Point process models for presence-only analysis

Ian Renner, Jane Elith, Adrian Baddeley, William Fithian, Trevor Hastie, Steven J. Phillips, Gordana Popovic, and David I. Warton

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

December 2, 2015

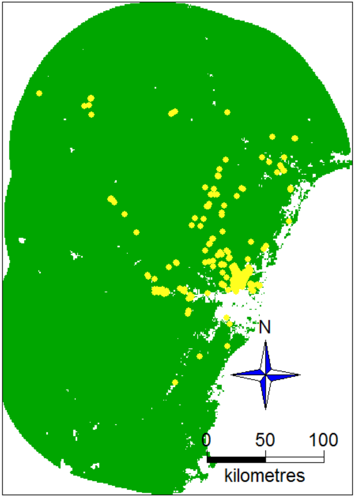
Outline

- Background (Species Distribution Models)
- Point Process Models
- Advances in Presence-Only Analysis
- Extensions of PPMs
- Future Work

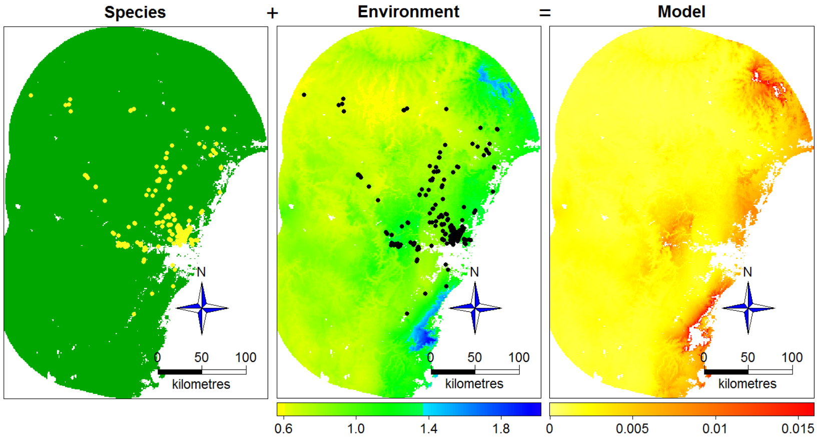
Presence-only Data

e.g. Reported locations of *Eucalyptus sparsifolia* in the Blue Mountains

Eucalyptus sparsifolia in the Blue Mountains



Species Distribution Modelling



Poisson point process models

Starting point: inhomogeneous Poisson point process model with intensity $\mu(s)$ defined over region $\mathcal{A}$, which assumes:

- Point locations $\mathbf{s}_P$ are independently distributed, conditional on environment
- Number of points m is a realisation of a Poisson random variable with mean $\int_{s\in\mathcal{A}} \mu(s)ds$

Intensity modelled as a log-linear function of environmental variables:

$$\ln \mu(s) = \beta_0 + \beta_1 \times \text{rain}(s) + \beta_2 \times \text{temp}(s) + \dots$$

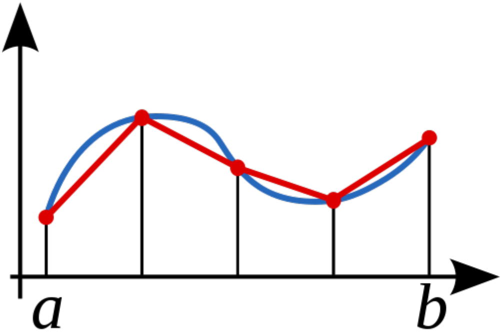
Maximise **log-likelihood** (using GLM software):

$$l(\boldsymbol{\beta}; \mathbf{s}_P) = \sum_{i=1}^{m} \ln \mu(s_i) - \int_{s\in\mathcal{A}} \mu(s)ds$$

Numerical Integration

$\int_{s \in \mathcal{A}} \mu(s) ds \approx \sum_{i=1}^{n} w_i \mu(s_i)$,

where $\mathbf{w} = \{w_1, \ldots, w_n\}$ are quadrature weights and $\mathbf{s}_0 = \{s_{m+1}, \ldots, s_n\}$ are quadrature points.



What is the intensity measuring?

Intensity is not a probability, but is related to abundance, but abundance of what?

What we want:

What we get:

Equivalence Results

Over the past 5 years, point process models have been linked to many other methods for fitting SDMs to presence-only data, *e.g.*:

Poisson point process models (ignoring weights) are equivalent to pseudo-absence logistic regression* and MAXENT†.

This links Poisson point process models to the two most common approaches to presence-only SDM!

Consequence of ignoring weights: Pseudo-absence logistic regression and MAXENT are scale-dependent (predicted probability depends on number of pseudo-absences/background points)

---

*Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Annals of Applied Statistics* **4**, 1383–1402.

†Renner, I.W. & Warton, D.I. (2013) Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69**, 274–281.

Fitting a Poisson PPM

Equivalence results mean there are many ways to fit Poisson PPMs:

- MAXENT software (ignoring weights)
- R packages spatstat, ppmlasso, and dismo (R version of MAXENT)
- "Infinitely weighted logistic regression" (IWLR)[‡]

```
>up.wt = (10^6)^(1 - Pres)
>iwlr = glm(Pres ~ X.des, family = binomial(), weights = up.wt)
```

- "Downweighted Poisson regression" (DWPR)[§]

```
>p.wt = rep(1.e-6, length(Pres))
>p.wt[Pres == 0] = Area/sum(Pres == 0)
>dwpr = glm(Pres/p.wt ~ X.des, family = poisson(), weights = p.wt)
```

[‡] Fithian, W. & Hastie, T. (2013) Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics* **7**, 1917–1939.

[§] Renner, I.W. *et al.* (2015) Point process models for presence-only analysis – a review. *Methods in Ecology & Evolution* **6**, 366–379.

Software properties

Property	spatstat	ppmlasso	IWLR	DWPR	MAXENT	R-INLA	lgcp
Regularisation	×	✓	✓	✓	✓[1]	×	×
Standard errors	✓[2]	×	✓[2]	✓[2]	×	✓	✓
Variable importance plots	×	×	×	×	✓	×	×
Diagnostic plots	✓	✓	×	×	×	×	×
Spatial dependence	✓	✓	×	×	×	✓	✓
Non-linearity (eg smoothers)	✓	✓	✓	✓	✓	✓	✓
Scale-invariant	✓	✓	×	✓	✓[3]	✓	✓

1 LASSO only

2 For Poisson models only

3 Raw output only

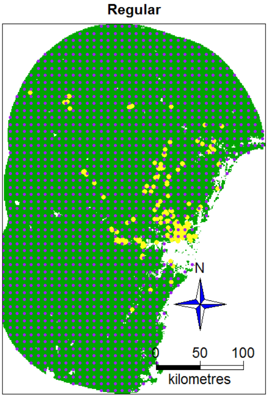
Why use PPMs?

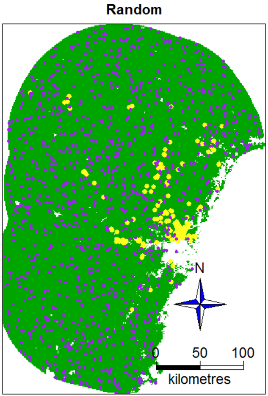
The point process model framework provides advances to presence-only SDM, including:

- Criteria for choice of pseudo-absences
- Checking assumptions
- Ecological insight
- Data-driven LASSO regularisation
- Accounting for observer bias

Choice of pseudo-absences

Most presence-only methods require pseudo-absences. But how many should be chosen? Where should they be placed?





Previous recommendations

Lots of literature on how to choose pseudo-absences:

- A fixed number (often 10,000)
- A fixed ratio of presence:pseudo-absence points
- Choose points more likely to be true absences

These recommendations are generally justified through simulation or by looking at only a few data sets.

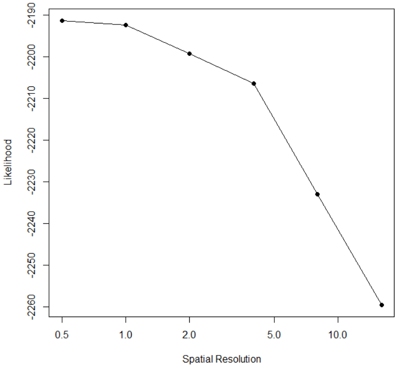
This has led to confusion about which of the (sometimes contradictory) approaches to take.

Likelihood convergence

PPM framework turns pseudo-absence choice into quadrature problem.

Regular grid: choose enough for likelihood convergence.

`findres` function in `ppmlasso`:



Random quadrature points (DWPR): standard error formula.

Required number for standard error within e: $\frac{|\mathcal{A}|^2 s^2}{e^2}$

For *E. sparisolia* using an initial fit of 10,000 random quadrature points, $s = 0.0103$, so the required number to reduce the standard error to below $e = 2$ is roughly 198,000.

Why use PPMs?

The point process model framework provides advances to presence-only SDM, including:

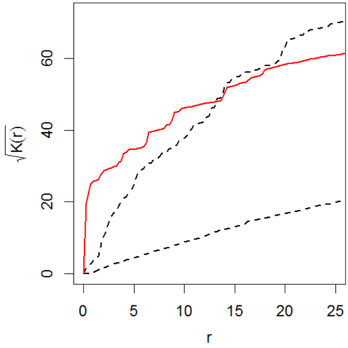
- Criteria for choice of pseudo-absences
  - Until likelihood convergence
  - Standard error formula
- Checking assumptions
- Ecological insight
- Data-driven LASSO regularisation
- Accounting for observer bias

Model diagnostics

Many SDM methods (MAXENT, P-A regression) currently have no way of checking model assumptions (particularly the independence assumption).

Lots of literature on checking assumptions for PPMs.

One check of independence assumption: K-envelope:



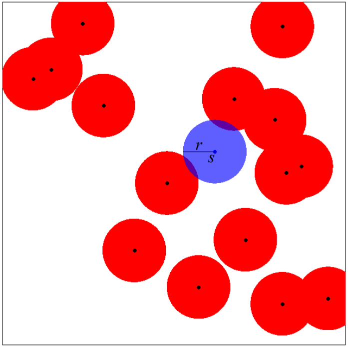
Poisson PPM (hence MAXENT/P-A regression) is not appropriate for *E. sparsifolia*!

More diagnostic (and other) tools available via `spatstat`.

Accounting for dependence: AI models

Other types of PPMs can account for point dependence:

An Area-interaction model of radius r



- fits conditional intensity at s as a log-linear function of environmental variables $\mathbf{x}(s)$ and point interaction $t_s(\mathbf{s_P})$
- $\ln \mu(s, \mathbf{s}_P) = \mathbf{x}(s)'\boldsymbol{\beta} + t_s(\mathbf{s}_P)\theta$
- available in `spatstat` and `ppmlasso`

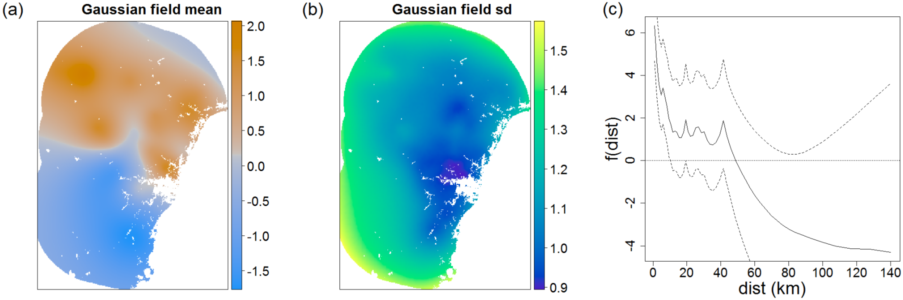
Accounting for dependence: Cox processes

A more flexible way to account for dependence.

Log-Gaussian Cox process models: intensity $\mu(s)$ modelled as a realisation of a stochastic Gaussian process $\xi(s)$:

$$\ln \mu(s) = \mathbf{x}(s)'\boldsymbol{\beta} + \xi(s)$$

Fitted via MCMC (`lgcp` package) or integrated nested Laplace approximation (`R-INLA` package)



Why use PPMs?

The point process model framework provides advances to presence-only SDM, including:

- Criteria for choice of pseudo-absences
  - Until likelihood convergence
  - Standard error formula
- Checking assumptions
  - Many diagnostic tools available
  - Alternative PPMs to account for point dependence
- Ecological insight
- Data-driven LASSO regularisation
- Accounting for observer bias

Explaining the distribution

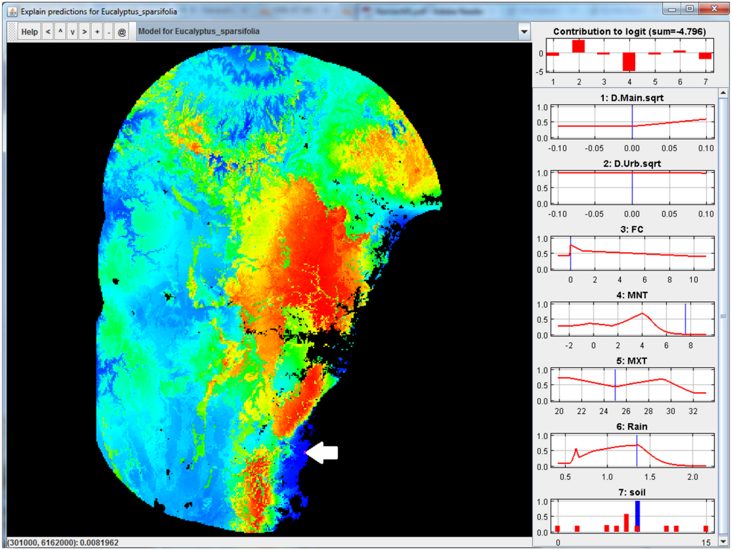
Eucalyptus sparsifolia is known to prefer "low nutrient soils, but some on medium and high nutrient soils, over a wide range of rainfall".⋆

Minimum temperature emerges as an important driver of the distribution of *Eucalyptus sparsifolia* that was previously unknown to ecologists, as evident from a significantly negative quadratic coefficient.

This variable has implications for climate change projections, suggesting a substantial decrease in *Eucalyptus sparsifolia* intensity at the southern end of its range under warming scenarios.

⋆Hager, T. & Benson, D. (2010) The Eucalypts of the Greater Blue Mountains World Heritage Area: distribution, classification and habitats of the species of Eucalyptus, Angophora and Corymbia (family Myrtaceae) recorded in its eight conservation reserves. *Cunninghamia* **10**, 425–444.

MAXENT's "explain" tool



Why use PPMs?

The point process model framework provides advances to presence-only SDM, including:

- Criteria for choice of pseudo-absences
  - Until likelihood convergence
  - Standard error formula
- Checking assumptions
  - Many diagnostic tools available
  - Alternative PPMs to account for point dependence
- Ecological insight
  - Tools to discover important environmental covariates
- Data-driven LASSO regularisation
- Accounting for observer bias

Why use PPMs?

The point process model framework provides advances to presence-only SDM, including:

- Criteria for choice of pseudo-absences
  - Until likelihood convergence
  - Standard error formula
- Checking assumptions
  - Many diagnostic tools available
  - Alternative PPMs to account for point dependence
- Ecological insight
  - Tools to discover important environmental covariates
- Data-driven LASSO regularisation
  - Various options available in `ppmlasso`
- Accounting for observer bias
  - Include covariates associated with site accessibility

Extensions of PPMs

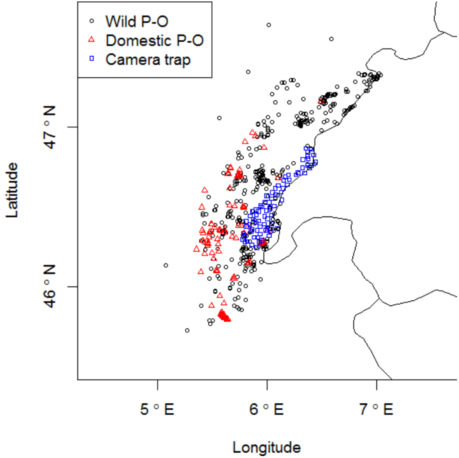
Work with Olivier Gimenez, who works in the CEFE in Montpellier, France:

- **Combined data sources**
- Dynamic SDM
- Tricky applications for brown bears and monk seals in Greece

Combined data sources

In many situations, there is more than one source of data.

Example: Lynx in the Jura Mountains in France



- Sightings in the wild (P-O)
- Domestic interferences (P-O)
- Camera traps (survey)

Combined Likelihood

Typically, people build a model using only one source of data.

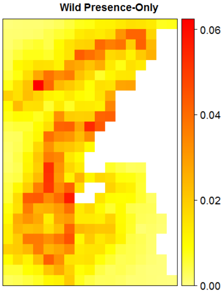
How might we build a model using multiple sources of data?

- **Presence-only and presence-absence**[▲]:
  $l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = l_{\text{PO}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}) + l_{\text{PA}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$
- **Presence-only and occupancy**[**]: $l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = l_{\text{PO}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + l_{\text{Occ}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$
- Goal for lynx:
  $l(\boldsymbol{\alpha_W}, \boldsymbol{\alpha_D}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = l_{\text{Wild PO}}(\boldsymbol{\alpha_W}, \boldsymbol{\beta}) + l_{\text{Domestic PO}}(\boldsymbol{\alpha_D}, \boldsymbol{\beta}) + l_{\text{Occ}}(\boldsymbol{\beta}, \boldsymbol{\gamma})$

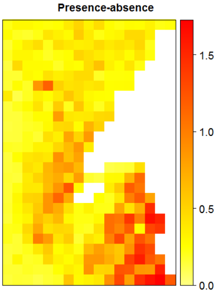
▲Fithian, W., Elith, J., Hastie, T., & Keith, D.A. (2015) Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution* **6**, 424–438.

Dorazio, R.M. (2014) Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography* **23, 1472–1484.

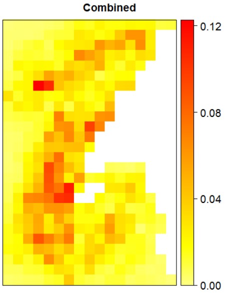
Presence-Only and Presence-Absence



μ_{PO}: intensity of reportings per unit area

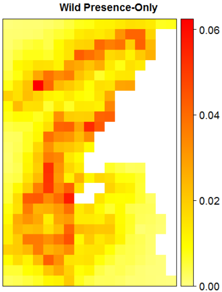


μ_{PA}: "intensity" of species per unit area

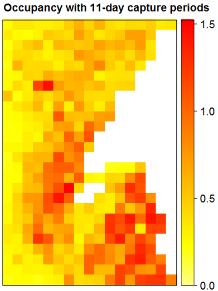


$\mu_{\mathrm{PO+PA}}$: ?

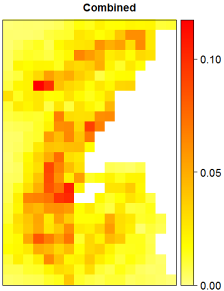
Presence-Only and Occupancy



μ_{PO}: intensity of reportings per unit area



μ_{Occ}: "intensity" of species per unit area



$\mu_{\mathrm{PO+Occ}}$: ?

Future Work

- PPMs
  - Data quality
    - Errors in covariates
    - Accuracy of location coordinates
  - Temporal aspect
    - Decades of observed locations
    - Environmental variation over observed timespan
    - Applications for telemetry, invasive species
- Combined likelihood
  - Occupancy model stability: LASSO on detection covariates?
  - Weighting: presence-only seems to dominate?
- Dynamic SDM: PPMs + HMMs
  - "Self-exciting" Poisson point processes to model wolf attack patterns?

Acknowledgements

A big thank you to:

- The organising committee for this great conference
- All of my co-authors
- All of you, for your attention!

References

- Renner, I.W. & Warton, D.I. (2013). Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. *Biometrics* **69**, 274–281.
- Renner, I.W., Baddeley, A., Elith, J., Fithian, W., Hastie, T., Phillips, S., Popovic, G. & Warton, D.I. (2015). Point process models for presence-only analysis – a review. *Methods in Ecology & Evolution* **6**, 366–379.
- Warton, D.I., Renner, I.W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS ONE* **8**, e79168.