

Day 2 – Joint modelling and Non-stationary processes

Olatunji Johnson

Day 2 goals

Today we focus on two ideas that come up constantly in disease mapping:

1. Joint modelling

- ▶ multiple related malaria outcomes
- ▶ shared and outcome-specific spatial structure
- ▶ *multiple likelihoods* (Binomial + Poisson)

2. Non-stationary spatial processes

- ▶ when “one Matérn everywhere” is not realistic
- ▶ two practical constructions:
 - ▶ mixture of SPDEs (smooth/rough)
 - ▶ **structured range** (range changes with a covariate)

Why do we need joint models?

You often have **multiple imperfect views** of the same latent risk:

- ▶ prevalence surveys (direct infection measurements)
- ▶ facility case counts (incidence proxy)

Part A — Joint modelling (same likelihood, two groups)

Example: two prevalence processes

Suppose we have:

- ▶ children <5 prevalence data
- ▶ pregnant women prevalence data

At each location s_i , group $g \in \{c, p\}$:

$$Y_{ig} \mid p_{ig} \sim \text{Binomial}(N_{ig}, p_{ig})$$

$$\text{logit}(p_{ig}) = \eta_{ig}$$

Decomposition: shared + group-specific spatial fields

A common and interpretable joint structure:

Identifiability and scaling

Potential issue: S_0 and S_g can compete.

Typical solutions:

- ▶ set priors so that group-specific fields are “smaller” than the shared field
- ▶ consider constraints or priors on variance:
 - ▶ $\sigma_{S_g}^2 < \sigma_{S_0}^2$ (probabilistically)

Conceptually: shared structure should capture the dominant signal.

What do we assume about $S_0(s)$ and $S_g(s)$?

A standard choice: **Matérn Gaussian random fields**

$$S(s) \sim \text{GRF}(0, \text{Matérn}(\rho, \sigma^2))$$

- ▶ ρ : range (how quickly correlation decays)
- ▶ σ^2 : marginal variance

Stationarity here means ρ, σ^2 are constant over space.

Mesh and SPDE

- ▶ Mesh nodes: basis functions
- ▶ Field represented as:

$$S(s) \approx \sum_{k=1}^K w_k \phi_k(s)$$

where w_k are GMRF weights.

Joint model uses the same mesh for S_0 and S_g , but:

- ▶ S_0 : one field
- ▶ S_g : replicated by group

Part B — Joint modelling with multiple likelihoods

Motivation: prevalence + case counts

Often you have:

- ▶ **Binomial** prevalence surveys:
 - ▶ Y_i positives out of N_i tested
- ▶ **Poisson** facility case counts:
 - ▶ C_i cases with exposure E_i (population, person-time, etc.)

Each dataset is incomplete on its own.

Joint modelling combines them while respecting their data-generating mechanisms.

The two likelihoods

Prevalence

$$Y_i \mid p_i \sim \text{Binomial}(N_i, p_i), \quad \text{logit}(p_i) = \eta_i^{(B)}$$

Counts

$$C_i \mid \lambda_i \sim \text{Poisson}(E_i \lambda_i), \quad \log(\lambda_i) = \eta_i^{(P)}$$

Different link functions:

- ▶ logit for probabilities
- ▶ log for rates

Linking them through shared latent structure

A flexible shared-structure joint model:

$$\eta_i^{(B)} = \beta_0^{(B)} + \beta^\top x_i + S_0(s_i) + S_B(s_i)$$

Why include outcome-specific fields S_B, S_P ?

Because the two processes are not identical:

- ▶ prevalence and incidence are related but not the same
- ▶ facility data can reflect access and reporting biases
- ▶ survey prevalence can reflect age structure and diagnostics

Outcome-specific fields mop up these systematic differences.

What can go wrong?

Two common practical pitfalls:

1. **Confounding:**

- ▶ covariates and spatial fields compete

2. **Mis-specified exposure E_i :**

- ▶ Poisson component can dominate if E_i is wrong scale

3. **Different spatial supports:**

- ▶ facility counts might be aggregated (administrative units)

Part C — Non-stationarity

Why stationarity can be unrealistic

Stationary Matérn assumes:

- ▶ same smoothness everywhere
- ▶ same correlation length everywhere

But malaria transmission may be:

- ▶ smooth in the north (broad ecological gradients)
- ▶ rough in the south (heterogeneous land use, urbanisation)
- ▶ different across ecozones

Non-stationarity: the covariance structure changes over space.

Two practical nonstationary constructions

1. Mixture of SPDEs (smooth + rough)

- ▶ simple, and robust

2. Structured range in the SPDE

Part C1 — Mixture of two SPDEs

Idea: blend a smooth and a rough field

Let:

- ▶ $S_{\text{smooth}}(s)$: long-range Matérn
- ▶ $S_{\text{rough}}(s)$: short-range Matérn
- ▶ $\omega(s) \in [0, 1]$: spatial weight (e.g., function of latitude)

Define:

$$S(s) = \omega(s)S_{\text{smooth}}(s) + (1 - \omega(s))S_{\text{rough}}(s)$$

Then:

$$\eta(s) = \beta_0 + \beta^\top x(s) + S(s)$$

What does $\omega(s)$ do?

- ▶ If $\omega(s) \approx 1$: field behaves like smooth long-range
- ▶ If $\omega(s) \approx 0$: field behaves like rough short-range
- ▶ If $\omega(s)$ changes with location: correlation structure changes

Interpretation - $\omega(s)$ is a *nonstationarity driver*.

Choosing $\omega(s)$

Simple choices:

- ▶ monotone latitude function (north–south structure)
- ▶ ecozone indicator (piecewise structure)
- ▶ logistic function of a covariate:

$$\omega(s) = \text{logit}^{-1}(a + bz(s))$$

In the practical session:

- ▶ start with latitude-based $\omega(s)$

Part C2 — Structured range in the SPDE

Goal: let the Matérn range vary with a covariate

Recall: for Matérn SPDE models, range is approximately:

$$\rho(s) \approx \frac{\sqrt{8}}{\kappa(s)}$$

So if we model:

$$\log \kappa(s) = \theta_0 + \theta_1 z(s)$$

then

$\rho(s)$ varies smoothly with $z(s)$.

What does the sign of θ_1 mean?

Why use mesh-vertex covariates?

In the SPDE approximation:

- ▶ the field is defined through weights on mesh vertices
- ▶ so nonstationary parameters need to be specified at the same support (vertices)

Thus $z(s)$ is evaluated at mesh nodes:

- ▶ $z_k = z(\text{vertex}_k)$

This makes the range a spatially varying parameter.

What changes compared to stationary SPDE?

Stationary SPDE:

- ▶ one κ , one τ for all space

Structured-range SPDE:

- ▶ $\kappa(s)$ and $\tau(s)$ can depend on covariates
- ▶ introduces additional parameters θ controlling how they vary

Conceptual result:

Practical guidance for modelling

- ▶ Start with a simple covariate $z(s)$ (e.g. northing or ecozone index)
- ▶ Scale it to have mean 0, sd 1
- ▶ Use weak-to-moderate priors on the slope parameter to avoid extreme range variation

In interpretation:

- ▶ show the covariate map
- ▶ show the predicted prevalence map
- ▶ (optionally) show a derived “implied range” surface

Summary

Joint models - share spatial information across related outcomes - can handle: - multiple groups (same likelihood) - multiple likelihoods (binomial + poisson)

Nonstationarity - mixture SPDE: simple and robust - structured range SPDE: mechanistic, covariate-driven

Take-away - choose the simplest model that answers your scientific question - use nonstationarity when stationary assumptions are visibly violated

