Point Pattern Modeling for Degraded Presence-Only Data over Large Regions

Avishek Chakraborty

Joint Work with
A. Gelfand, J. Silander, A. Latimer & A. Wilson

April 7, 2010

Nature of “data”

- In large scale field experiments, difficult to keep records of the entire sampling.
- Instead, only information related to some event(s) of interest are stored
- For example, in ecological surveys, often we
  - save only the location where one or more species of interest were observed.
  - No clue about ”non-occurrence” or absence of events from these records,
- Like success-only realizations of bernoulli trial.

The problem

- For cell i, we are given only $n_i =$ number of sites where a presence was observed
- Usually presence-only data is opportunistic, informative on overall extent of sampling of the region, thus absences would not be known
- Current approaches includes putting "some" absence locations in the cell.
- Ward *et. al.*(2009) : model absence as missing data and do an EM application, under the assumption of known probability of marginal presence

A different question

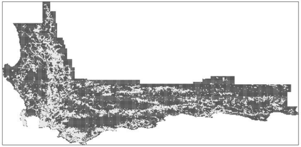
- Try to approach from a different direction, possible to estimate their distribution over the landscape
- Can compare which regions are more likely to have the species, so its relative rather than absolute
- Phillips *et. al.*(2006) proposed a optimization routine to to choose the distribution with maximum entropy constrained on the feature information
- Output of this “MaxEnt” method a single probability (*not* intensity) vector, has no measure of uncertainty attached

Biased sampling

- Different sources of bias involved in the data :
  - Don't visit the cells where species is unlikely to occur
  - Places difficult to access (interiors of reserve, mountaneous regions etc.)
  - Can miss a presence, depending on density of growth, time of the year etc.
  - Places highly sampled are close to roads or towns
- These biases often act in conflicting directions
- Extent of bias vary from one region to another, so not possible to use training data from other regions.

CFR sampling

- CFR data :: Sites within 10158 cells; around 28% of whole CFR
- Even within a cell, number of sampled sites vary a lot – negates the assumption in Ward *et. al.*(2009)
- Samples are spatially biased and "randomness" assumption is over simplifying
- covariates like town distance, road distance, presence of reserve etc., that may influence chance of sampling



Point process approach

- Instead of viewing it as a "success-only" realization of a binary distribution, we like to
  - model the set of presence locations as one "outcome" of a point process
  - think of a intensity surface which controls where the presences are likely to be found and in what frequency
- Also addressed in Warton and Shepherd (2010) – linked to the pseudo-absence model.
- In a hierarchical structure, the intensity function can be modeled with a Gaussian process depending on set of covariates
- Huge amount of data forces us to use techniques for handling the "big N" problem for Gaussian processes

High Dimensional Spatial Data

- For areal units Markov Random Field(MRF) is used where spatial parameters are updated sequentially from their full conditionals
- Although each update is simple, sequential generation makes the algorithm slow as well as highly correlated.
- However, MRF has no predictive property, inefficient for significantly large prediction set.
- For point referenced data, one generally use gaussian Process.
- In a large region, areal units can often be treated as “points”, ignoring within cell heterogeneity.
- To deal with high dimensional GP, used Bias Corrected Predictive Process as in Finley et. al (2009)

Point Process Model

- Let $\lambda(\cdot)$ be the intensity function governing the distribution of presence locations within CFR
- With cell level covariate information, not possible to get point level variations in $\lambda(\cdot)$
- Instead assume a tiled surface $\lambda(s) \equiv \lambda_i$ if $s \in$ cell i
- So homogeneous Poisson process inside each block, nonhomogeneity across blocks
- In the second stage of hierarchy, model $\lambda_{1,2,...,k} \sim GP(0, \sigma^2 R(\phi, t_1, t_2, ..., t_k))$, where t's are the cell centers

Degradation

- Degradation from actual intensity due to land use and/or sampling bias
- In cell i, $u_i =$ (known) proportion of land available for species prevalence
- Only p_i proportion area of the cell was searched for the species
- Observed intensity at cell i $= \lambda_i u_i p_i$

Degradation (contd.)

- Ideally like to augment a level for p_i using factors influencing sampling
- Such a model not well defined due to unidentifiability of $\lambda_i p_i$ upto scale.
- used a tiled surfaced, with 0 or 1 within each cell.

Species Richness

- $r(A)$ = Richness of a set $A \equiv =$ No of distinct species present inside A.
- For L species, r(A) $= \sum_{l=1}^{L} 1(n_l(A) > 0)$
- For $A \cap B = \phi$, $r(A \cup B) \neq$ r(A) + r(B)
- $E(r(A)) = \sum_{l=1}^{L} P(n_l(A) > 0)$

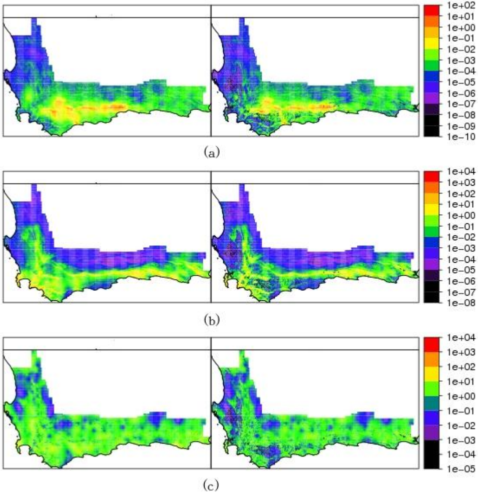
Richness (contd.)

- With the current model,
  - $E(r(A)) = \sum_{l=1}^{L}(1 - exp(-\lambda_l(A)))$
- For $A \cap B = \phi$, $\lambda(A \cup B) = \lambda(A) + \lambda(B)$
- Start with a 'base' collection of sets and infer for any A.
- More flexible than direct model

Model Specification

$$\begin{aligned}
y_i &\overset{ind}{\sim} Poi(\lambda_i u_i \Delta),\ i = 1, 2, ..., k \\
log(\lambda_i) &= x_i^T \beta + \tilde{w}_i + \epsilon_i^*,\ i = 1, 2, ..., k \\
\tilde{w}_{1:k} &= C(\phi) C^{*-1}(\phi) w_{1:m}^* \\
w_{1:m}^* &\sim N(0, \sigma^2 R(\phi, s_1^*, s_2^*, ..., s_m^*)) \\
\epsilon^* &\sim N(0, \sigma^2 Diag(C_b(\phi) - C(\phi) C^{*-1}(\phi) C^T(\phi))) \\
(\beta, \sigma^2, \phi) &\sim \pi(\beta)\pi(\sigma^2)\pi(\phi)
\end{aligned}$$

Intensity surfaces

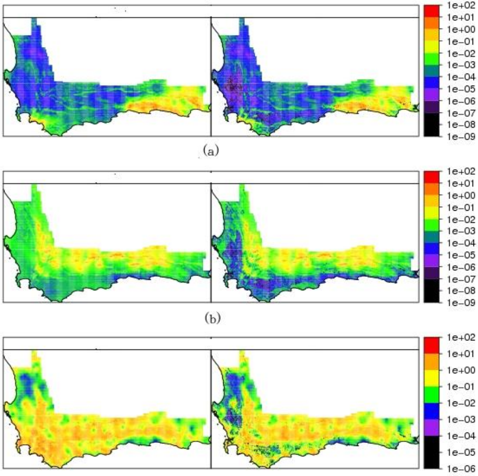


(a)

(b)

(c)

Intensity surfaces

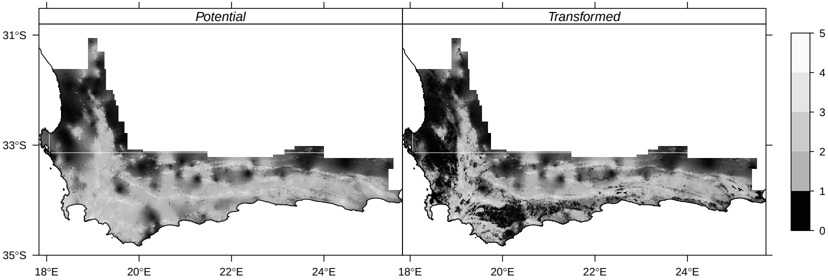


(a)

(b)

(c)

Richness surfaces



Comparison

Compared MaxEnt and our method for sensitivity to the omission of variables, for a synthetic dataset with three covariates, under two different loss functions and biased/unbiased sampling

Variable subset	Exhaustive sampling				Biased sampling			
	Loss 1		*Loss 2*		*Loss 1*		*Loss 2*	
	GP	Maxent	GP	Maxent	GP	Maxent	GP	Maxent
x_1	2.428e-06	5.919e-06	4.756e-04	2.133e-03	7.725e-07	5.006e-06	1.489e-04	1.738e-03
x_2	2.637e-06	5.999e-06	5.214e-04	1.449e-03	1.017e-06	5.649e-06	2.233e-04	1.402e-03
x_3	2.633e-06	5.143e-06	5.395e-04	1.443e-03	1.335e-06	4.678e-06	2.902e-04	1.196e-03
x_2, x_3	2.548e-06	4.428e-06	5.127e-04	9.389e-04	9.029e-07	4.194e-06	2.077e-04	8.904e-04
x_1, x_3	2.304e-06	2.976e-06	4.539e-04	1.041e-03	6.554e-07	2.864e-06	1.246e-04	8.678e-04
x_1, x_2	2.296e-06	3.570e-06	4.377e-04	9.480e-04	4.690e-07	3.616e-06	1.135e-04	9.990e-04

Extensions

- Simultaneous modeling of multiple species prevalence
- Identifying clusters of potentially co-existing species
- Explore the richness map with respect to large number of species
- Associating an individual species prevalence to overall richness

References

- Finley A.O., Sang, H., Banerjee, S. and Gelfand, A.E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational statistics and Data Analysis*, **53**, 8, 2873-2884
- Phillips, S.J., Anderson, R.P. and Schapire, R.E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231-259
- Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J.R. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**,2, 554-563
- Warton, D.I. and Shepherd, L.C. (2010). Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *Annals of Applied Statistics*. in press