Species Distribution Modeling

A statistical review with focus on spatio-temporal issues

Prepared by Evgeny Noi

Applications in ecology and epidemiology

- identifying critical habitats
- risk assiciated with invasive species
- effects of climate change
- designing protected areas
- distribution of infectious decease (COVID, monkey pox)

Background

- Modeling is traditionally within GIS framework
- Non-independence
- Non-gaussianity
- Big data poses challenges in: storage, algorithms, non-linearities
- Imperfect survey data
- Gaps
- Etc.

Disclaimer on Types of Spaces

- Attribute Space
- Parameter Space
- Reference Space

Definitions and scope of this survey

- Geostatistical data (point-referenced data)
- Spatially-continuous phenomena
- Model-based geostatistics approach

Bayesian Hierarchical Framework

- Allows incorporating prior information about phenomena
- Discriminates spatial autocorrelation effects from ordinary non-spatial linear effects
- Uses integrated nested Laplace approximations (INLA) --> Speeding up computations

Sources of Information

- Biological data (distribution of phenomena in space)
- Environmental data (predictors)

Gaussian Fields and Hierarchical Modeling

- presence-only datasets (max enthropy algorithm, envelope methods)
- ML algorithms (ANN, ensembles, generalized linear / additive models, multivariate adaptive regression splines)
    - GAM and MARS model spatial and temporal autocorrelation using smoothing splines

Spatial Processes

- $y(s)$, $s \in \mathcal{D}$, where s is a spatial index
- For any set of n locations ($n \geq 1$): $(s_1, \ldots, s_n)$; vector $(y(s_1), \ldots, y(s_n))$ follows multivariate Normal distribution with mean $\mu = (\mu(s_1), \ldots, \mu(s_n))$ and ccovariance matrix Σ defined by a covariance function $C(\cdot, \cdot)$ such that
- $\Sigma_{ij} = Cov(y(s_i), y(s_j)) = C(y(s_i), y(s_j))$

Stationarity and Anisotropy

- If the mean is constant in space, the generic spatial covariance matrix elements depends only on the difference vector $(\boldsymbol{s}_i - \boldsymbol{s}_j) \in \mathbb{R}^2$ the spatial process is **second-order stationary**
- If covariance function only depends on the Eulclidean distance $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ the process is said to be **isotropic**

Computational Difficulties

- 'big n problem
- Spatial covaraince represented by dense matrices
- Use SPDE to cut down on computational expenses

General Modeling Structure

$$\eta_i = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + \sum_{l=1}^{L} f_l(z_{li})$$

where β_0 - intercept, β - coefficients quantifying linear effect of covariates on the response, and f are unknown functions of the covariate, that can assume different forms: smooth non-linear effects, time trends, and seasonal effects, random intercept and slopes, spatial and temporal random effects. Equivalent to GAMM.

Bayesian Framework

- data level: process and parameters of the model
- latent components (spatial/dynamic, uni/multivariate)
- priors of the parameters

INLA and SPDE

- Spatial distribution models $\sim$ latent Gaussian models
- Identify distribution of the observed data and link its mean to the linear predictor (previous formula)

What is INLA?

- Approximate Bayesian Inference for latent Gaussian (GMRF) models
- pros: low computational cost, high accuracy, LGM!, complex models, applied and applicable, works well with hard-to-fit models (prior to INLA)

Latent Gaussian Models

- Observations y_i
- Gaussian random field x_j
- non-Gaussian hyperparameters θ_k

$$\pi(x, \theta | y) \propto \pi(\theta) \times \pi(x | \theta) \times \prod \pi(y_i | x_i, \theta)$$

- The main task is compute posterior marginal distributions (via integrated Laplace approximations)

$$\pi(x_j | y) \quad \text{and} \quad \pi(\theta_k | y)$$

Examples of LGMs

dynamic linear models, GLM, GAM, measurement error models, spline smoothing, functional data analysis, log-Gaussian Cox-processes, model-based geostatsics, survival models.

LGMs only a way to compute, not to model!

Laplace approximation

- Approximating integrals

$$\int \exp(-ng(x))dx \approx \sqrt{\frac{2\pi}{ng''(x_0)}}$$

- High accuracy: relative error $\mathcal{O}(1/n)$
- Algo: for each x_1:
  - compute mode and curvature at the mode
  - use the Laplace approximation to integrate out x_2

GMRF - 'Gaussian on graphs'

- conditional independence
- sparse precision matrix q
- numerical algorithms for sparse matrices

Under conditional independence:

$$p(y|\theta,\psi) = \prod_{i=1}^{n} p(y_i|\theta_i,\psi)$$

where θ - set of latent components (latent field) and ψ denotes vector of K hyperparameters. Normal prior distribution on θ with mean 0 and precision matrix $\mathcal{Q}(\psi) : \theta \sim N(0, \mathcal{Q}^{-1}\theta)$, with density:

$$p(\theta|\psi) = (2\pi)^{-n/2}|\mathcal{Q}(\psi)|^{1/2}exp(-\frac{1}{2}\theta'\mathcal{Q}(\psi)\theta)$$

INLA

- When the precision matrix $\mathcal{Q}(\psi)$ is sparse a GF becomes Gaussian Markov random field (GMRF), yielding computational benefits.
- INLA Cannot be applied when dealing with models that have geostatistical data (continuously indexed Gaussian Fields) because parametric covariance function needs to be specified and fitted based on data, which determines covariance matrix Σ and enables predictions at unsampeld locations
- The cost of factorizing dense covariance matrix Σ is cubic in its dimensions (inverse and determinant).

Alternative SPDE formulations

- approxximate stochastic weak solution to SPDE is GMRF approximation with Matern covariance structure:

$$(k^2 - \Delta)^{\alpha/2}(\tau\xi(s) = \mathcal{W}(s)$$

where $s \in \mathbb{R}^2, \alpha = v + \delta/2, k > 0, v > 0$, Δ is the Laplacian, α controls the smoothness, k is the scale parameter, τ controls the variance and $\mathcal{W}$ is a Gaussian spatial white noise process.

Matern Covariance Function

$$Cov(\xi(s_i)m, \xi(s_j)) = C(\xi_i, \xi_j)$$

$$= \frac{\sigma^2}{2^{v-1}\Gamma(v)}(k||s_i - s_j||)^v K_v(k||s_i - s_j||)$$

where $||s_i - s_j||$ is Euclidean distance between pairs of locations and σ^2 is the marginal variance. K_v is the modified Besssel function of the second kind and order $v > 0$, which measures the degree of smoothness of the process. $k > 0$ is a scaling parameter related to the distance at which the spatial correlateion becomes almost null.

Temporal Autocorrelation

- evolution of epidemics, temperature, air polution
- same pricniple as spatial correlation
- improve overall data fit
- more than simple addition to the continuous spatial domain

Spatio-Temporal Modeling Structure

$$\eta_{ij} = g(\mu_{it}) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mit} + \sum_{n=1}^{N} f_k(z_{kit}) + u_{it}$$

where β_0 - intercept, β - coefficients quantifying linear effect of covariates, u_{it} spatio-temporal structure of the model, z_{kit} is the k-th explanatory variable at a given location and time, and f represents any latent model applied to covariates.

Types of spatio-temporal models

- Opportunistic spatial distribution
- Persistent spatial distribution with random intensity changes over time
- Persistent spatial distribution with temporal intensity trend
- Progressive spatio-temporal distribution

Opportunistic spatial distribution

Different spatial realizations w_t of the same spatial field for each time unit, sharing common covariance function (same k and τ) to avoid overfitting. This structure is a good approximation for processes where the spatial distribution varies considerably among different time units and unrelatedly among neighboring times.

$$u_{it} = w_{it}$$

$$w_t \sim N(0, \mathcal{Q}^{-1}(k, \tau))$$

Persistent spatial distribution with random intensity changes over time

Varying scales of intensity and a time structure is a zero mean Gaussian random noise effect (v_t). For processes where spatial component persists in time.

$$u_{it} = w_{it} + v_t$$

$$w_t \sim N(0, \mathcal{Q}^{-1}(k, \tau))$$

$$v_t \sim N(0, \tau_v^{-1})$$

Persistent spatial distribution with temporal intensity trend

Temporal progression in the mean for situations where temporal trend (h(t)) is present.

$$u_{it} = w_i + h(t)$$

$$w \sim N(0, \mathcal{Q}^{-1}(k, \tau))$$

Progressive spatio-temporal distro

Spatial realizations change in a related manner over time (!! for cases with moderate amount of variation) with r_{it} as an autoregressive temporal term for correlation among temporal neighbors of order K

$$u_{it} = w_{it} + r_{it}$$

$$w_t \sim N(0, \mathcal{Q}^{-1}(k, \tau))$$

$$r_{it} \sim N(\sum_{k=1}^{K} p_k r_{i(t-k)}, \tau_r^{-1})$$

Most common issues and additions

- *Preferential Sampling* - the sampling process that determines the data locations and the species observations are not independent
- *Spatial Misalignment* - response is observed in locations which are different from the spatial points where covariate data are available
- *Non-stationarity* - heterogeneity in space (i.e. non-stationarity) occurs when a latent global process is also affected by some underlying local processes. Typically handled via GWR
- *Imperfect detection* (environmental conditions)
- *Excess of zeroes*

Questions?

References

Martínez-Minaya, J., Cameletti, M., Conesa, D., & Pennino, M. G. (2018). Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stochastic environmental research and risk assessment, 32(11), 3227-3244.