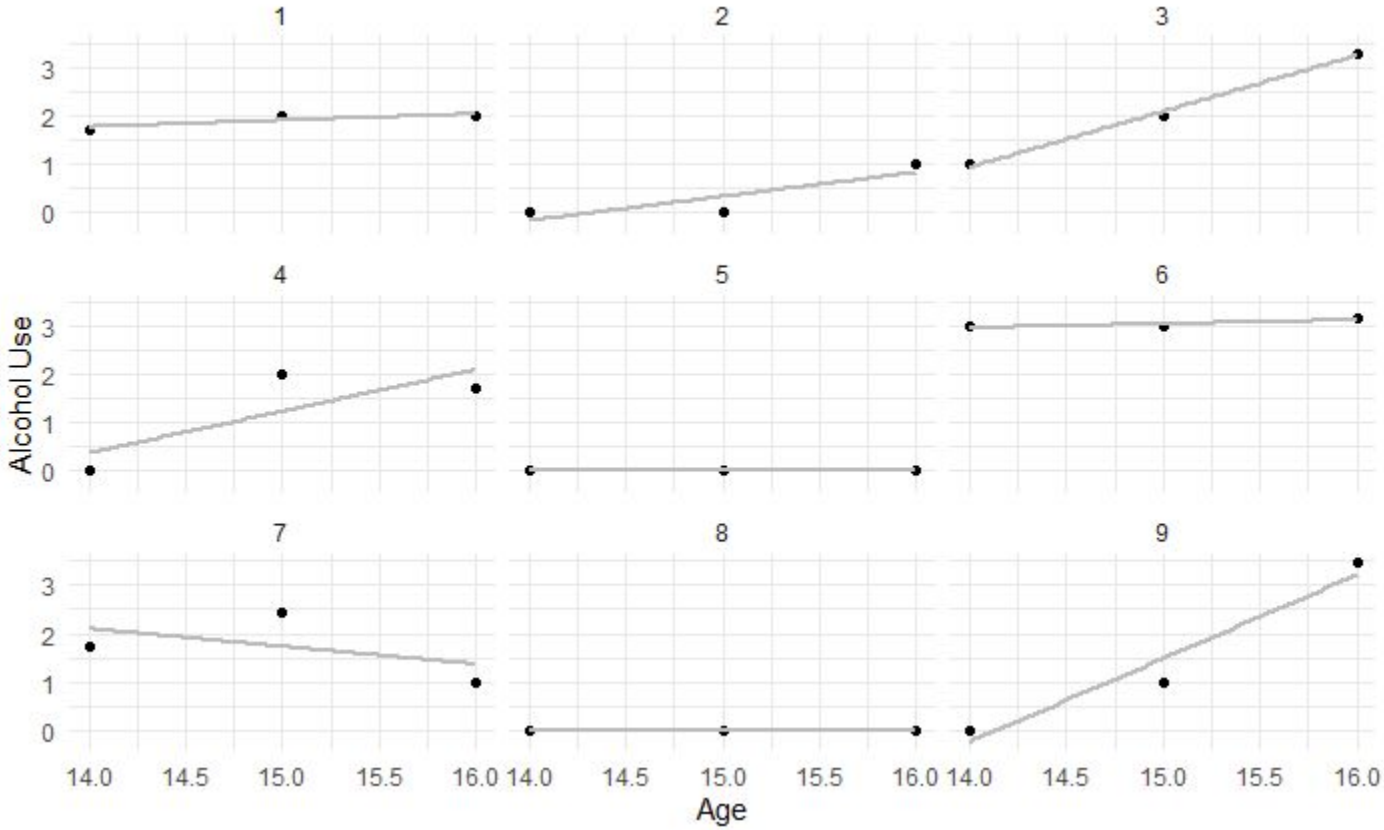
- ➔ Not all models can be formulated as latent Gaussian models, for example mixture models.
- ➔ The computational advantages are reduced when the parameters in $\boldsymbol{\theta}$ are dependent on each others (the precision matrix for the Gaussian prior on $\boldsymbol{\theta}$ needs to be sparse).

Applications

Data:

- ➔ Longitudinal observations of 82 young adults at ages 14,15, and 16.
- ➔ Alcohol use on a 0-7 scale was recorded as well as information of parents' alcohol use (binary) and proportion of peers consuming alcohol (0-5 scale).

Basic Visualization



Implementing the Model

We decide to estimate a model with uncorrelated random intercepts and slopes at the individual level.

`lme4` implementation:

```
model.lme <- alcData %>%
  lmer(alcuse ~ 1 + age_14 + cpeer + ccoa
               + (1 | id)
               + (0 + age_14 | id), data=.)
```

Implementing the Model

stanarm implementation:

```
model.stanarm <- alcData %>%
  stan_lmer(alcuse ~ 1 + age_14 + cpeer + ccoa
                     + (1 | id)
                     + (0 + age_14 | id),
            data=.)
```

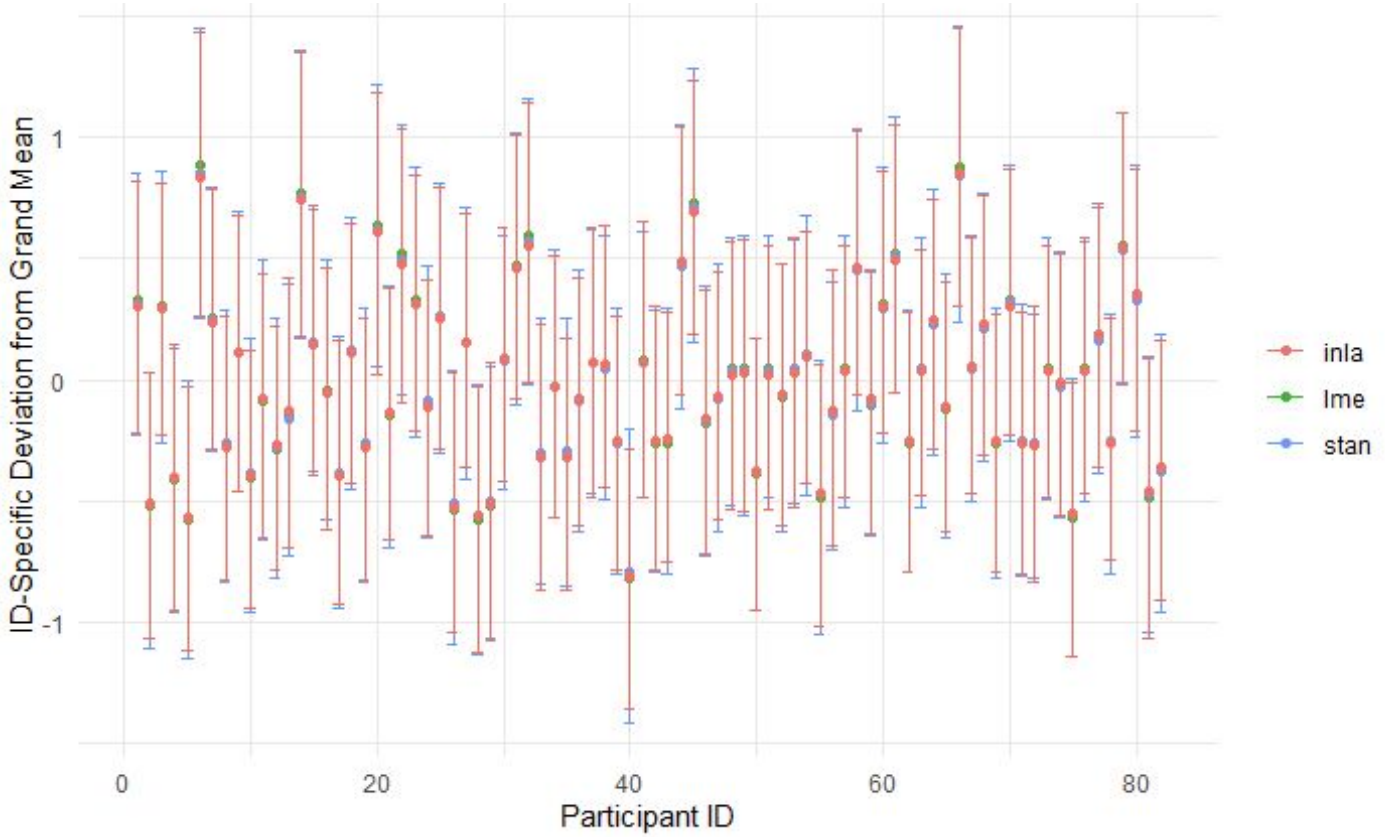
Notice how the stan team has mirrored the lme4 syntax so that no change is needed.

Implementing the Model

inla implementation:

```
model.inla <- alcData %>%
  mutate(id2 = id) %>% #We need to duplicate the id
  inla(alcuse ~ 1 + age_14 + cpeer + ccoa
                + f(id, model='iid')
                + f(id2, age_14, model='iid'),
    control.compute=list(config = TRUE),
    data=.)
```

Results


ID-Specific Deviation from Grand Mean
1
0
-1
0
20
40
60
80
Participant ID
inla
lme
stan

Execution Times

	`lme4`	`stanarm (1 core)`	`stanarm (4 cores)`	`inla`
Execution Time in secs	0.1047199	15.7162008	12.5480168	0.8972831

Bottom line:

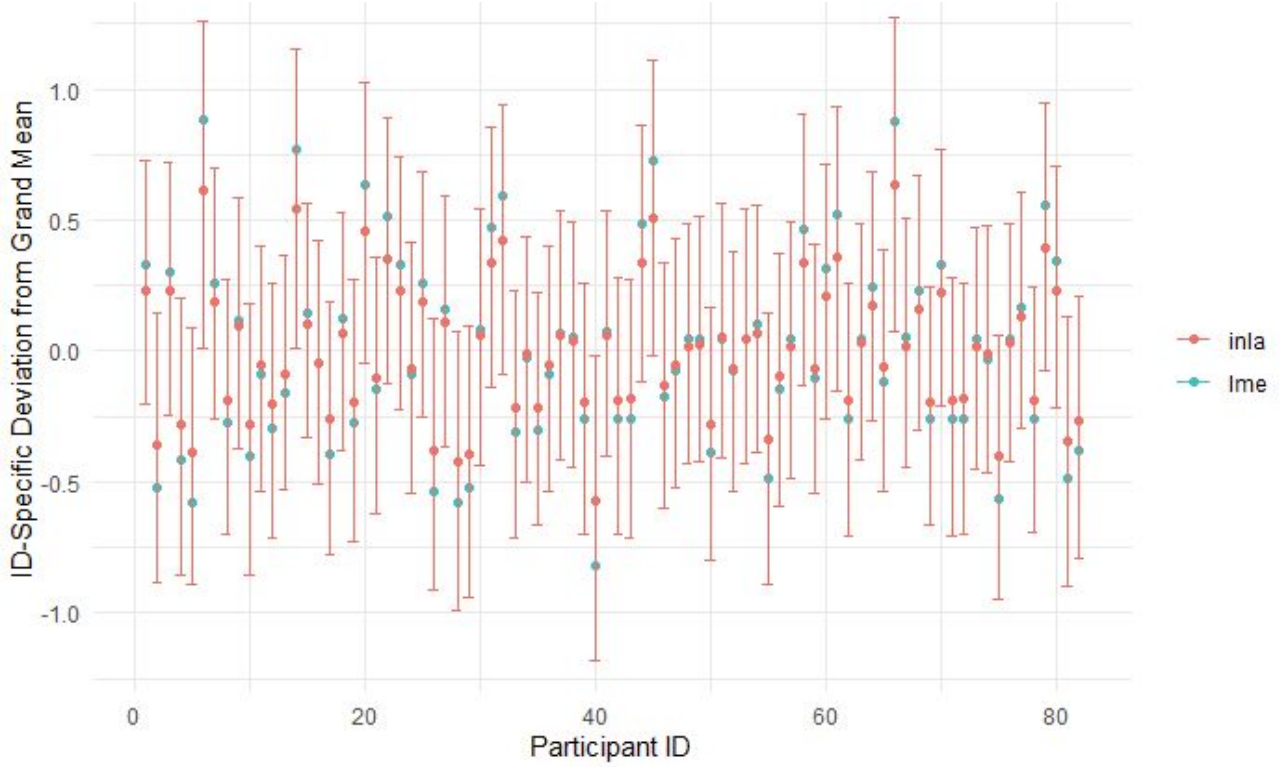
- ➔ lme4 is incredibly fast but has no reliable way of producing uncertainty intervals on random effects and is relatively rigid.
- ➔ Inla and stanarm have a similar amount of flexibility and precision but stanarm is at least one order of magnitude slower than inla (even for this relatively simple model).

What Do We Mean by “More Flexible”

In inla (but also in stan) we can play with priors to alter the estimated coefficients.

Here, for example, I’ve increase the amount of shrinkage performed by the model by increasing the precision of the prior.

This is not possible with lmer.



The Besag-York-Mollié Model

BYM model

Bayesian conditional autoregressive methods model the prior distribution in a way that accounts for spatial dependence in the posterior distribution (more in later slides)

A popular prior distribution for modeling spatial structure was introduced by Besag, York, and Mollié (BYM) where they decompose the random effect of the model into two parts:

- θ are the non-structured errors, iid
- Φ are the random effects with a spatially structured prior

$$\psi_i = \phi_i + \theta_i$$

$$\phi_i|\phi_j \sim N\left(\frac{\sum_{i=1}^{n} w_{ij}\phi_i}{\sum_{i=1}^{n} w_{ij}}, \frac{r^2}{\sum_{i=1}^{n} w_{ij}}\right)$$

$$\theta_i \sim N(0, \sigma^2)$$

$$\tau^2, \sigma^2 \sim \text{Inverse-Gamma}\ (a, b)$$

Example using BYM

Research Question: Where are White populations concentrated?

Data: tract-level 2010 Census data

- Dependent Variable: Number of White residents in 2010
- Determinents: % Black, % More than HS educated, % New Housing

$$\mu_{it} = \exp\left(X_i\beta + \psi_i\right)$$
$$\log\left(N_i(t)\right) = X_i\beta + \psi_i$$

Modeling the formula

```
f1 <- WHITE ~ B_PCT + MHS_PCT + NEWHOUSE_PCT
```

glm implementation:

```
mod2 <- glm(f1, family="poisson", data = data)
```

Carbayes implementation:

```
mod2 <- S.CARbym(f1, family="poisson", data = data,
                 W = W.mat, burnin = burn.in,
                 n.sample = N, thin = thin)
```

R-INLA implementation:

```
mod3 <- inla(update(f1, .~. , +
                    f(GEOID10, model = “bym”, graph = W.mat) +
                    f(GEOID10_2, model = “iid”)),
             family="poisson", data = data,
             control.predictor = list(compute = TRUE)
             )
```

Comparing Moran's I

Global Moran's

```
moran.mc(data$WHITE, nb.Q1, nsim = 999)
# 0.64867 ; p-value = 0.001
```

Aspatial Moran's

```
moran.mc(mod1$residuals, nb.Q1, nsim = 999)
# 0.47194 ; p-value = 0.001
```

BYM Moran's

```
resid.2 <- residuals(nh.CARflow.a1.z, type="response")
moran.mc(resid.2, nb.Q1, nsim=999)
# -0.082635 ; p-value = 0.999
```

Conclusion

- ➔ Spatial models are important to better understand phenomena with a geographic component.
- ➔ Ignoring spatial correlation might severely bias the results and lead to the wrong conclusions.
- ➔ Bayesian models have many advantages in this area and have been used successfully in many cases.
- ➔ INLA represents an exciting new method allowing for fast inference even with large datasets and more complex models.