ADIM: Bayesian Spatial Statistics

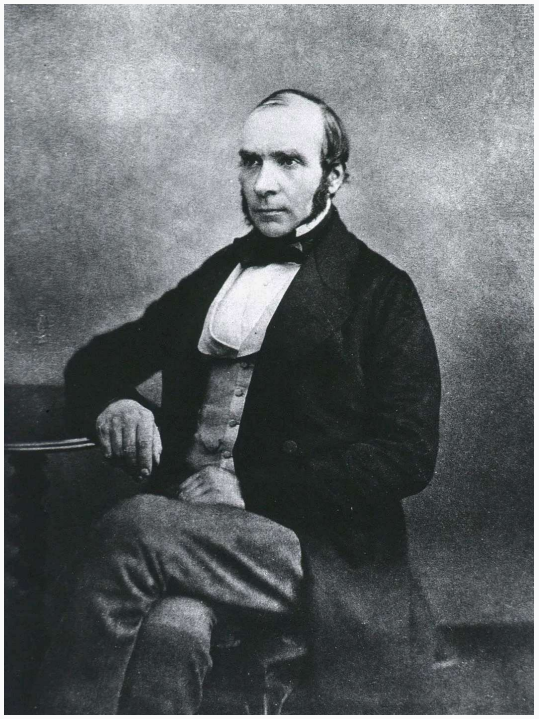
Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering

Joaquín Martínez-Minaya, 2024-12-16

VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante
jmarmin@eio.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

John Snow's Cholera Map in London 1854'



Outline

1. Spatial statistics. Types of spatial data
2. Disease mapping
3. Geostatistics
4. Penalized complexity priors (PC-priors)
5. References

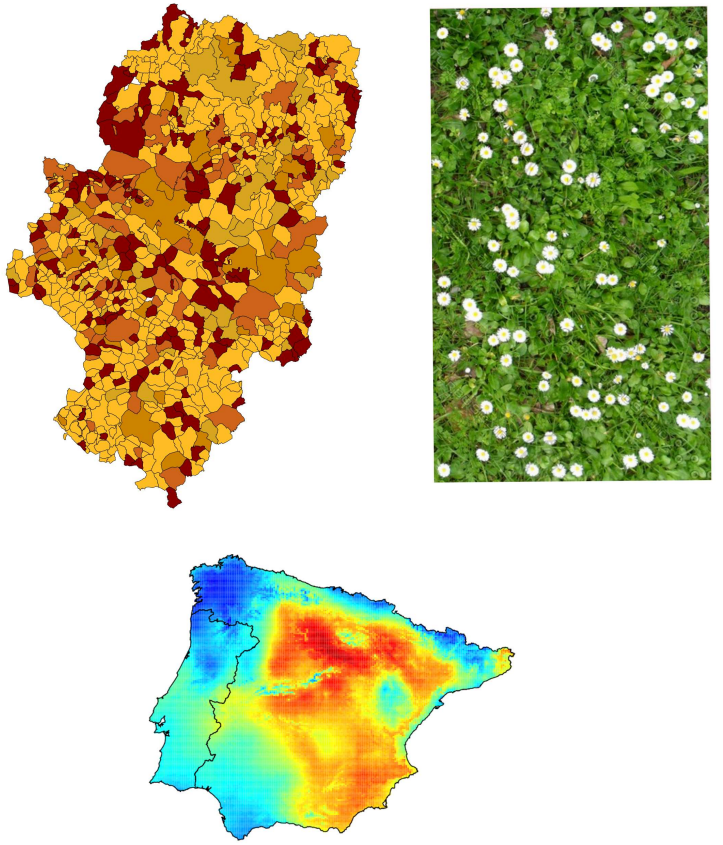
1. Spatial statistics. Types of spatial data

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Spatial statistics. Types of spatial data

Spatial statistics is defined as the part of statistics which deal with spatial data and study spatial patterns.

- **Lattice or areal data**: observations are taken at a finite number of sites whose whole constitutes the entire study region (discrete space), e.g. number of sick people by provinces.
- **Point pattern**: the interest is study the process which generates the points. e.g. distribution of trees in a mountain.
- **Geostatistical data**: consist of a collection of data in a fixed set locations over a continuous spatial field, e.g. amount of fish in the ocean or presence/absence of a plant in a country.

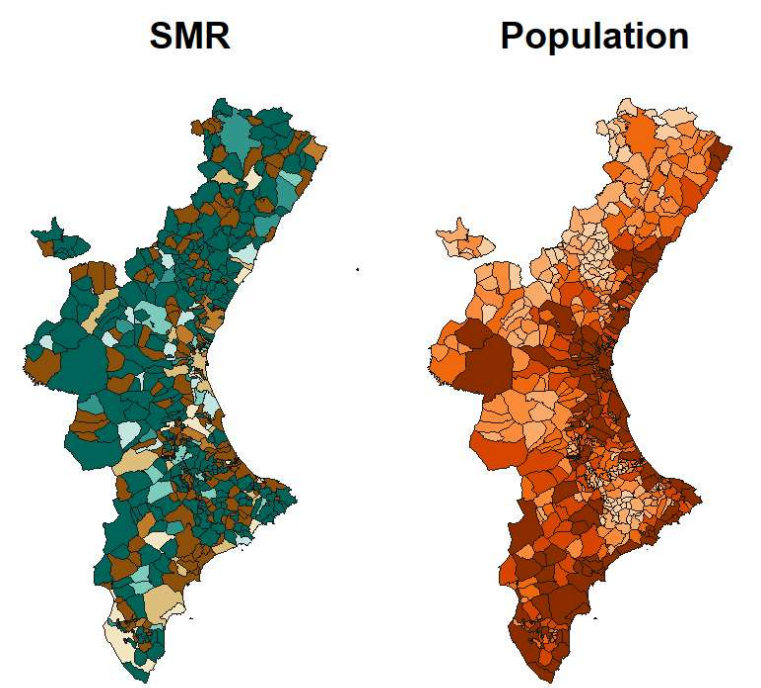


2. Disease mapping

Oral Cancer mortality in Valencian Region

- In this analysis, we study **oral cancer mortality in the municipalities of the Valencian Region** using a disease mapping model. The aim is to understand the spatial distribution of risks and identify high-risk areas while accounting for variability due to population size and random noise. The variables are:

- **Obs**: the number of observed deaths from oral cancer in the study period.
- **Exp**: the number of expected deaths, based on population size and age-specific rates.
- **SMR**: the standardized mortality ratio, calculated as $\text{SMR} = \frac{\text{Obs}}{\text{Exp}} \cdot 100$
  - **SMR = 100**: Risk is equivalent to the standard population.
  - **SMR > 100**: **excess risk**.
  - **SMR < 100**: **reduced risk**.



The model

- A conditional independent **Poisson** likelihood function is assumed:

$$y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = E_i \rho_i, \quad \log(\rho_i) = \eta_i \ , i = 1, \dots 32$$

- We assume that $\eta_i = \beta_0 + u_i + v_i$, being $\boldsymbol{u}$ the **independent random effect** and $\boldsymbol{v}$ the **spatially structured random effect**:

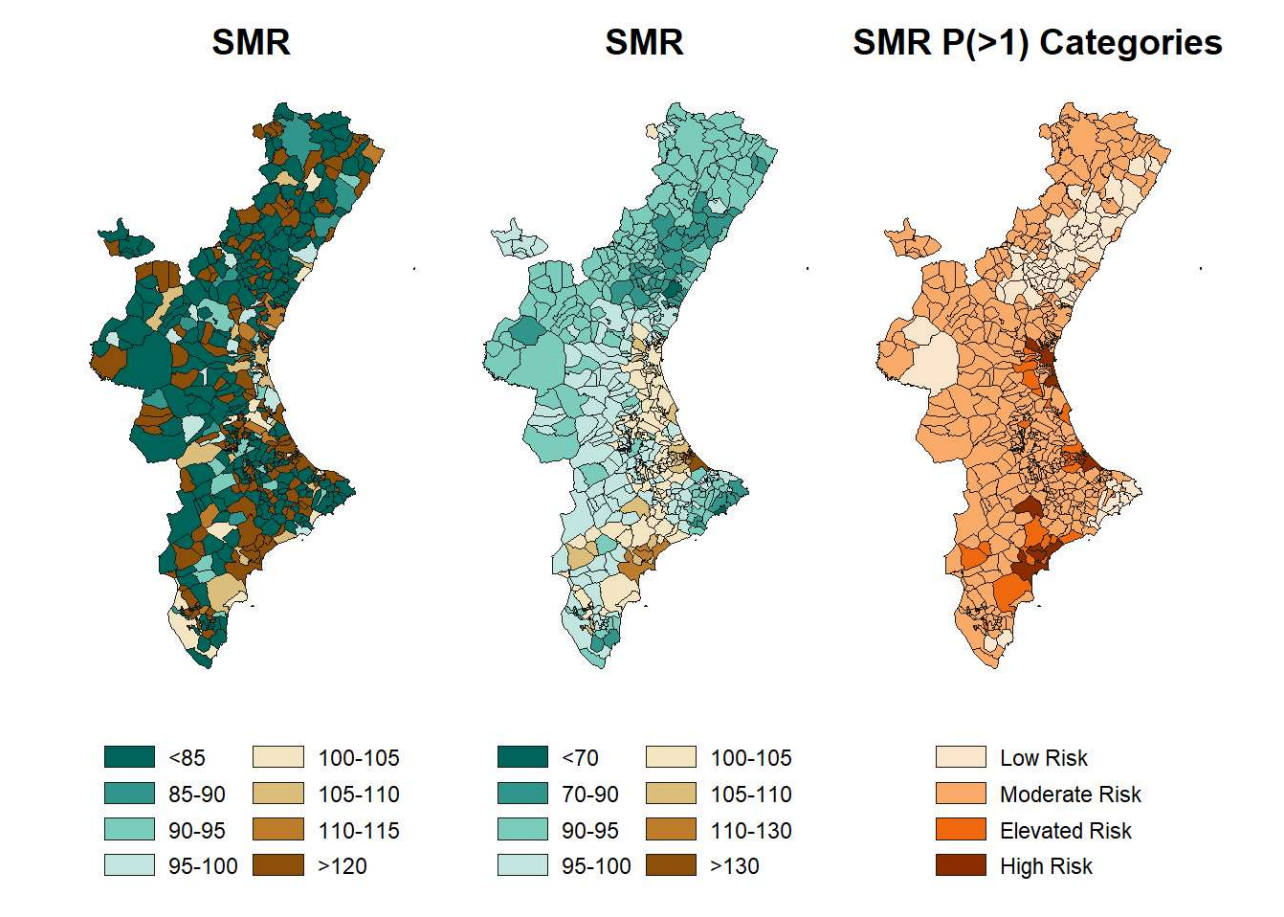
$$u_i \sim \mathcal{N}\left(0, \tau_{\boldsymbol{u}}^{-1}\right), \ v_i \mid \boldsymbol{v}_{-i} \sim \mathcal{N}\left(\frac{1}{n_i} \sum_{i \sim j} v_j, \frac{1}{n_i \tau_{\boldsymbol{v}}}\right) \ .$$

In this case $\boldsymbol{\theta} = (v_1, \dots, v_{32}, u_1, \dots, u_{32})$, and $\boldsymbol{\theta} \mid \boldsymbol{\psi}$ is Gaussian distributed.

- **Hyperpriors** for the standard deviation parameters σ_u and σ_v follow uniform priors:

$$\sigma_u, \sigma_v \sim \text{Uniform}(0, \infty)$$

Predicting Risk in Valencia Region

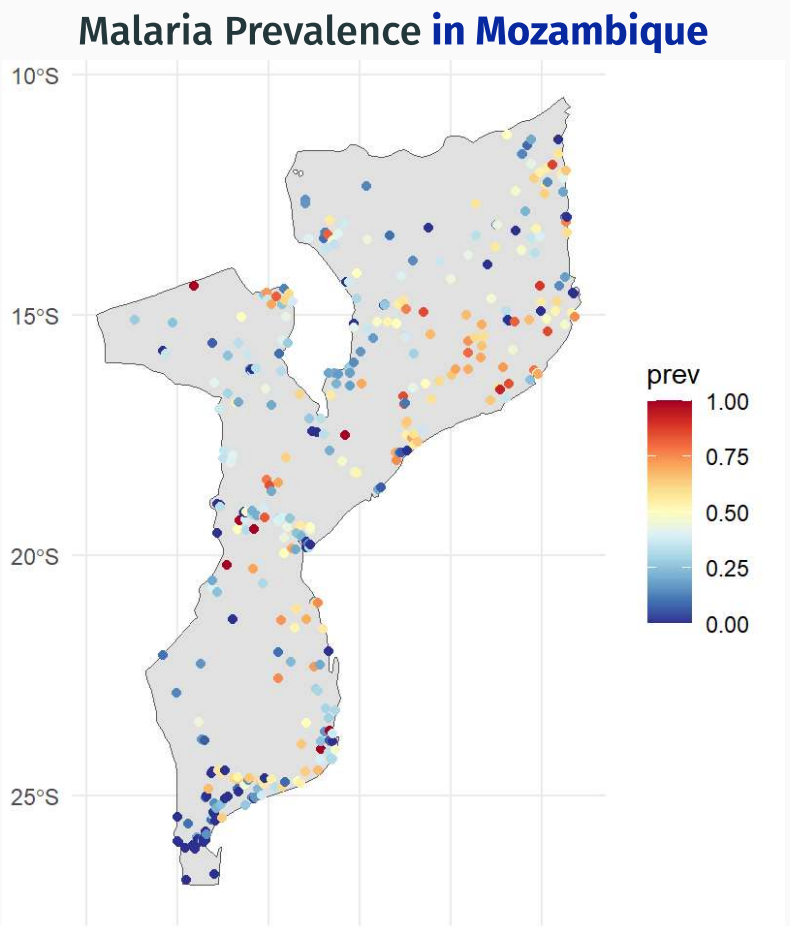


3. Geostatistics

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Continuous spaces

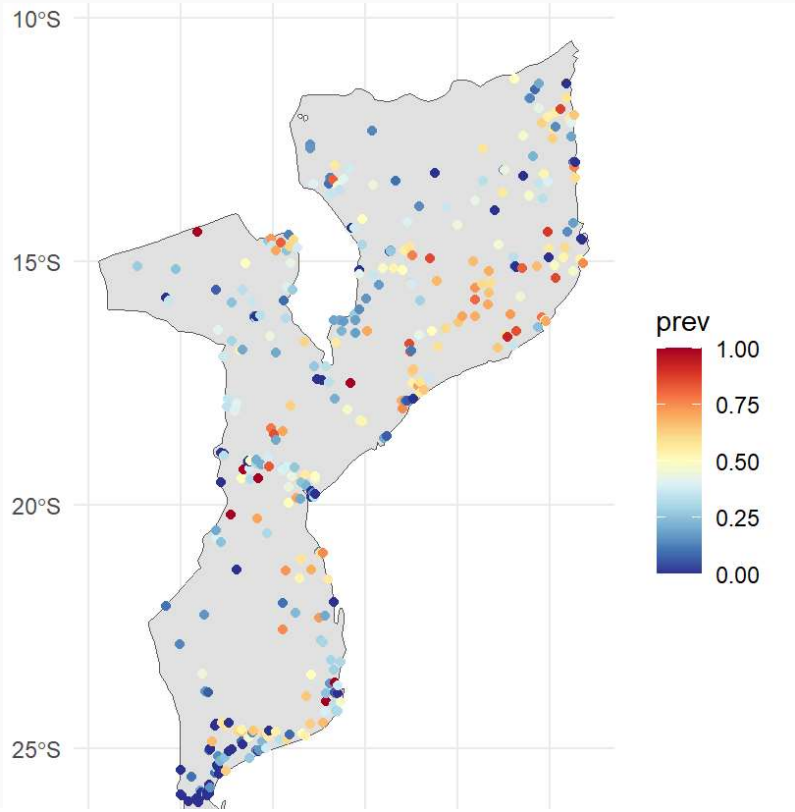
- Sometimes, the assumption that the observations have been collected over **discrete time** points have to be removed.
- The same happen in **space**.
    - If we are studying the **presence of a disease**, **pollution in an area** or the temperature of a country, the locations where the phenomenon of interest is measured are not frequently allocated in a lattice.
- Then, we are dealing with **continuous spaces** in 1D and 2D



Malaria Prevalence in Mozambique

- This analysis studies **malaria prevalence in Mozambique** using a spatial Bayesian model. The goal is to predict malaria risk and evaluate the effects of environmental and demographic covariates.

- **Examined**: the number of individuals examined for malaria.
- **Positive**: the number of individuals testing positive for malaria.
- **Covariates**:
  - **Altitude**: Elevation of the study location (in meters).
  - **Temperature**: Average temperature (in °C).
  - **Proximity to water bodies**: Distance to the nearest water source (in kilometers).



Geostatistics. Basis

- Geostatistical models assume that the observations are correlated.

- They are based on the following principle

Everything is related to everything else, but near things are more related than distant things

- So, two close locations tend to **co-vary** more than those far from each other.

Let's be a bit more formal

- A random spatial effect $w(\boldsymbol{s})$ at a location $\boldsymbol{s} \in \mathcal{D}$ can be considered as a **stochastic process** characterized by a spatial index $\boldsymbol{s}$ which varies continuously in the fixed domain $\mathcal{D}$, where $\mathcal{D}$ is a fixed subset of r-dimensional Euclidean space.

- The spatial process $w(\boldsymbol{s})$ is Gaussian if for any $n \geq 1$ and any set of sites $\boldsymbol{s} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, $w = \{w(\boldsymbol{s}_1), \ldots, w(\boldsymbol{s}_n)\}$ has a multivariate normal distribution with mean $\mu = E(w(\boldsymbol{s}))$ and a structured covariance matrix Σ. Usually μ is assumed to be $\mathbf{0}$. In the literature, this process is widely known as a **Gaussian field (GF)**.

- The key issue in spatial statistics is the covariance function $\mathcal{C}$, which determines the covariance between random variables in two different points. If $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are two locations in space, then the **covariance function** is defined as

$$\mathcal{C}(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j)) = Cov(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j))$$

- It defines the covariance matrix $\boldsymbol{\Sigma}$ of the GF. Each element of the matrix $\boldsymbol{\Sigma}_{ij}$ is defined as:

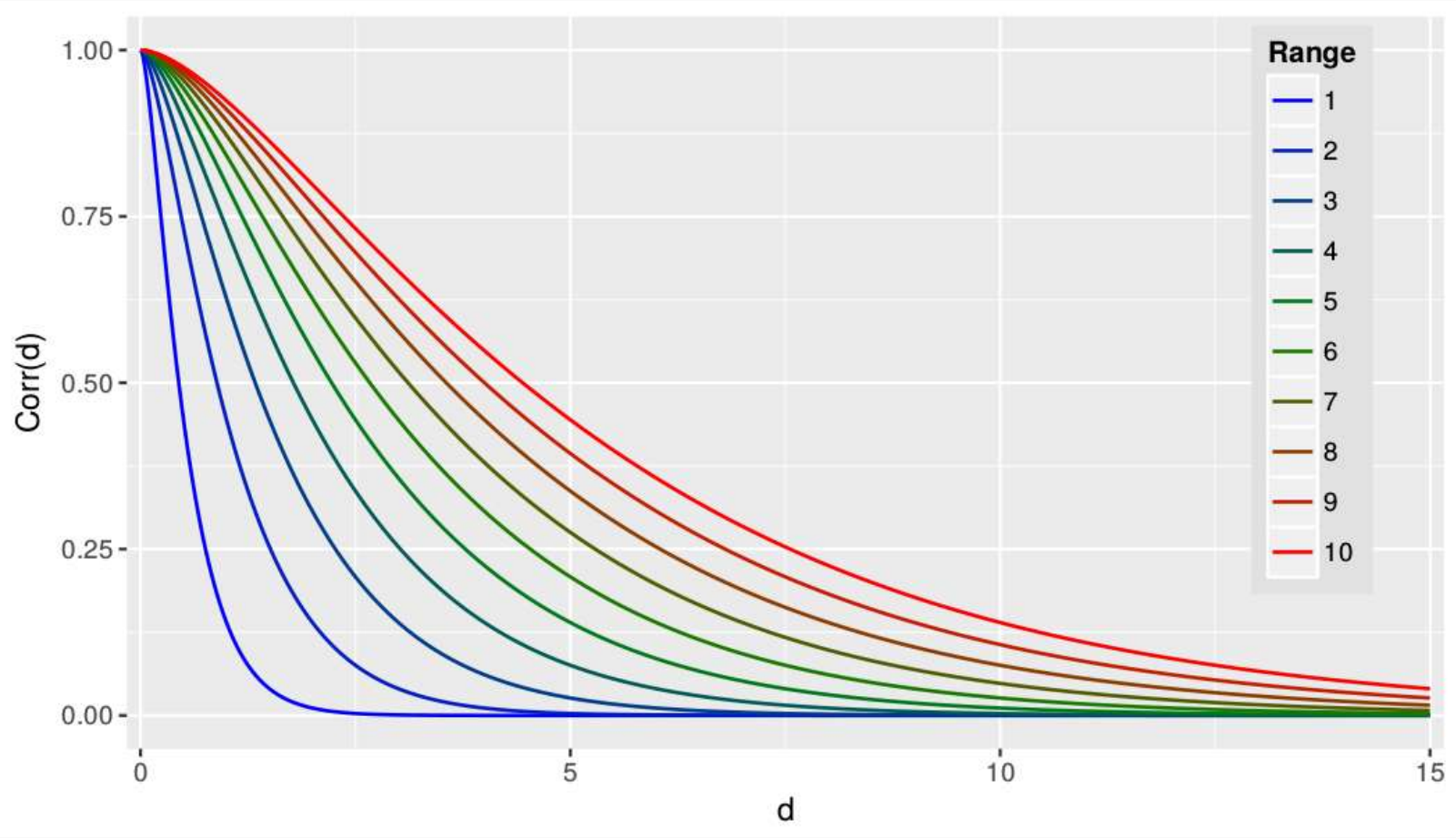
$$\boldsymbol{\Sigma}_{ij} = \mathcal{C}(w(\boldsymbol{s}_i), w(\boldsymbol{s}_j))$$

Matérn

- **Stationarity**. We say that the GF is second-order stationary if $\mu(s) = \mu$ and $Cov(w(s), w(s+h)) = \mathcal{C}(h)$ for all $h \in \mathcal{R}$ such that s and $s+h$ lie within $\mathcal{D}$. The covariance function in two different locations depends on the distance vector between these two locations.
  - An example could be the spread of a pathogen in plants. If there is a road close to the crop, maybe this pathogen could spread faster along the road in cars or trucks than in the crop, it would depend on the direction.
- **Isotropy**. We say that the GF is isotropic if the covariance function depends only on the Euclidean distance between points, i.e., $Cov(w(s), w(s+h)) = C(||h||)$.
  - For instance, if we think again in the spread of a pathogen in a crop, it would mean that the spread does not depend on the direction, just on the distance.
- **Matérn correlation** function is very common.

$$\mathcal{C}(||h||) = \sigma_w^2 \left(\frac{\sqrt{8}}{\phi} ||h|| \right) K_1 \left(\frac{\sqrt{8}}{\phi} ||h|| \right)$$

Matérn correlation function



Geostatistics in the context of LGMs

Likelihood

- A conditional independent **Binomial likelihood** function is assumed:

$$y_i \mid \pi_i \sim \text{Binomial}(n_i, \pi_i),\ \eta_i = \text{logit}(\pi_i) = \beta_0 + \beta_1 Temp + w_i\ , i = 1, \ldots 447$$

Latent Gaussian field

$$\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{\Sigma}(\sigma_{\boldsymbol{w}}, \phi)),\ \beta_j \sim \mathcal{N}(0, \tau = 0.001)$$

$\boldsymbol{\theta} = (\beta_0, \beta_1, w_1, \ldots, w_{447})$, and $\boldsymbol{\theta} \mid \boldsymbol{\psi}$ is Gaussian distributed.

- $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\sigma_w}, \boldsymbol{\phi}))$, i.e., the spatial effect is assumed to be a **continuous Gaussian field (GF)** with Matérn covariance structure, where:
- $\Sigma(\sigma_{\boldsymbol{w}}, \phi)$ is a **covariance matrix** depending on the distance between locations, $\sigma_{\boldsymbol{w}}$ is the **variance** of the spatial effect, and ϕ is the **range** of the spatial effect.

Hyperparameters $\boldsymbol{\psi} = (\sigma_{\boldsymbol{w}}, \phi)$

Problem: INLA can not fit continuous GFs

Solution: approximate the continuous GFs using the Stochastic Partial Differential Equation approach (SPDE)

The SPDE approach

Likelihood

$$y_i \mid \pi_i \sim \text{Ber}(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 Temp + w_i$$

Latent Gaussian field

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau = 0.0001)$$

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\sigma_w}, \boldsymbol{\phi}))$$

Hyperparameters

$$p(\sigma_{\boldsymbol{w}}, \phi)$$

Likelihood

$$y_i \mid \pi_i \sim \text{Ber}(\pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 Temp + w_i$$

Latent Gaussian field

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau = 0.0001)$$

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\sigma_w}, \boldsymbol{\phi}))$$

Hyperparameters

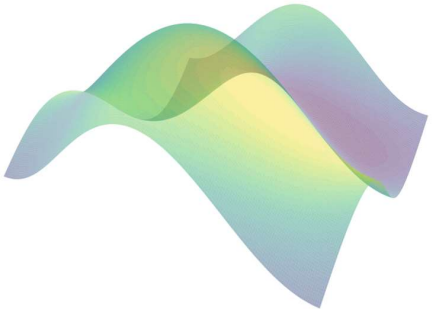
$$p(\sigma_{\boldsymbol{w}}, \phi)$$

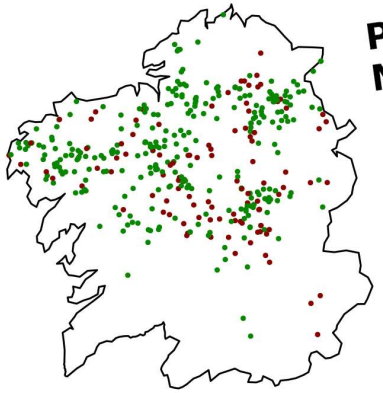
How is the approximation conducted?

SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau w(\boldsymbol{s})) = \mathcal{U}(\boldsymbol{s}), \ \boldsymbol{s} \in \mathbb{R}^d$$

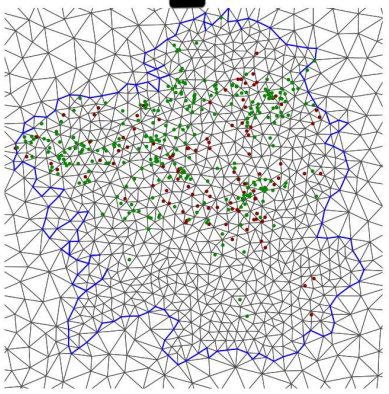
SURFACE



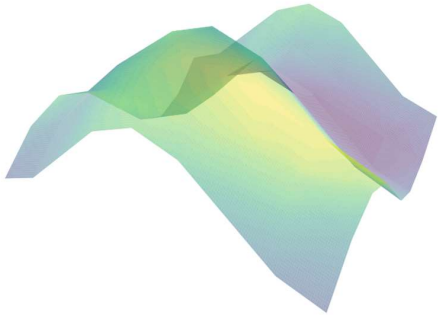


PROJECTION MATRIX

FEM

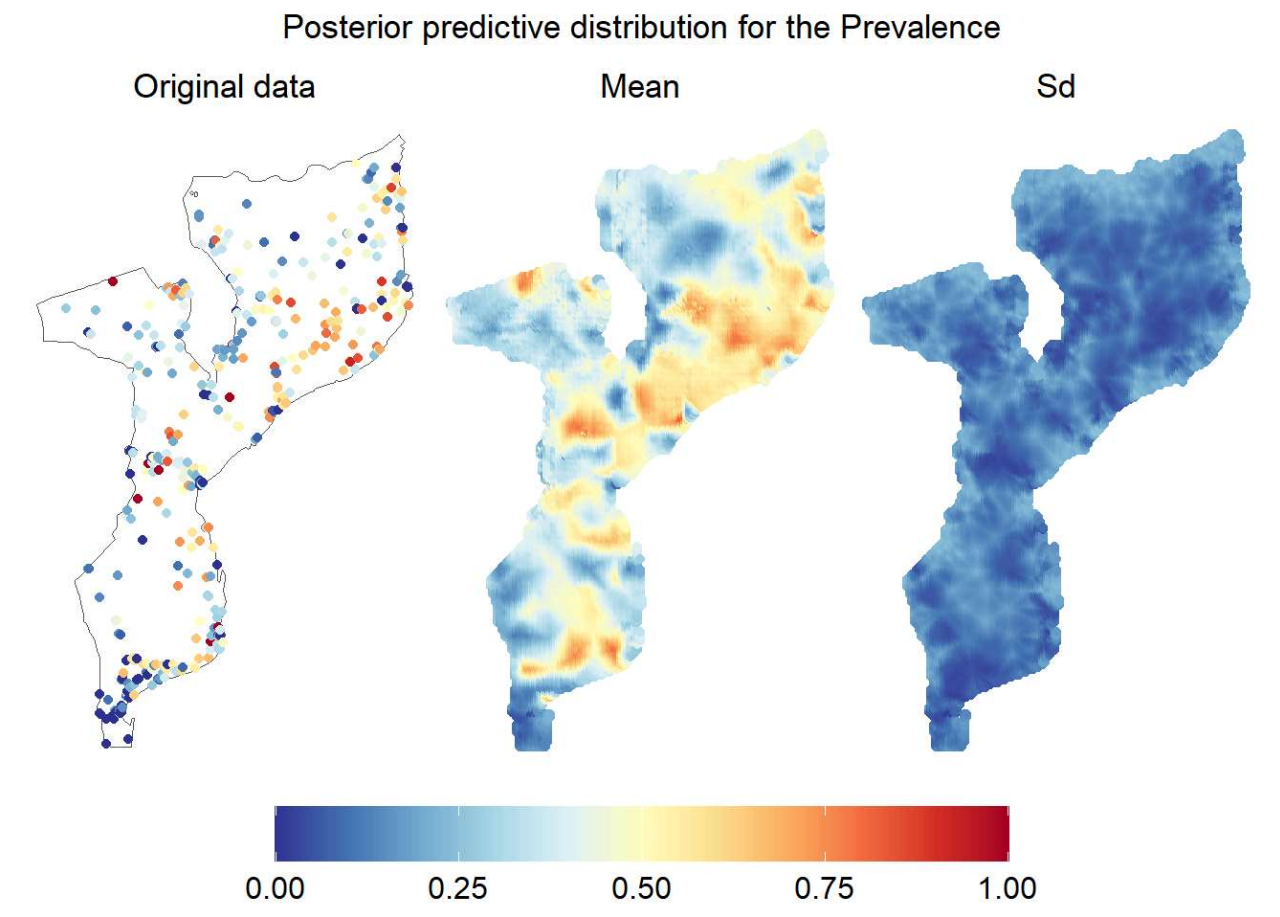


MESH



$Q(\sigma^2, \phi)$

Malaria Prevalence in Mozambique



4. Penalized complexity priors (PC-priors)

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Penalizing departure from the base model

- Simpson et al. (2017) propose priors that penalize departure from a base model and for this reason they are called **Penalized Complexity (PC) priors**.
- The prior favors the base model unless evidence is provided against it, following the principle of parsimony.
- Distance from the base model is measured using the **Kullback-Leibler** distance, and penalization from the base model is done at a **constant rate on the distance**.
- Finally, the PC prior is defined using **probability statements** on the model parameters in the appropriate scale.

Hyperpriors for the standard deviation in an iid

- The **PC-prior for the precision** τ has density:

$$p(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp(-\lambda\tau^{-1/2}),\ \tau > 0,$$

where

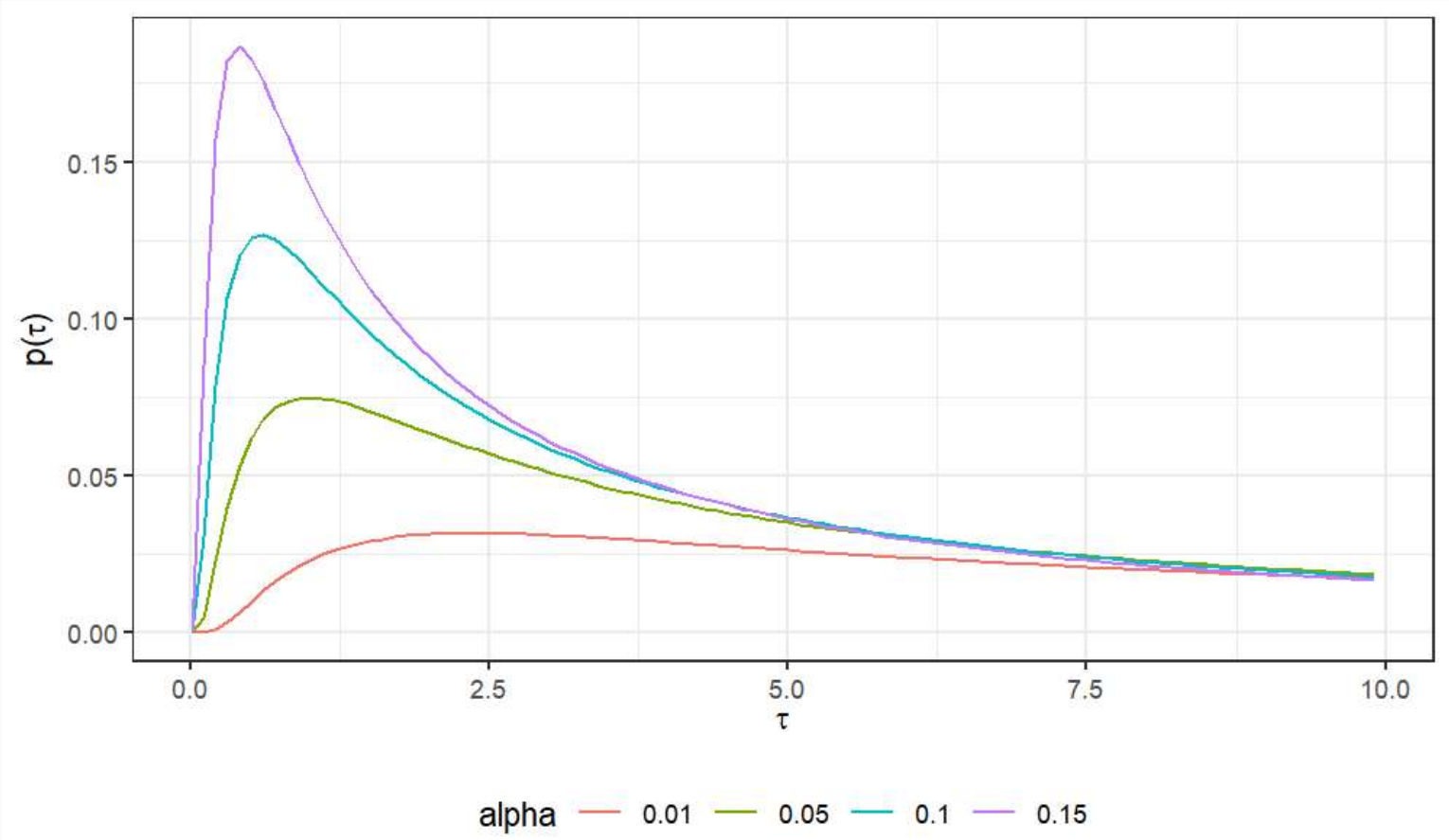
$$\lambda = -\frac{ln(\alpha)}{u},$$

and (u, α) are the parameters to this prior. The interpretation of (u, α) is that:

$$Prob(\sigma > u) = \alpha,\ u > 0,\ 0 < \alpha < 1.$$

- Functions `inla.pc.{d,p,q,r}.prec` allow us to **deal with this priors**.
- If we want to plot the prior in terms of the **standard deviation** σ, remember that using function `inla.tmarginal` we can go from the τ parameter to σ parameter.

Hyperpriors for the standard deviation in an iid.



Spatial effect: priors

- The PC-prior for the **range** is defined in terms of ϕ_0 and p_1 so that

$$Prob(\phi < \phi_0) = p_1$$

- The PC-prior for the **standard deviation** is defined in terms of σ_0 and p_2 so that

$$Prob(\sigma_{\boldsymbol{w}} > \sigma_0) = p_2$$

- In order to define the SPDE using PC-priors, the following command have to be used:

```
spde ← inla.spde2.pcmatern(
mesh = ...,
prior.range = c(phi0, p1),
prior.sigma = c(sigma0, p2))
```

5. References

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

This material has been constructed based on:

- Moraga, P., Dean, C., Inoue, J., Morawiecki, P., Noureen, S. R., & Wang, F. (2021). Bayesian spatial modelling of geostatistical data using INLA and SPDE methods: A case study predicting malaria risk in Mozambique. Spatial and Spatio-temporal Epidemiology, 39, 100440.
- Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.
- Fuglstad, G. A., Simpson, D., Lindgren, F., & Rue, H. (2019). Constructing priors that penalize the complexity of Gaussian random fields. Journal of the American Statistical Association, 114(525), 445-452.
- INLA tutorials
- INLA book by Virgilio Gómez-Rúbio
- INLA book by Paula Moraga
- SPDE book by Krainski et al.

ADIM: Bayesian Spatial Statistics

Master's Degree in Data Analysis, Process Improvement and Decision Support Engineering

Joaquín Martínez-Minaya, 2024-12-16
VAlencia BAyesian Research Group
Statistical Modeling Ecology Group
Grupo de Ingeniería Estadística Multivariante

jmarmin@eio.upv.es

UNIVERSITAT POLITÈCNICA DE VALÈNCIA