Lecture 4

The Binomial geostatistical model

1. Formulation of linear geostatistical models and assumptions.
2. Brief introduction to Gaussian processes.
3. Understanding the nugget effect.
4. Parameter estimation via the maximum likelihood method.

Defining geostatistical problems

The ingredients:

Defining geostatistical problems

The ingredients:

- S = process of nature (e.g. disease risk)

Defining geostatistical problems

The ingredients:

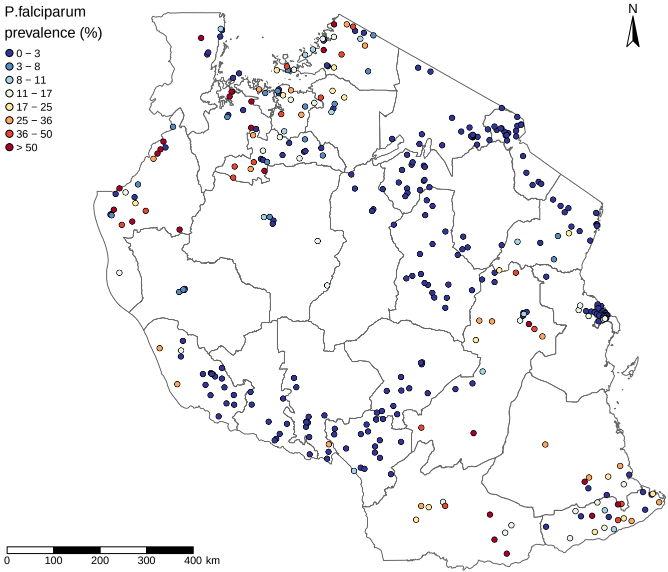
- S = process of nature (e.g. disease risk)
- Y = data

Defining geostatistical problems

The ingredients:

- S = process of nature (e.g. disease risk)
- Y = data
- A statistical model $[S, Y] = [S] \times [Y|S]$

Example: Malaria in Tanzania



Standard model for prevalence mapping

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process
- Z_i = unstructured random effects

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process
- Z_i = unstructured random effects
- Assumption: $Y_i|S(x_i), Z_i \sim \text{Bin}(n_i, p(x_i))$

$$\log\left\{\frac{p(x_i)}{(1-p(x_i)}\right\} = d(x_i)^\top\beta + S(x_i) + Z_i$$

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process
- Z_i = unstructured random effects
- Assumption: $Y_i | S(x_i), Z_i \sim \text{Bin}(n_i, p(x_i))$

$$\log\left\{\frac{p(x_i)}{(1 - p(x_i)}\right\} = d(x_i)^\top \beta + S(x_i) + Z_i$$

- Public health question: where are areas that are hihgly likely to exceed 30%?

Standard model for prevalence mapping

- Data: x_i = location of the cluster; n_i = number of sampled individuals at x_i; y_i=number of positively tested individuals at x_i
- $d(x_i)$ = vector covariates
- $S(x)$ = spatial stochastic process
- Z_i = unstructured random effects
- Assumption: $Y_i | S(x_i), Z_i \sim \text{Bin}(n_i, p(x_i))$

$$\log\left\{\frac{p(x_i)}{(1-p(x_i)}\right\} = d(x_i)^\top \beta + S(x_i) + Z_i$$

- Public health question: where are areas that are hihgly likely to exceed 30%?

Objectives of the study: 1) How do we incorporate $d(x)$ into $p(x)$? 2) How do we assess the impact of $d(x)$ on spatial prediction?

To predict or to explain?

Galit Shmueli (2010) "To Explain or to Predict?." Statistical Science 25 (3) 289 - 310 https://doi.org/10.1214/10-STS330

- **Explanatory modelling:** maximize the predictive performance of the model
- **Predictive modelling:** emphasis is placed on understanding the relationships between the health outcome and risk factors

Stages of a (geo)statistical analysis

1. Exploratory analysis
2. Model formulation and parameter estimation
3. Spatial prediction

Modelling non linear relationships

- Patterns:
  - **Unimodal:** an increasing trend is followed by a decreasing one, or vice-versa.
  - **Saturation:** monotonic relationship that appears to flatten for increasing (or decreasing) values of the covariate.

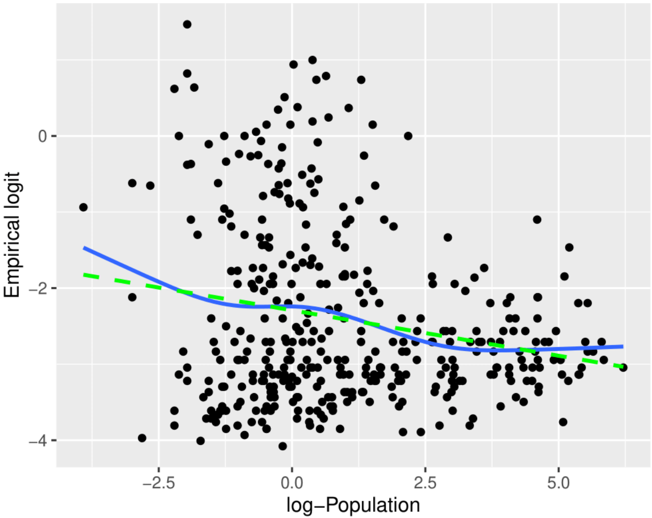
Modelling non linear relationships

- Patterns:
  - **Unimodal:** an increasing trend is followed by a decreasing one, or vice-versa.
  - **Saturation:** monotonic relationship that appears to flatten for increasing (or decreasing) values of the covariate.
- Empirical logit: $\log\{(y_i + 0.5)/(n_i - y_i + 0.5\}$

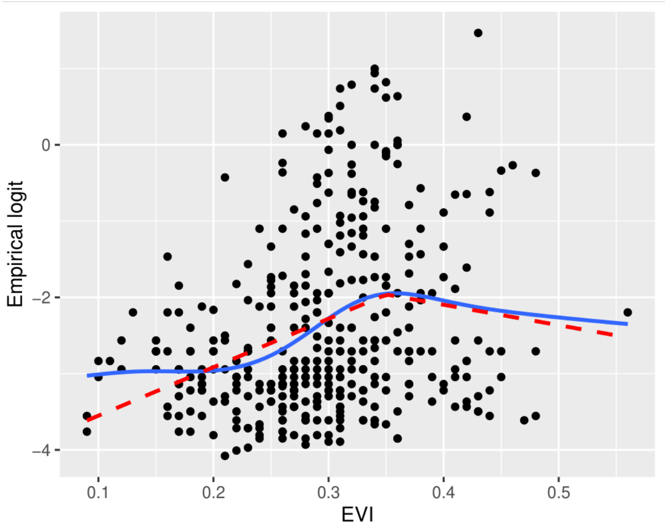
Modelling non linear relationships

- Patterns:
  - **Unimodal:** an increasing trend is followed by a decreasing one, or vice-versa.
  - **Saturation:** monotonic relationship that appears to flatten for increasing (or decreasing) values of the covariate.
- Empirical logit: $\log\{(y_i + 0.5)/(n_i - y_i + 0.5\}$
- Linear spline: $f\{d(x)\} = \beta d(x) + \gamma \max\{d(x) - c, 0\}$

Covariate: population



Covariate: enhanced vegetation index



Standard geostatistical model for prevalence mapping

- $S(x_i)$ is stationary and isotropic zero-mean Gaussian process with covariance
$$\mathrm{cov}\{S(x_i), S(x_j)\} = \sigma^2 \exp\{-u/\phi\}$$
- Z_i is an unstructured random effect mean 0 and variance τ^2
- $Y_i | S(x_i), Z_i \sim \mathrm{Bin}(n_i, p(x_i))$

$$\log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = d(x_i)^\top \beta + S(x_i) + Z_i$$

- The likelihood function for $\theta = (\beta, \sigma^2, \phi, \tau^2)$:

$$L(\theta) = \int [W][Y|W] \, dW$$

where $W = (S(x_1) + Z_1, \ldots, S(x_n) + Z_n)$.

A non-spatial model for prevalence survey data

Design

- Sample communities $i = 1, ..., n$.
- In community i, sample m_i individuals of whom Y_i test positive for disease of interest.
- Associated covariates w_i

Model

- $p_i =$ probability that a randomly sampled individual
  in community i will test positive
- $\log\{p_i/(1-p_i)\} = \alpha + w_i'\beta$
- $Y_i \sim \mathrm{Binomial}(m_i, p_i)$, mutually independent

Spatial prediction

- Define a **predictive target**:
  - Surface of prevalence: $\{p(x) : x \in A\}$
  - District-level estimates:

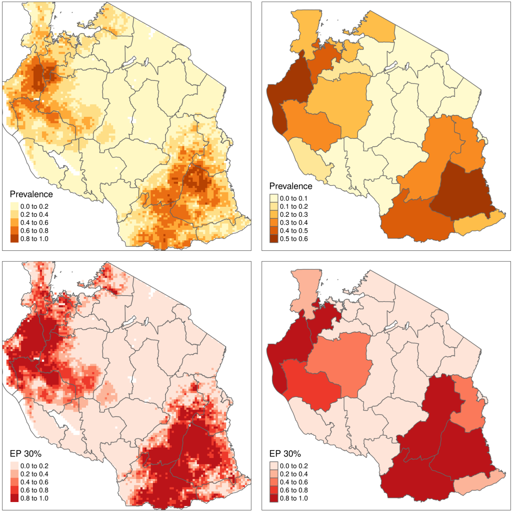
$$p(R_k) = \frac{\int_{R_k} w(x)p(x)\,dx}{\int_{R_k} w(x)\,dx}, k = 1, \ldots, K \tag{1}$$

  where $w(x)$ is the population density at location x.
- Quantify uncertainty
  - Classical summaries: standard errors, quantiles, coefficient of variation,...
  - **Exceedance probability**

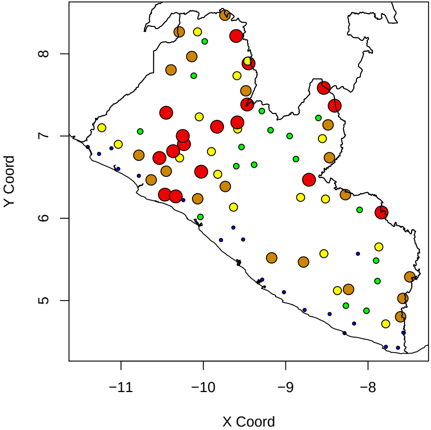
$$\mathrm{Prob}(p(x) < l | y_1, \ldots, y_n)$$

  where l is 0.3 (or 30%) is in our example.

Spatial prediction



Exploratory analysis of onchocerciasis data



- patches of high and low prevalence
- increasing trend away from coast?