Day 3 – Hybrid Machine Learning + Geostatistics

Olatunji Johnson

Overview

What Day 3 is about

We want **high-quality malaria prevalence maps** using:

- **Machine learning** to learn complex covariate effects
- **Geostatistics** to model residual spatial dependence
- **INLA/SPDE** to quantify uncertainty efficiently

The main idea:

$$\text{Observed data} \approx \underbrace{\text{Complex covariate signal}}_{\text{ML}} + \underbrace{\text{Residual spatial structure}}_{\text{SPDE/INLA}}$$

We will:

- let **ML** learn flexible, nonlinear covariate effects
- let **SPDE/INLA** capture remaining **spatial dependence**

Why hybrid ML + geostatistics?

Part 0 — Data and notation

Data structure (prevalence surveys)

At locations s_i, we observe:

- Y_i: number positive
- N_i: number tested

$$Y_i \mid p(s_i) \sim \text{Binomial}(N_i, p(s_i))$$

We work on the linear predictor scale:

$$\eta(s) = \text{logit}\{p(s)\}$$

Hybrid modelling target

We decompose the linear predictor as:

Part 1 — Baseline model (reference point)

Baseline model (geostatistics only)

We already covered MBG/SPDE in Day 1, so here it's a reference:

$$Y_i \mid p(s_i) \sim \text{Binomial}(N_i, p(s_i))$$

$$\text{logit}\{p(s)\} = m(x(s)) + S(s) + \epsilon(s)$$

Strengths

- coherent uncertainty
- explicit spatial dependence

Limitation

- covariate effects often too rigid (linear terms)

Part 2 — Machine learning for m(x)

What ML is doing here

We treat ML as a flexible estimator of the systematic component:

$$m(x) \approx E\{\eta(s) \mid x(s) = x\}$$

ML learns:

- nonlinearities
- interactions
- splitting rules / ensembles

Then we compute residuals:

$$r_i = \tilde{\eta}_i - \hat{m}(x_i)$$

If r_i is spatially correlated $\rightarrow$ we need a spatial residual model. If not???

Random Forest: core idea

RF model uncertainty: what it is and isn't

RF offers some uncertainty tools:

- variability across trees
- quantile regression forests
- conformal prediction

But a standard RF fit does not produce a coherent spatial posterior.

In this workshop:

- we use RF mainly to get a strong mean function $m(x)$
- we rely on INLA/SPDE to quantify spatial residual uncertainty

Part 3 — Diagnosing spatial structure after ML

Key diagnostic question

After ML, compute residuals:

$$r_i = \eta_i - \hat{m}(x_i)$$

If r_i is spatially correlated, ML has not captured all structure.
Then we fit:

$$r_i = S(s_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

This is a clean and interpretable hybrid decomposition.

Variogram for ML residuals

Empirical variogram:

$$\hat{\gamma}(u) = \frac{1}{2|N(u)|} \sum_{(1,j)\in N(u)} (r_i - r_j)^2$$

Interpretation:

- increasing variogram $\rightarrow$ decreasing correlation with distance
- near-zero at small distances $\rightarrow$ strong local similarity

Permutation envelope (quick test)

Procedure:

1. permute residuals across locations
2. compute variogram
3. repeat many times
4. build 95% envelope

If observed variogram lies outside envelope $\rightarrow$ evidence of spatial correlation.
This justifies adding the spatial residual field $S(s)$.

Part 4 — Spatial residual model (what INLA adds)

Spatial residual model (Gaussian)

Residuals are approximately continuous:

$$r_i \mid S(s_i) \sim N(S(s_i), \sigma^2)$$

$$S(s) \sim MaternSPDE$$

Outputs we get from INLA:

- posterior mean of $S(s)$
- posterior sd of $S(s)$
- posterior of hyperparameters (range, σ)
- fast prediction on dense grids

This is where the hybrid gets principled spatial uncertainty.

Why modelling residuals works well

Part 5 — Hybrid prediction and uncertainty

Hybrid predictor

At a new location s:

$$\eta_{hyb}(s) = \hat{m}(x(s)) + \hat{S}(s)$$

Convert back to prevalence:

$$\hat{p}_{hyb}(s) = \text{logit}^{-1}(\eta_{hyb}(s))$$

What uncertainty does the hybrid provide?

In principle there are three components:

1. Binomial sampling noise (from $Y \mid p$)
2. Spatial residual uncertainty (from INLA: $S(s)$)
3. ML uncertainty (from learning $m(x)$)

In this workshop, we quantify (2) very clearly, and discuss how to extend to (3).

Spatial residual uncertainty (what we can map)

INLA gives:

$$S(s) \mid data \approx \text{posterior with mean and sd}$$

So we can map:

- $E\{S(s) \mid y\}$
- $\text{sd}\{S(s) \mid y\}$

Propagation to prevalence (approx):

- uncertainty in η transforms through logistic curve

Why the hybrid improves uncertainty communication

Baseline geostatistical model uncertainty mixes:

- covariate uncertainty (via linear model)
- spatial field uncertainty

Part 6 — Comparison to baseline

What changes compared to baseline?

Baseline:

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

Hybrid:

$$\eta(s) = m(x(s)) + S(s)$$

So differences are entirely in how we model the systematic component:

- linear predictor vs flexible ML predictor

What should improve?

Hybrid often improves:

- predictive accuracy (RMSE / log score)
- realism of covariate-response relationships
- interpretability of residual spatial signal
- robustness when covariate effects are highly nonlinear

But hybrid can fail if:

- ML overfits
- covariates shift out-of-distribution at prediction locations
- residual model is mis-specified

A practical evaluation checklist

1. Does ML reduce residual spatial structure?
   - compare variograms (pre-ML vs post-ML residuals)
2. Does hybrid improve held-out prediction?
   - RMSE on prevalence or logit-prevalence
3. Are maps plausible?
   - no obvious artefacts, aligned with known epidemiology
4. Is uncertainty meaningful?
   - higher in sparse regions, lower in data-rich areas

Where to go next (research directions)

Full Bayesian hybrid:

- BART + SPDE
- Bayesian deep learning + spatial random effects
- Joint models with multiple outcomes + ML mean functions
- Nonstationary residual fields (Day 2 ideas + ML)

Spatio-temporal hybrids (Bamidele Toba):

$$\eta(s,t) = m(x(s,t)) + S(s,t)$$