Bayesian Hierarchical Spatiotemporal Modeling for Citizen Science Data in Biodiversity

Jorge Sicachá Parada

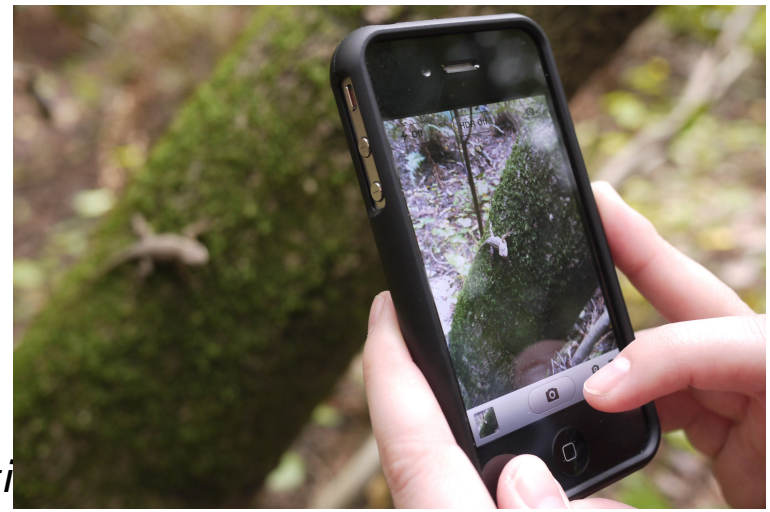
March 18, 2019

Table of contents

Citizen Science

«*Citizen science typically refers to research collaborations between scientists and volunteers, particularly (but not exclusively) to expand opportunities for scientific data collection and to provide access to scientific information for community members.*» [1]

The Cornell Lab of Ornithology
(*http://www.birds.cornell.edu/citscitoolkit/about/definiti*



[1] The Cornell Lab of Ornithology
(*http://www.birds.cornell.edu/citscitoolkit/about/definition*)

Citizen Science in Ecology

Several CS projects have been released in order to contribute to biodiversity conservation by collecting data about the location and abundance of several species.

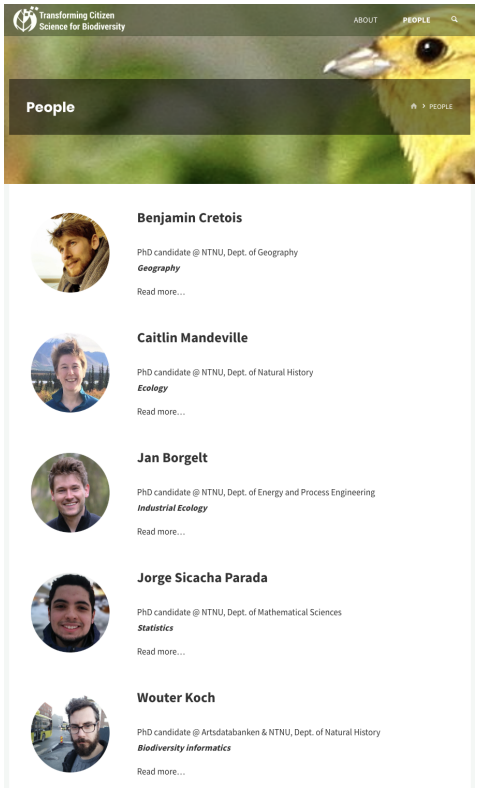
Potential of this data:

- Species Distribution Models
- Identification of critical habitats
- Study the risk of invasive species
- Protection of threatened species
- Determine the potential distribution of infectious diseases

Transforming Citizen Science for Biodiversity

- Make correct use of data collected by citizen scientists in ecology, model them, report results back to them and encourage their work.
- In order to achieve these goals TCSB integrates knowledge in citizen science, ecology, biodiversity informatics, industrial ecology, geography and statistics.
- For more information, visit **citizenscience.no**

Transforming Citizen Science for Biodiversity



Citizen Science Data

Pros

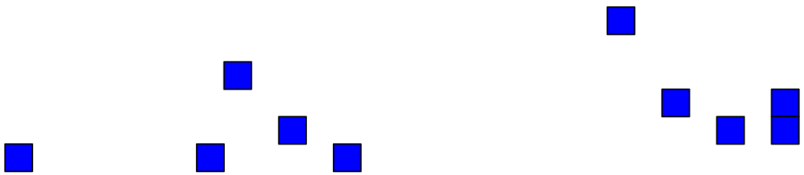
- Large volume of data
- Much easier and cheaper to collect than professional surveys
- Easily accessible (i.e. eBird, GBIF, Artsobservasjoner)

Cons

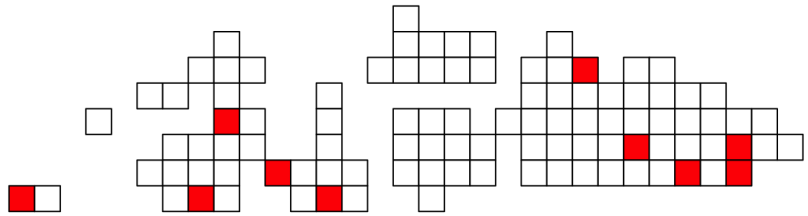
- Obtained through unsystematic sampling designs
- Their quality depend on how skilled the observer is
- Potentially biased inferences are drawn from them

Classification of Citizen Science Data

Presence-only data



Presence/absence data



Does a '1' have the same meaning regardless of the kind of data we are analyzing?

No...

Sources of bias in Citizen Science Data

According to Ruete et al (2016):

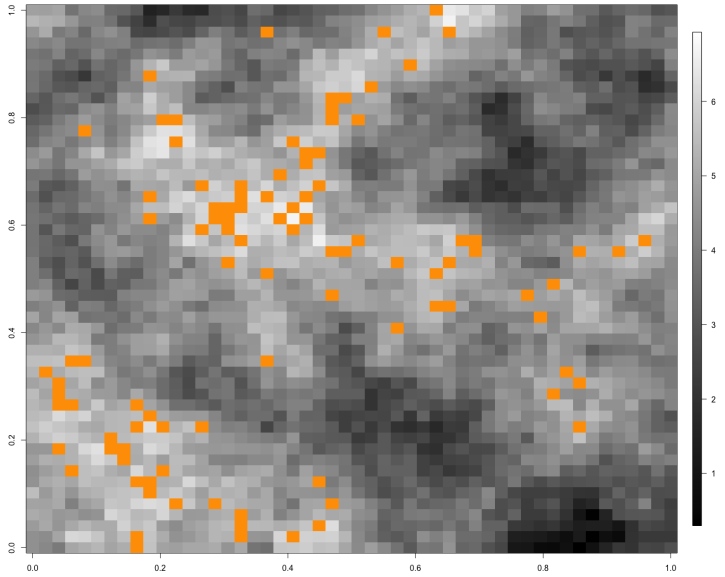
- More reports in more convenient locations - Preferential sampling
- Misclassification of observed species - Observer error
- Variation in sampling effort
- Preference for sampling some species - Taxonomical bias
- More willingness to collect observations in specific moments of the year - temporal bias

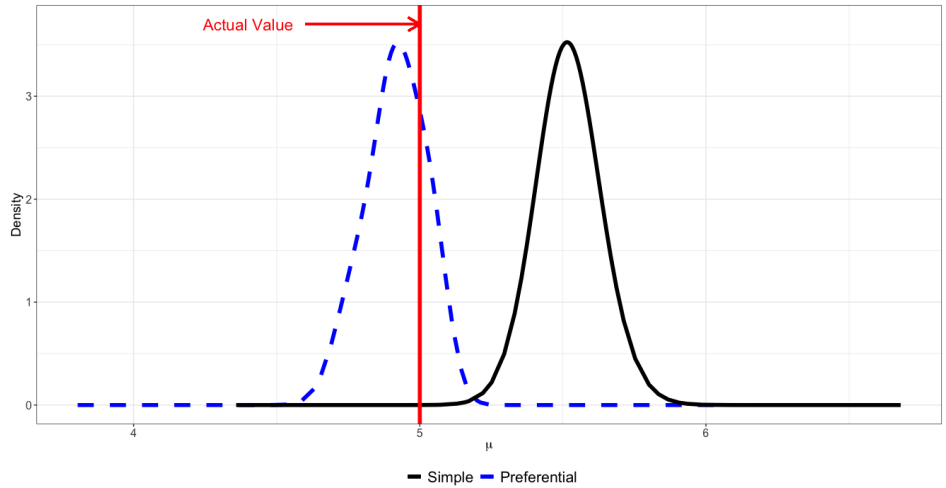
Question...

Can we characterize citizen data, their biases and their potential implication on inference, according to the sampling scheme used?

Motivating Example

- Random field with exponential covariance function with N=2500
- n=100 locations preferentially sampled (Orange cells)
- Two scenarios were considered
  - "Naïve" approach
  - Accounting for preferential sampling





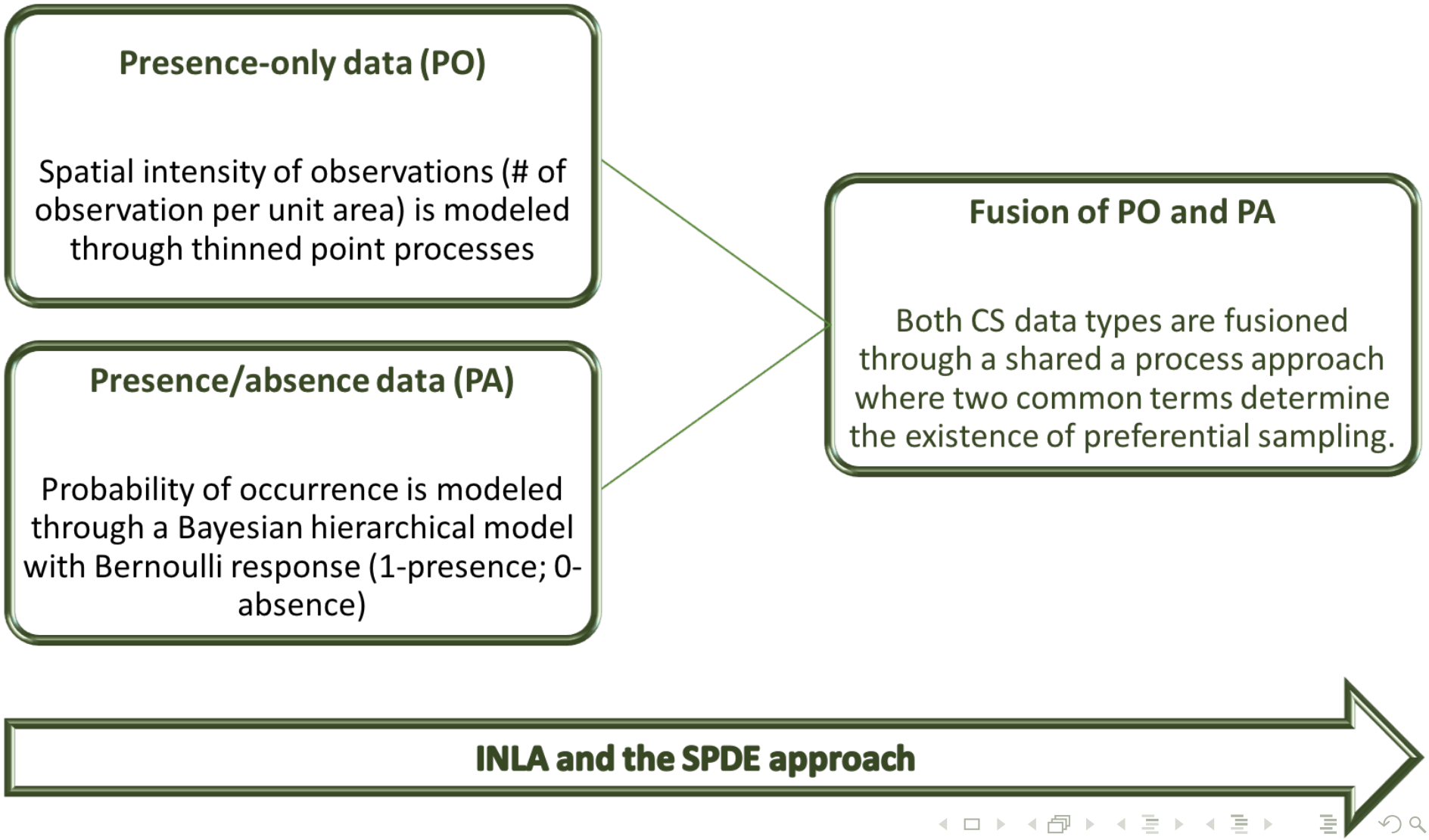
How to address this problem?: Existing approaches

According to Martinez-Minaya et al (2018):

1. Maximum entropy algorithm, MAXENT
2. Approaches based on General Additive Models

How to address this problem?

Approach based on Shirota and Gelfand (2018)



How to address this problem?

- **Presence-only data**
  - Point pattern subject to degradaton
  - Multiple sources of degradation (e.g. variable sampling effort, land transformation, etc.)

Assume $Y_{PO}(\mathbf{s}) \sim Poisson(\lambda_{PO}(\mathbf{s}))$. $\lambda_{PO}(\mathbf{s})$ is modeled as a LogGaussian Cox Process:

$$log\{\lambda_{PO}(\mathbf{s})\} = \mathbf{w}^T(\mathbf{s})\beta + \omega_{PO}(\mathbf{s}) \quad (1)$$

with $\omega_{PO}(\mathbf{s})$ a Gaussian Process.

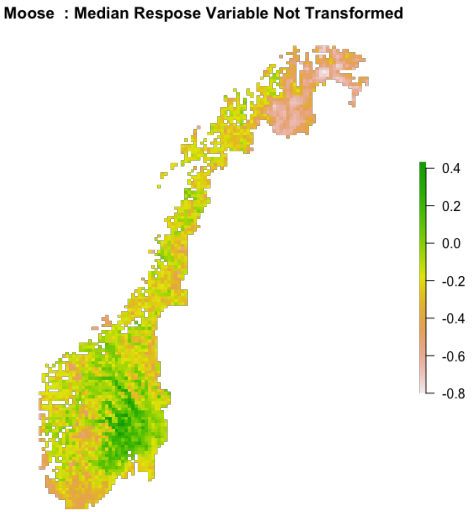
Accounting for bias...

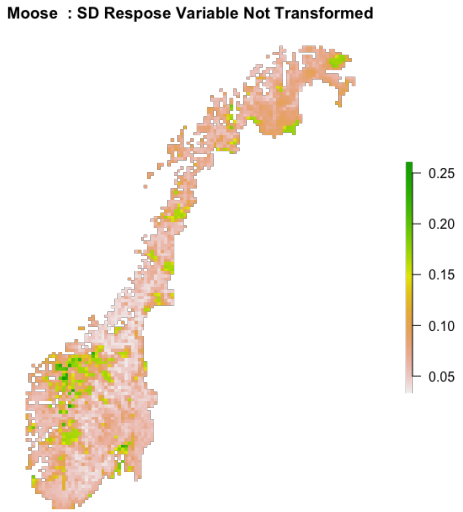
Thinned point process with intensity $\lambda_{PO}(\mathbf{s})q(\mathbf{s})$, where:

$q(\mathbf{s})$: Probability of degradation at location $\mathbf{s}$.

Ongoing research

Potential intensity of 13 ungulates in Norway, based on Citizen Science data.





Log-Gaussian Cox process considering as covariates

- Snow coverage duration
- Climatic factors
- Terrain ruggedness
- Human footprint

How to address this problem?

- **Presence/absence data**

$$Y(\mathbf{s}) \sim Bernoulli(p(\mathbf{s}))$$

with $p(\mathbf{s})$ the probability of occurrence at location $\mathbf{s}$. Under a probit link, $Y(\mathbf{s}) = 1(Z(\mathbf{s}) > 0)$ with $Z(\mathbf{s}) > 0$ a Gaussian latent process. It can be modeled as:

$$Z(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\alpha} + \phi_{PA}\omega_{PA}(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{2}$$

Accounting for bias...

$$log\{\lambda_{PA}(\mathbf{s})\} = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta_{PA}} + \omega_{PA}(\mathbf{s}) \tag{3}$$

Here, if $\phi_{PA} = 0$, then we have non-preferential sampling. Otherwise, we have it, Diggle et al (2010).

How to address this problem?

- **Fusioning presence/absence and presence-only data**
  Both types of data are fusioned based on the "shared process" perspective, Diggle and Milne (1983). For example:

$$Z(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\boldsymbol{\alpha} + \phi_{PA}\omega_{PA}(\mathbf{s}) + \phi_{PO}\omega_{PO}(\mathbf{s}) + \epsilon(\mathbf{s}) \quad (4)$$

  while the intensity of both point patterns can be modeled as in (1) and (3):

$$log\{\lambda_{PA}(\mathbf{s})\} = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta_{PA}} + \omega_{PA}(\mathbf{s})$$

$$log\{\lambda_{PO}(\mathbf{s})\} = \mathbf{w}^T(\mathbf{s})\boldsymbol{\beta_{PO}} + \omega_{PO}(\mathbf{s})$$

ϕ_{PO} has the same interpretation as ϕ_{PA}

Key questions for future papers

- How to incorporate temporal bias to this model approach?
- Based on models that account for multiple sources of bias, is it possible to model interaction between species or sets of species? How could it impact conservation management?
- How could we highlight locations where additional sampling effort is needed?

Course plan

- Fall 2018
  - MA8704 Probability Theory and Asymptotic Methods (7.5 ECTS)
  - GEOG8523 GIS Data Capture and Mapping 2 (10 ECTS)
- Spring 2019
  - MA8701 General Statistical Methods (7.5 ECTS)
- Fall 2019
  - Project course or Ethics and Communication

References


Shirota, S., & Gelfan, A. (2018)
Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data.
arXiv:1611.08719 [stat.AP]


Diggle, P., Menezes, R. & Su, T. (2010)
Geostatistical inference under preferential sampling.
Journal of the Royal Statistical Society: Series C (Applied Statistics) 59: 191-232.


Diggle, P. & Milne R. (1983)
Bivariate Cox Processes: Some Models for Bivariate Spatial Point Patterns.
Journal of the Royal Statistical Society. Series B 45(1), 11-21.


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


Ruete, A., Pärt, T., Berg, Å., Knape, J. (2016)
Exploiting opportunistic observations to estimate changes in seasonal site use: An example with wetland birds.
Ecology and Evolution 7:5632– 5644..