SDM

GeoComput & ML

05 May 2022

ReCap

- Joint Distr. : [A,B]
- Cond. Prob. : [A|B] = [A,B]/[B]
- LTP : $[B] = \sum [B|A_i][A_i]$
- Bayes : $[B_i|A] = \dfrac{[A|B_i][B_i]}{\sum [A|B_i][B_i]} = \dfrac{[B_i, A]}{[A]}$

Spat. Stat.

Set Up

Z(s)

where location s vary over the index set $D \subset \mathbb{R}^d$

- random
- dependency

Stationarity

$$E(Z(s+h) - Z(s))$$

$$V(Z(s+h) - Z(s)) = 2\gamma(h)$$

$2\gamma(h)$: variogram

Process

$$[Y, \theta | Z] = \frac{[Z | Y, \theta][Y | \theta][\theta]}{[Z]}$$

Data update

$$(Z^{(1)}, Z^{(2)})$$

$$[Z^{(1)}|Y,\theta][Y|\theta][\theta]$$

$$[Y,\theta|Z^{(1)}, Z^{(2)}] = \frac{[Z^{(1)}, Z^{(2)}|Y,\theta][Y,\theta]}{Z^{(1)}, Z^{(2)}}$$

$$= \frac{[Z^{(2)}|Z^{(1)}, Y,\theta][Y,\theta|Z^{(1)}]}{Z^{(2)} Z^{(1)}}$$

Process Decomposition

$$[\cap_{i=1}^{T} Y_i] = [Y_1] \prod_{i=1}^{T-k} \left[Y_{T-k+1} | \cap_{i=1}^{T-k} Y_i\right], \quad k \in [1, T-1]$$

1st order Markov Assumption

$$[Y_T | \cap_{i=1}^{T-1} Y_i] = [Y_T | Y_{T-1}]$$

Therefore :

$$[\cap Y_i] = [Y_1] \prod_{i=2}^{T} [Y_i | Y_{i-1}]$$

Cautions

$$[Y_2|Y_1]$$

$Y = (Y_1, Y_2, F)$, where F is the level of spat. aggregation

$$[Y] = [Y_2, Y_1|F][F] = [Y_2|Y_1, F][Y_1, F][F]$$

Cautions

likelihood inference $[Z|\theta]$

missing the fundamental importance of Y which is rooted in physics, chemistry, biology, economics, etc.

Cautions

$$[\tilde{Z}] = f(Z) \approx \langle Z \rangle$$

$$[Y|\tilde{Z}, \theta_D, \theta_p] \propto [\tilde{Z}|Y, \theta_D][Y|\theta_p]$$

$$[\tilde{Z}|\theta_D, \theta_p] = \int [\tilde{Z}|Y, \theta_D][Y|\theta_p]dY$$

SDM

Domain Knowledge

> *Statisticians can not evade their responsibilities for understanding the processes they apply or recommend.*
>
> *– Ronald Fisher*

Community Ecology

Whittaker : Multiplicative Law of Diversity

$$\gamma = \alpha\,\beta$$

- γ : tot. # species
- α : ave. # species in a locality
- β : variation in the species composition between locations

Community Ecology

- organismic : traits determining persistence
- individualistic : species-specific response to env.

Community Ecology

Niche Theory

- Huntchinson, 1960s
- n-d hypervolume, resources and env.
- hard to define, species interaction w/ env.

Community Ecology

Equilibrium Theory of Island Biogeography

- MacArther and Wilson, 1960s
- # species determined by the balance of immigration and extinction

Community Ecology

Unified Neutral Theory of Biodiversity and Biogeography

- Hubbell, 2000s
- diversity rising and organised at random
- all individuals ecologically identical
- stochastic random process

Community Ecology

Theory of Ecological Community

- selection : fitness
- drift : random
- dispersal : movement
- speciation : variation

SDM

types

- correlative
- mechanistic

SDM

- tendency

predictors $\uparrow \rightarrow$ model complexity $\uparrow \rightarrow$ overfitting $\uparrow$

- right-way

var. selection based on ecol. relevance

Linear Models

$$\langle \boldsymbol{Z} \rangle = \boldsymbol{\beta}\,\boldsymbol{X},\ \boldsymbol{Z} \sim (\boldsymbol{\beta}\,\boldsymbol{X}, \sigma^2 \boldsymbol{I})$$

Linear Models

GLM : $g(f(\boldsymbol{Z})) = \boldsymbol{\beta}\,\boldsymbol{X}$

link function $g(\cdot)$

- identity
- logit : $log\left(\frac{p}{1-p}\right) \Rightarrow \text{logit}^{-1}(p) = \frac{1}{1+e^{-x}}$
- log

Linear Models

Mixed models

$$Z_i = \sum X_i \beta_i + \alpha_{p(i)} + \epsilon_i$$

- $r_i = \epsilon_i + \alpha_{p(i)}$
- $C(i,i) = \sigma_p^2; C(i,j) = 0$

Presence-Absence Data

Challenges

- detectability
- sample bias

Presence-Absence Data

MaxEnt

$$\pi(s)$$

$$[Z=1|s] = \frac{[s|Z=1][Z=1]}{[s]}$$

Jaynes principle : max entropy

$$H(x) = -\sum \pi(x) ln \pi(x)$$

Presence-Absence Data

Feature space : $\{f_1, ..., f_n\}$

$$\hat{\pi}(x) = \sum \pi(s) f_j(x)$$

$$\tilde{\pi}(x) = \langle f_j \rangle$$

$$\hat{\pi}(x) \sim q_\lambda(s) = \exp(\lambda f(x))/Z_\lambda$$

$min(-ln(q_\lambda))$ with regularisation

Presence-Absence Data

Pseudo absence data

- random
- random with exclusion : weighted by env. or geo.

core question : spat. extent

Joint SDM

$$Z = XB$$

Acknowledgement

Thanks for Your Attention

References

- J. Franklin and J. Miller, Mapping Species Distributions, 2010
- O. Ovaskainen and N. Abrego, Joint Species Distribution Modelling w R, 2020
- N. Cressie and C. K. Wikel, Statistics for Spatio-Temporal Data, 2011
- N. Cressie, Statistics for Spatial Data, 1991
- R. Larsen and M. Marx, An introduction to mathematical statistics and its applications, 2018
- A. Agresti, Foundations of Linear and Generalized Linear Models, 2015
- C. McMulloch, et.al., Generalised, Linear and Mixed Models, 2008
- https://www.sciencedirect.com/science/article/pii/S030438000500267X
- https://onlinelibrary.wiley.com/doi/10.1111/j.0906-7590.2008.5203.x
- https://onlinelibrary.wiley.com/doi/full/10.1111/j.1472-4642.2010.00725.x
- https://onlinelibrary.wiley.com/doi/full/10.1111/j.1600-0587.2013.07872.x
- https://besjournals.onlinelibrary.wiley.com/doi/10.1111/j.2041-210X.2011.00172.x
- https://www.sciencedirect.com/science/article/pii/S0304380008005486