From Points to Polygons: Modeling Spatial Autocorrelation

Muhammad Rehan David Goldstein
University of California, Los Angeles
rehanmuh@ucla.edu, dgoldstein@humnet.ucla.edu

April 7, 2026

Workshop on the Geography of Linguistic Evolution
EVOLANG XVI, Plovdiv, Bulgaria

Introduction

Space and typological variables

- Common theme across workshop's talks: spatial autocorrelation in linguistic traits

Space and typological variables

- Common theme across workshop's talks: spatial autocorrelation in linguistic traits
- How should we model spatial dependencies and on what representation of languages?

Space and typological variables

- Common theme across workshop's talks: spatial autocorrelation in linguistic traits
- How should we model spatial dependencies and on what representation of languages?
  - Traditional response: stratified sampling or areal controls (macroareas) (e.g., Bickel 2013; Perkins 1989)

Space and typological variables

- Common theme across workshop's talks: spatial autocorrelation in linguistic traits
- How should we model spatial dependencies and on what representation of languages?
  - Traditional response: stratified sampling or areal controls (macroareas) (e.g., Bickel 2013; Perkins 1989)
  - Recent advances: model space directly as a latent random effect over language coordinates (Guzmán Naranjo, Becker, et al. 2025; Hartmann and Nichols 2025; Verkerk et al. 2025)

This talk & case study: population size and phoneme inventory size

1. Argue that Conditional Autoregressive (CAR) priors might be better suited for the modeling of spatial autocorrelation in some linguistic data than Gaussian processes (GPs)

This talk & case study: population size and phoneme inventory size

1. Argue that Conditional Autoregressive (CAR) priors might be better suited for the modeling of spatial autocorrelation in some linguistic data than Gaussian processes (GPs)
2. Examine the alleged effect of population size on phoneme inventory size

This talk & case study: population size and phoneme inventory size

1. Argue that Conditional Autoregressive (CAR) priors might be better suited for the modeling of spatial autocorrelation in some linguistic data than Gaussian processes (GPs)
2. Examine the alleged effect of population size on phoneme inventory size

- Research Question: do larger speech communities have larger phoneme inventories?

This talk & case study: population size and phoneme inventory size

1. Argue that Conditional Autoregressive (CAR) priors might be better suited for the modeling of spatial autocorrelation in some linguistic data than Gaussian processes (GPs)
2. Examine the alleged effect of population size on phoneme inventory size

- Research Question: do larger speech communities have larger phoneme inventories?
- Some claims of a positive effect (e.g. Hay and Bauer 2007; Atkinson 2011) vs. failed replications/criticisms (e.g. Donohue and Nichols 2011; Moran, McCloy, and Wright 2012)

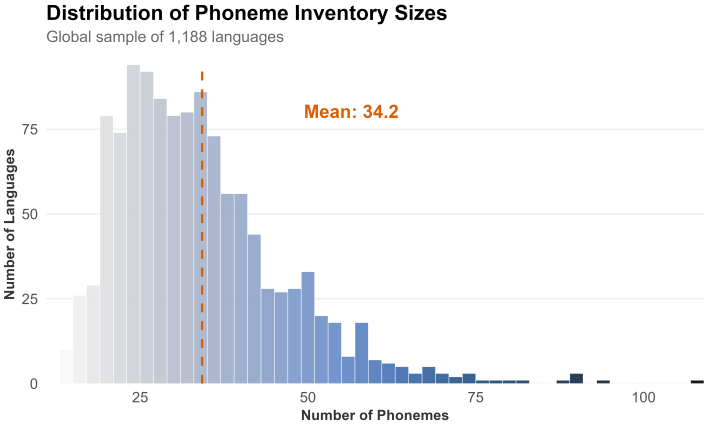
This talk & case study: population size and phoneme inventory size

1. Argue that Conditional Autoregressive (CAR) priors might be better suited for the modeling of spatial autocorrelation in some linguistic data than Gaussian processes (GPs)
2. Examine the alleged effect of population size on phoneme inventory size

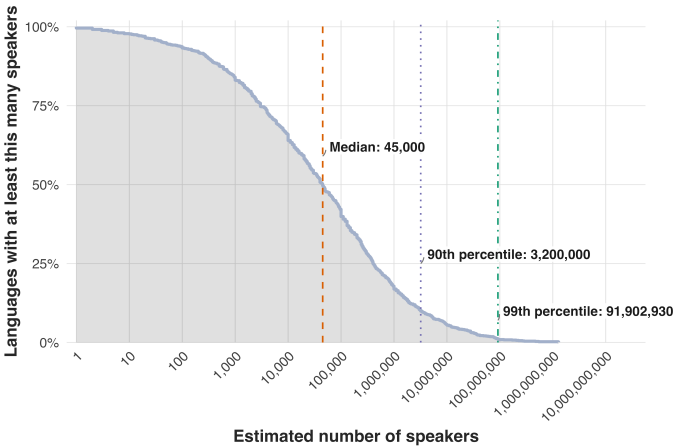
- Research Question: do larger speech communities have larger phoneme inventories?
- Some claims of a positive effect (e.g. Hay and Bauer 2007; Atkinson 2011) vs. failed replications/criticisms (e.g. Donohue and Nichols 2011; Moran, McCloy, and Wright 2012)
- Test the hypothesis more adequately by controlling for phylogenetic and spatial autocorrelation

Data

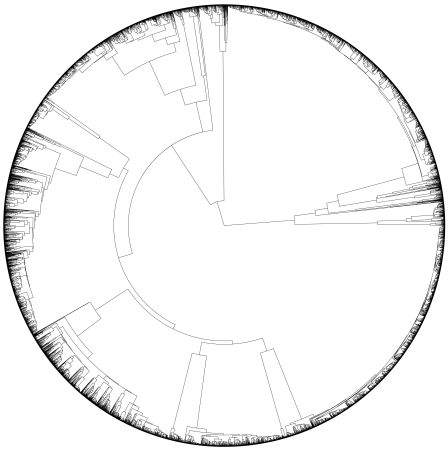
Phoneme inventory data (N = 1,188; Moran and McCloy 2019)



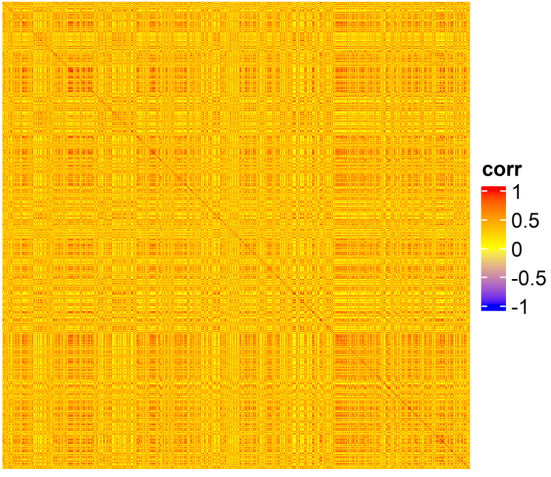
Estimates of population sizes (Ritchie et al. 2024)



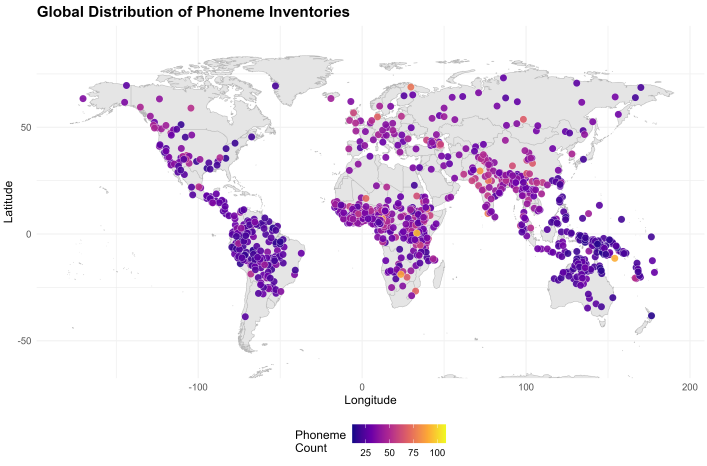
Global phylogeny (Bouckaert et al. 2022)



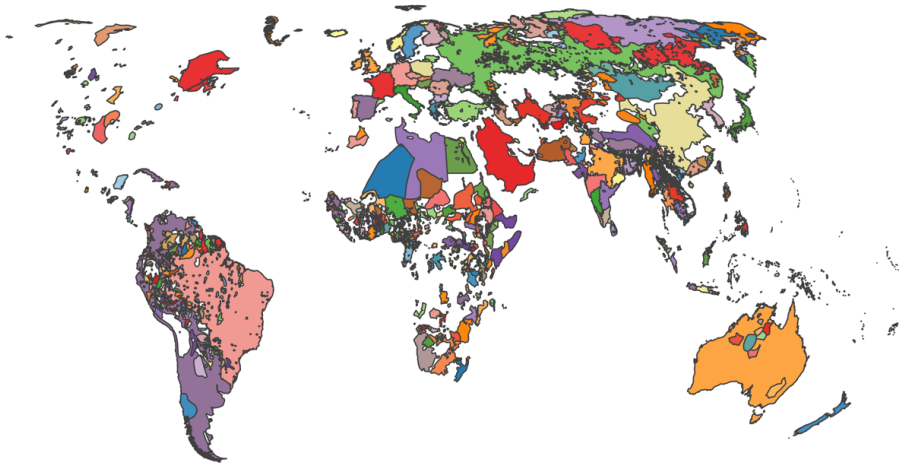
Phylogenetic correlation matrix (Revell and Harmon 2022)



Point coordinates (Hammarström et al. 2024)



Language polygons (Ranacher et al. 2025)



Methods

Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)

Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)
- **Likelihood (overdispersed count):**

$$y_i \mid \mu_i, \theta \sim \text{NegBin}(\mu_i, \theta), \qquad \theta > 0$$

Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)
- **Likelihood (overdispersed count):**

$$y_i \mid \mu_i, \theta \sim \text{NegBin}(\mu_i, \theta), \qquad \theta > 0$$

- **Mean structure:** the expected inventory size is modeled on the log scale as a function of:

Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)
- **Likelihood (overdispersed count):**

$$y_i \mid \mu_i, \theta \sim \text{NegBin}(\mu_i, \theta), \qquad \theta > 0$$

- **Mean structure:** the expected inventory size is modeled on the log scale as a function of:
  - standardized log population size (fixed effect; $\beta_1 x_i$),

Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)
- **Likelihood (overdispersed count):**

$$y_i \mid \mu_i, \theta \sim \text{NegBin}(\mu_i, \theta), \qquad \theta > 0$$

- **Mean structure:** the expected inventory size is modeled on the log scale as a function of:
  - standardized log population size (fixed effect; $\beta_1 x_i$),
  - a **dataset** random intercept (absorbs systematic differences; $b_{d[i]}$),

Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)
- **Likelihood (overdispersed count):**

$$y_i \mid \mu_i, \theta \sim \text{NegBin}(\mu_i, \theta), \qquad \theta > 0$$

- **Mean structure:** the expected inventory size is modeled on the log scale as a function of:
  - standardized log population size (fixed effect; $\beta_1 x_i$),
  - a **dataset** random intercept (absorbs systematic differences; $b_{d[i]}$),
  - a **phylogenetic** random intercept (shared ancestry; $\gamma_{p[i]}$, with Brownian/tree covariance)

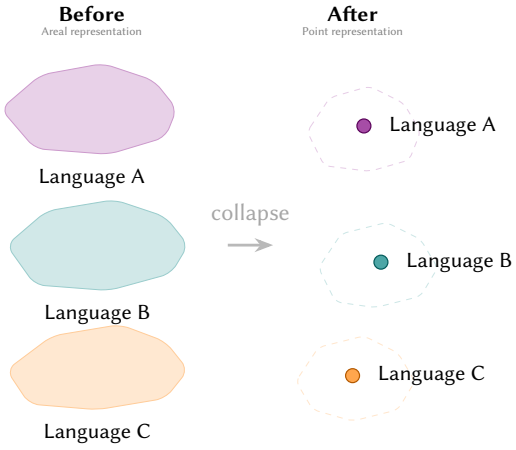
Bayesian GLMM

- **Outcome:** phoneme inventory size for each language (a count)
- **Likelihood (overdispersed count):**

$$y_i \mid \mu_i, \theta \sim \text{NegBin}(\mu_i, \theta), \qquad \theta > 0$$

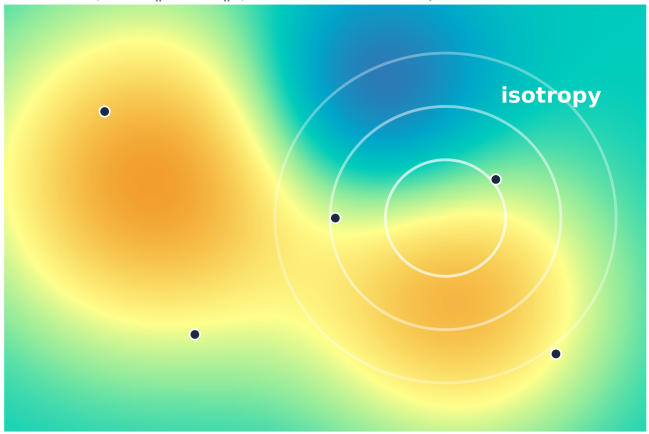
- **Mean structure:** the expected inventory size is modeled on the log scale as a function of:
  - standardized log population size (fixed effect; $\beta_1 x_i$),
  - a **dataset** random intercept (absorbs systematic differences; $b_{d[i]}$),
  - a **phylogenetic** random intercept (shared ancestry; $\gamma_{p[i]}$, with Brownian/tree covariance)
- Inference via INLA for scalability (Rue et al. 2024; R-INLA).

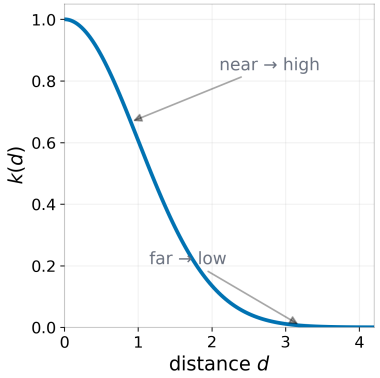
Languages encoded as points



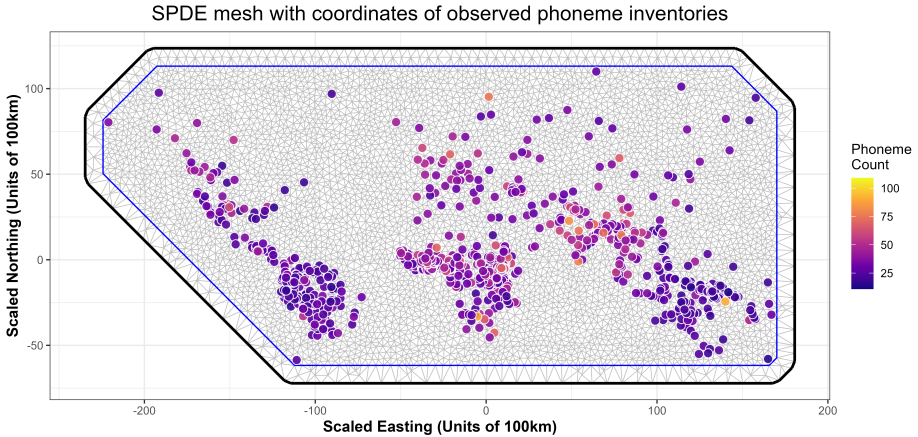
Gaussian Process on a spatial random field

$f(\mathbf{s}) \sim \mathcal{GP}(0,\ k(\|\mathbf{s}-\mathbf{s}'\|))$ (smooth, isotropic)

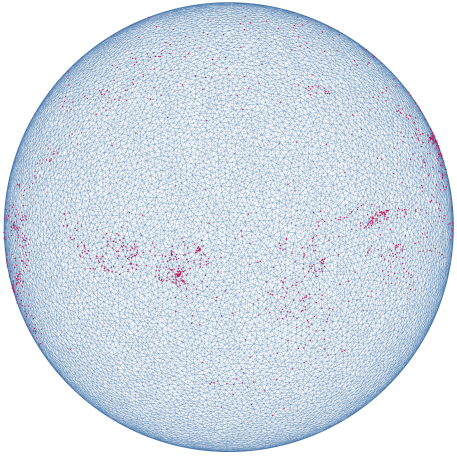




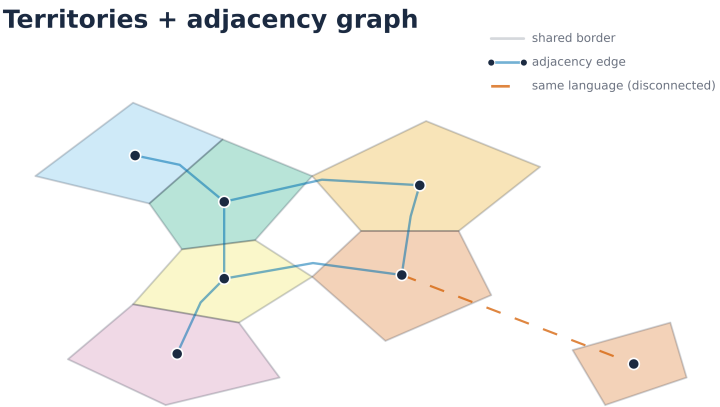
2D INLA SPDE mesh for the GP field



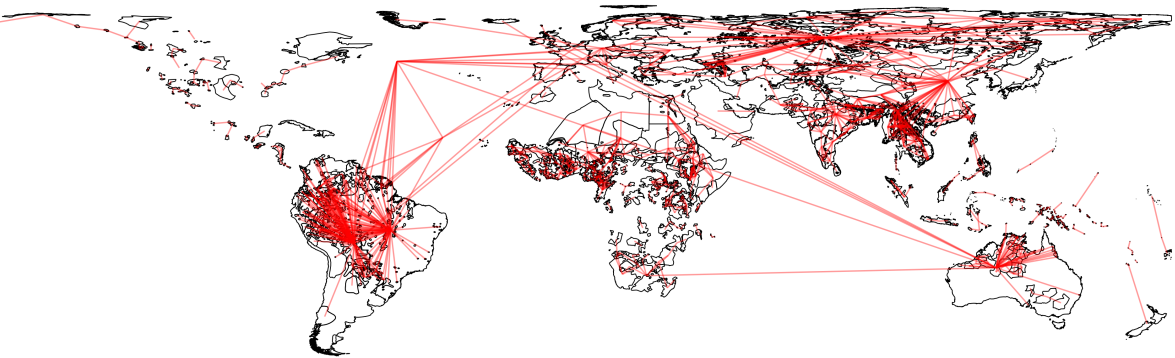
3D INLA SPDE mesh for the GP field



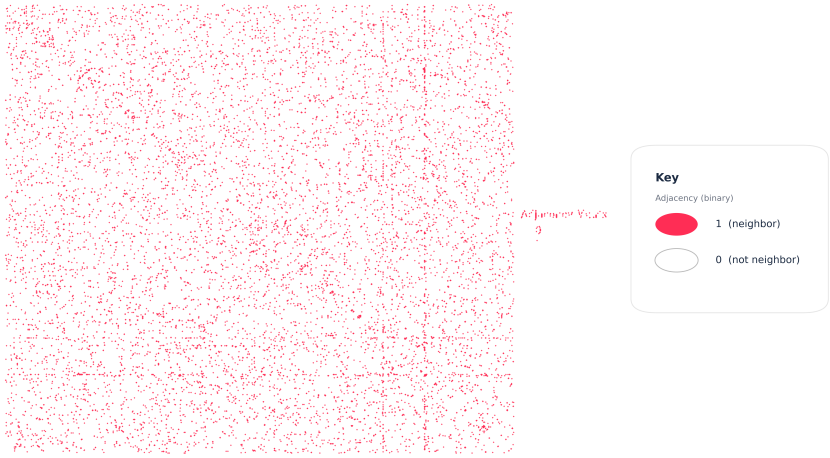
Adjacency encoded over shared boundaries



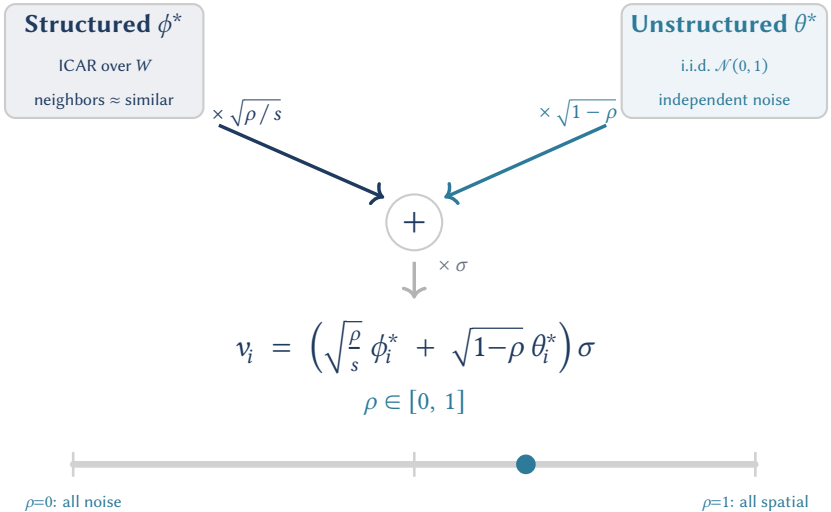
Neighborhood graph for the polygon data



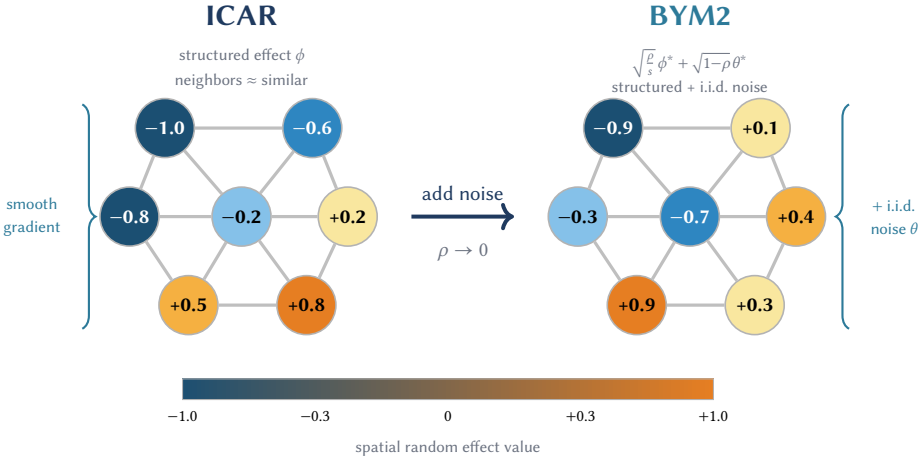
Weight matrix



BYM2: structured + unstructured areal effects



ICAR / BYM2 lattice prior



Model comparison

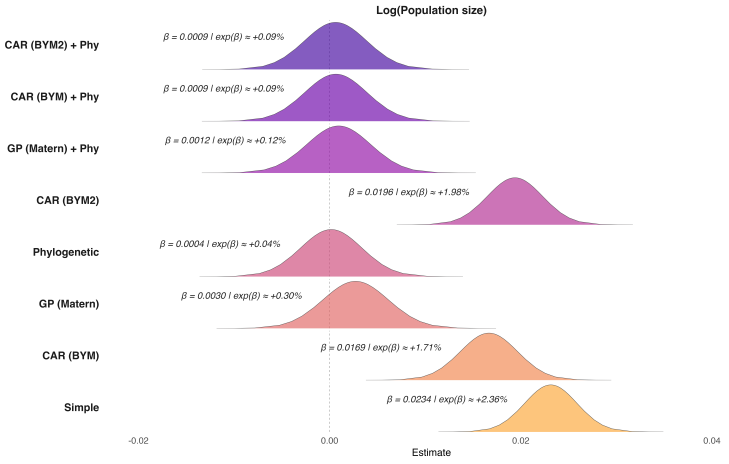
- Baseline NB regression (no phylogeny, no space)
- + **Phylogeny only** (Brownian covariance on the tree)
- + **Space only**:
  - Matérn GP (points)
  - ICAR / BYM2 (polygons)
- + **Phylogeny** + **space** (all combinations)

Model comparison

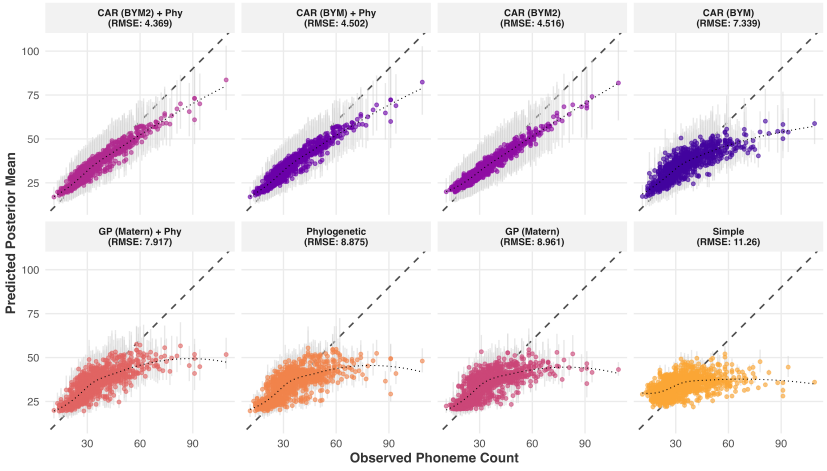
- Baseline NB regression (no phylogeny, no space)
- + **Phylogeny only** (Brownian covariance on the tree)
- + **Space only**:
  - Matérn GP (points)
  - ICAR / BYM2 (polygons)
- + **Phylogeny** + **space** (all combinations)
- Compare with INLA diagnostics: DIC, and WAIC

Results

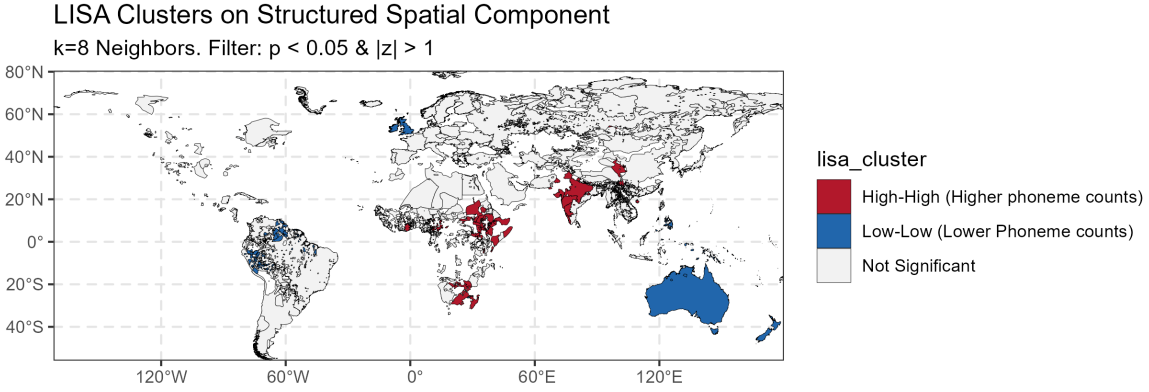
Estimates of the fixed effect



Posterior predictive check



BYM2+Phy inferred spatial effect



Discussion

Population size and phoneme inventory size

- Once phylogeny and space are modeled, the effect of population size on inventory size is very small / near zero (Donohue and Nichols 2011; Jäger 2025; Moran, McCloy, and Wright 2012)

Population size and phoneme inventory size

- Once phylogeny and space are modeled, the effect of population size on inventory size is very small / near zero (Donohue and Nichols 2011; Jäger 2025; Moran, McCloy, and Wright 2012)
- For the inference of typological variables, space and phylogeny must be controlled

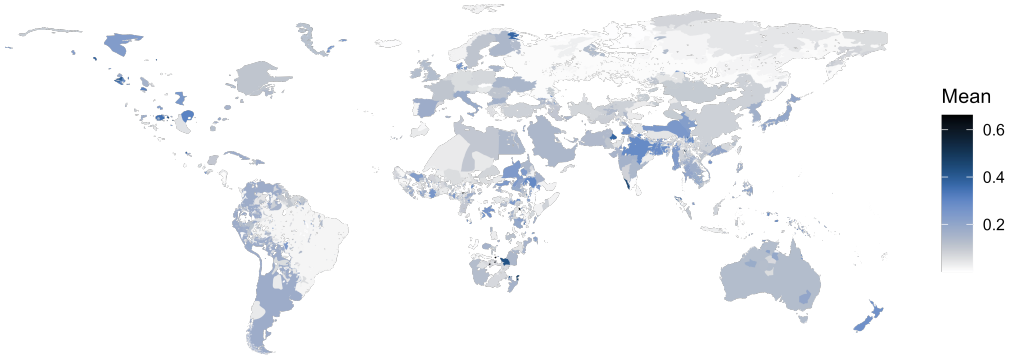
Spatial autocorrelation in phoneme inventory sizes

- CAR models allow us to move away from a binary classification of discrete linguistic areas and quantify the gradience in spatial autocorrelation (cf. Guzmán Naranjo, Mertner, et al. 2024, p. 5)

Spatial autocorrelation in phoneme inventory sizes

- CAR models allow us to move away from a binary classification of discrete linguistic areas and quantify the gradience in spatial autocorrelation (cf. Guzmán Naranjo, Mertner, et al. 2024, p. 5)
- Compare and rank linguistic areas according to strength of spatial autocorrelation

BYM2+Phy inferred structured spatial effect



Mapping the South Asian Sprachbund

I The core bloc — Central / Peninsular India · Deccan · Indo-Gangetic

Retroflexion (Emeneau 1956) — A "solid bloc" across Indo-Aryan, Dravidian, and Munda.

Mapping the South Asian Sprachbund

I **The core bloc** — Central / Peninsular India · Deccan · Indo-Gangetic

Retroflexion (Emeneau 1956) — A "solid bloc" across Indo-Aryan, Dravidian, and Munda.

II **The NW wedge** — Eastern Pakistan · Indus border · Transition zone

The B-complex (Tikkanen 2005) — Triple sibilant contrast (*ś* : *ṣ* : *s*) and retroflex affricates.

Mapping the South Asian Sprachbund

I The core bloc — Central / Peninsular India · Deccan · Indo-Gangetic

Retroflexion (Emeneau 1956) — A "solid bloc" across Indo-Aryan, Dravidian, and Munda.

II The NW wedge — Eastern Pakistan · Indus border · Transition zone

The B-complex (Tikkanen 2005) — Triple sibilant contrast (*ś* : *ṣ* : *s*) and retroflex affricates.

III The NE polygon — Himalayan / Tibetan plateau · Sino-Tibetan contact

Areal attenuation — "Missing" retroflexes in the NE cluster due to Tibeto-Burman/Daic influence.

Conclusion

Conclusions & takeaways

- **Substantive:** Population size → phoneme inventory size near zero once **phylogeny** + **space** modeled.

Conclusions & takeaways

- **Substantive:** Population size → phoneme inventory size near zero once **phylogeny** + **space** modeled.
- **Model comparison:** In our case study, **polygon-based CAR/BYM2 improves fit** relative to point-based GP; GPs remain useful as alternative diffusion priors

Conclusions & takeaways

- **Substantive:** Population size → phoneme inventory size near zero once **phylogeny** + **space** modeled.
- **Model comparison:** In our case study, **polygon-based CAR/BYM2 improves fit** relative to point-based GP; GPs remain useful as alternative diffusion priors
- **Recommendation:** Compare GP vs. CAR models in the same analysis

Conclusions & takeaways

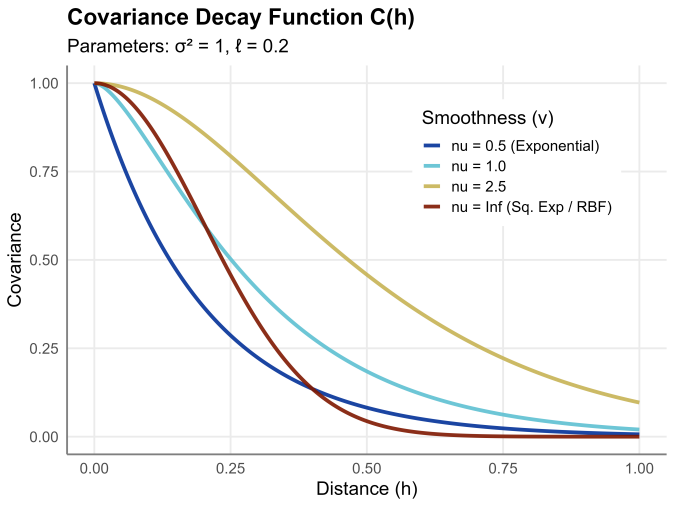
- **Substantive:** Population size → phoneme inventory size near zero once **phylogeny** + **space** modeled.
- **Model comparison:** In our case study, **polygon-based CAR/BYM2 improves fit** relative to point-based GP; GPs remain useful as alternative diffusion priors
- **Recommendation:** Compare GP vs. CAR models in the same analysis
- **Broader takeaway:** Ranacher et al. (2025)'s database of polygons enables worldwide inference of spatial autocorrelation on an areal representation of languages.

Благодаря ви за вниманието![1]

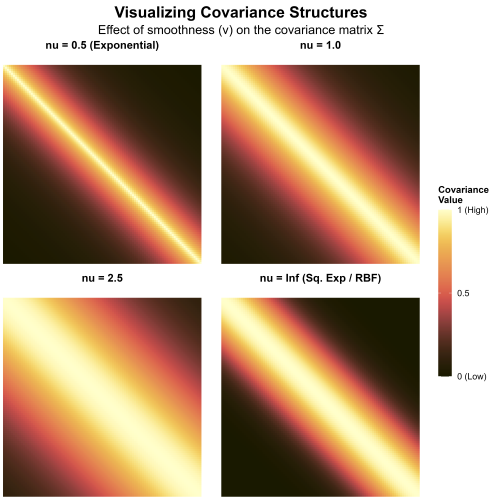
[1]Thank you to the members of the PIES Graduate Seminar and UCLA Phonology Seminar for their feedback and support.

Appendix

Covariance decay



Covariance matrices



Negative binomial baseline (INLA)

- Response y_i = phoneme inventory size (count), modeled as NB:

$$y_i \sim \text{NegBin}(\mu_i, \kappa)$$

- Linear predictor (baseline):

$$\log \mu_i = \beta_0 + \beta_1 \log(\text{speakers}_i) + b_{\text{dataset}[i]}.$$

- Dataset random intercepts: $b_d \sim \mathcal{N}(0, \sigma^2_{\text{dataset}})$.

Phylogenetic random effect (Brownian covariance)

- Let C be the phylogenetic correlation matrix induced by branch lengths under Brownian motion on a tree.
- Add a phylogenetic random effect:

$$u \sim \mathcal{N}(0, \sigma^2_{\text{phy}} C).$$

- This captures family-level dependence without forcing arbitrary family fixed effects.

Two spatial priors over “areal effect”

Continuous space (points)

- Spatial Gaussian process over coordinates
- Matérn covariance kernel implemented via SPDE in INLA
- Assumes smooth decay of covariance with distance

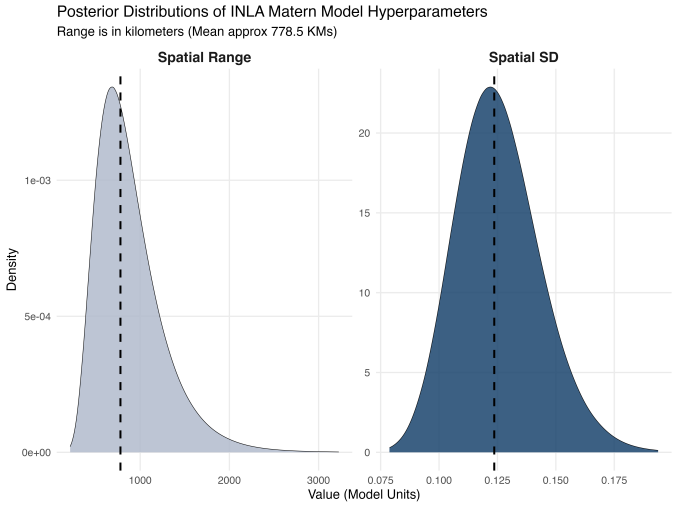
Discrete space (polygons)

- Conditional autoregressive (CAR) prior over areas
- Binary neighbor encoding in a graph representing shared polygon borders
- Captures patchiness / discontinuities naturally

Matérn GP (points)

- Matérn covariance controls:
  - range (how far correlation persists),
  - marginal variance,
  - smoothness ν (often fixed / weakly informed).
- In INLA: SPDE representation yields sparse precision and scalable inference.

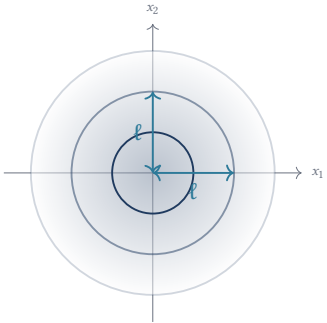
Matérn GP inferred hyperparameters



Isotropic vs. Anisotropic Kernels

Isotropic Prior

Circular Symmetry

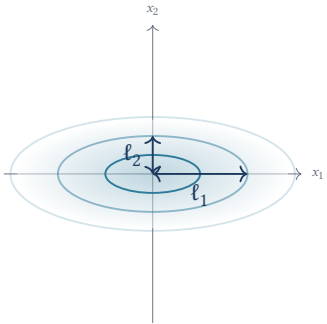


One lengthscale fits all.

$$\ell_{x_1} = \ell_{x_2}$$

Anisotropic (ARD) Prior

Axis-aligned Scaling



Independent scaling.

$$\ell_{x_1} \neq \ell_{x_2}$$

INLA mesh + SPDE ⇒ Gaussian spatial field

Continuous field

Let $x(s)$ be a mean-zero Gaussian spatial field on $D \subset \mathbb{R}^2$, often specified as a Matérn GRF:

$$x(\cdot) \sim \mathcal{GP}\big(0,\ C_{\text{Matérn}}(\cdot,\cdot;\nu,\kappa,\sigma^2)\big).$$

SPDE representation (equivalent characterization)

A Matérn GRF can be characterized as the solution to the SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}\, x(s) \;=\; \tau W(s), \qquad \alpha = \nu + \tfrac{d}{2}\ \ (d = 2),$$

where $W(s)$ is spatial Gaussian white noise.

Mesh approximations

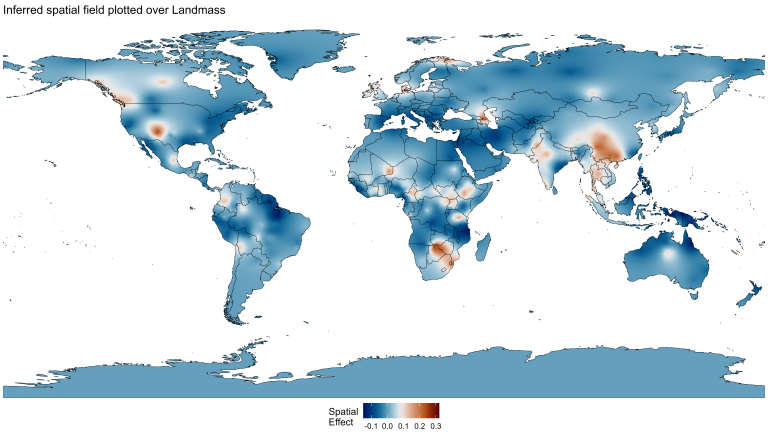
Mesh-based GMRF approximation (INLA)

Triangulate D with nodes $\{s_k\}_{k=1}^m$ and basis functions $\{\psi_k(s)\}$ (piecewise linear):

$$x(s) \approx \sum_{k=1}^{m} w_k \psi_k(s), \qquad \mathbf{w} = (w_1, \dots, w_m)^\top \sim \mathcal{N}\big(\mathbf{0},\ \mathbf{Q}(\kappa, \tau)^{-1}\big),$$

with sparse precision matrix $\mathbf{Q}$ induced by the FEM discretization of the SPDE.

Spatial field under Matérn+Phy



ICAR prior: neighborhood smoothing

- Define adjacency W from polygon borders; $w_{ij} = 1$ if areas share a boundary.
- Let $n_i = \sum_j w_{ij}$. ICAR is specified by conditionals:

$$\phi_i \mid \phi_{-i} \sim \mathcal{N}\Big(\frac{1}{n_i}\sum_{j\sim i}\phi_j,\ \frac{1}{\tau n_i}\Big).$$

- Equivalent joint form uses a sparse precision matrix $Q = \tau(D - W)$ (intrinsic / improper without constraints).

BYM2 (scaled) recap

$$v_i = \left(\sqrt{\frac{\rho}{s}}\,\phi_i^* + \sqrt{1-\rho}\,\theta_i^*\right)\sigma, \quad \rho \in [0,1].$$

- ϕ^*: scaled ICAR component (structured).
- θ^*: i.i.d. component (unstructured).
- ρ: interpretable mixing (how much structure).
- σ: overall marginal SD.

Extensions: add phylogenetic and/or spatial latent effects

- Phylogenetic random intercept (Brownian/tree-derived precision):

$$\eta_i \;=\; \beta_0 + \beta_1\,\texttt{log(pop)}_i + b_{d[i]} + \gamma_{p[i]}.$$

- Areal (lattice) spatial effects via polygon adjacency graph:

$$\eta_i \;=\; \beta_0 + \beta_1\,\texttt{log(pop)}_i + b_{d[i]} + u_{\texttt{idx}[i]} \quad \text{(proper ICAR; Besag 1974),}$$

$$\eta_i \;=\; \beta_0 + \beta_1\,\texttt{log(pop)}_i + b_{d[i]} + v_{\texttt{idx}[i]} \quad \text{(BYM2; Riebler et al. 2016).}$$

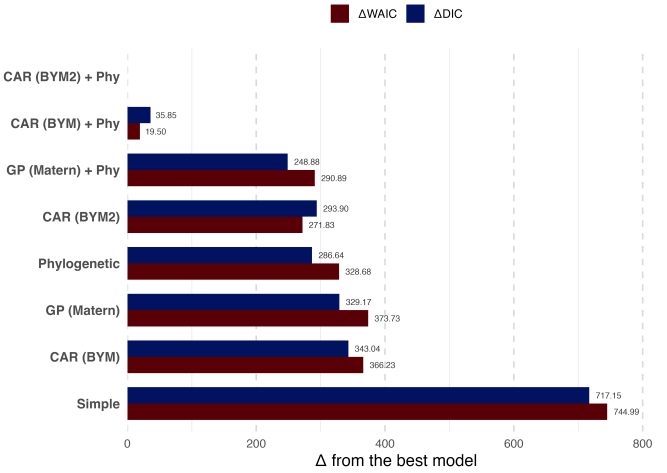
- Continuous spatial field (Matérn SPDE on mesh):

$$\eta_i \;=\; \beta_0 + \beta_1\,\texttt{log(pop)}_i + b_{d[i]} + w(s_i).$$

- Combined (examples):

$$\eta_i = \beta_0 + \beta_1\,\texttt{log(pop)}_i + b_{d[i]} + v_{\texttt{idx}[i]} + \gamma_{p[i]},$$

Model comparison results



Lower delta values are better.

Prior sketch

- Fixed effects: weakly informative Gaussian priors.
- Random-effect SDs: penalized complexity (PC) style priors (default in INLA / common in spatial work).
- BYM2: PC priors for (ρ, σ)
- Sensitivity: vary hyperpriors to check stability of fixed-effect conclusions.

Observed vs. predicted scatter