# Is my sdm good enough? insights from a citizen science dataset in a point process modeling framework

Camila Leandro, Pierre Jay-Robert, Bruno Mériguet, Xavier Houard, Ian W. Renner

# Is my SDM good enough? Insights from a citizen science dataset in a Point Process Modeling framework

**Authors**: Camila LEANDRO, Pierre JAY-ROBERT, Bruno MERIGUET, Xavier HOUARD, Ian RENNER

*Corresponding author*: Camila LEANDRO camila.leandro@cefe.cnrs.fr; Bat. J. Dpt. BEE, Université Paul-Valéry Montpellier 3, 34199 Montpellier, FRANCE

*Affiliations:*

Camila LEANDRO [orcid.org/0000-0003-1153-326X] and Pierre JAY-ROBERT: CEFE, Univ. Paul Valéry Montpellier 3, Univ. Montpellier, EPHE, CNRS, IRD; Montpellier, France

Bruno MERIGUET and Xavier HOUARD: Office pour les Insectes et leur Environnement (Opie), BP 30, 78041 Guyancourt cedex, France

Ian RENNER: School of Mathematical and Physical Sciences, The University of Newcastle, Australia

## *Abstract and keywords*

Citizen science programs, and particularly atlas schemes based on opportunistic biological records, are very important sources of data for species distribution models and conservation. Nevertheless, these data are prone to bias, particularly when they come from less popular or hard to detect/identify species, such as insects. With such biased data, it is important to evaluate the stability of the model predictions. In recent years, point process models (PPMs) have shown their strength as a unifying framework to fit presence-only species distribution models with many advantages in model implementation and interpretation; PPMs are closely connected to methods already in widespread use in ecology such as MaxEnt and to logistic regression and benefit from being more transparent about resource selection and absence handling. Moreover, there is a well-developed set of tools to fit these models and assess various features of the underlying model, including model

stability. However, such tools are currently unavailable when point process models are fitted with a lasso penalty, which has been shown to improve predictive performance. Based on the French citizen science program “Stag beetle Quest”, we propose new methods to assess model stability in this context. The ultimate goal was to develop a set of functions to analyze PPM models with lasso penalties fitted with presence-only data. To assess model stability, we randomly sampled different subsets of locations with varying size from the whole dataset and used the proposed tools to compare fitted intensities and model coefficients. All the developed measures are complementary and can be used to identify at what number of point locations the model stabilizes, which will be dependent on the dataset. Our work presents a new toolbox to explore questions around model stability based on the number of locations in the context of point process models with a lasso penalty and confirms once more the use of the point process modelling framework as a flexible and unifying framework to fit presence-only species distribution models.

Key-words: Species distribution models; Point process models; LASSO; diagnostic tools; R functions; Lucanus cervus

# *I. Main text*

## 1. Introduction

To be able to estimate accurately the decline of biodiversity, we need to be equipped with reliable tools and methods allowing a good characterization of population trends. Methods should provide a picture of the distribution of species through space and time from data which represent a subsample of the true species populations. This is especially necessary for

47 organisms that are hard to detect in their environment, such as insects (Donaldson et al.
48 2017; Leandro et al. 2017).

49 Species distribution models (SDMs) have become important methods to inform policy
50 makers and conservation practitioners about biodiversity trends. Mapping the patterns of
51 biodiversity, SDMs can be used in land use planning, leading to prioritization of conservation
52 strategies (Devictor et al., 2010; Guisan et al. 2013). They have also been put forward as
53 pivotal tools for the appropriate evaluation of conservation status of insects (Diniz-Filho et
54 al., 2010; Cardoso et al., 2011; Leandro et al., 2017).

55 In order to fit a SDM, a substantial number of recorded locations is typically necessary. One
56 source of data that can be used to fit an SDM is a list of locations found in biodiversity atlas
57 schemes and citizen science programs, but such data involve the attendance of particular
58 questions related to the observation process (Alabri, 2010; Isaac & Pockok, 2015; Powney &
59 Isaac, 2015). Indeed, data can come in a number of formats, the two most common being:
60 (1) presence-absence data, which implies a clear sampling protocol and a greater effort from
61 the observer when cryptic species are considered and (2) presence-only data. Presence-only
62 data are cheaper and consequently more widely available than presence-absence data.
63 However, they are more prone to bias due to the way they are collected: presence-only data
64 can be opportunistic observations whose distribution is highly correlated with the
65 observation process (Warton et al. 2013; Guillera-Arroita, 2017).

66 Let us put ourselves in the place of a practitioner wanting to model the distribution of a
67 species whose observations come from citizen science with presence-only data. Let us say
68 that the ecology of the species is relatively well known. First we have to address the
69 question “Which is the best statistical framework to model my data?” This question has

been largely explored (Aguirre-Gutiérrez et al. 2013; Guillera-Arroita et al. 2015; Duque-Lazo et al. 2016) and in recent years, point process models (PPMs) have shown their strength as a unifying framework to fit presence-only species distribution models (SDMs) with many advantages in model implementation and interpretation, which can be obscured in popular software platforms such as MaxEnt (Renner et al. 2015; Stirling et al. 2016). Indeed, easy to use "click-button" platforms such as MaxEnt (Philips et al. 2017) and the Biomod R package (Thuiller et al. 2009) have been described as "black box techniques" because users can ignore the details and nuances of their models and default parameters (Renner & Warton 2013; Ahmed et al. 2015; Philips et al, 2017). Point process models, on the contrary, let the user have complete control over what its being modelled (Renner et al. 2015). In particular, PPMs provide clearer interpretations of the model output as an intensity of reported observations per unit area and as well as clarity regarding necessary choices to implement presence-only models such as the choice of quadrature points (also referred to as "pseudo-absences" or "background points").

Then comes the crucial question "do I have enough data to model the distribution of the species?" (Virgili et al. 2018), a question that is not new and which can be translated into the important matter of "trust in models" or model accuracy and particularly in their specific contexts (Stockwell & Peterson 2002; Guillera-Arroita et al. 2015; Ross et al. 2015). When fitting a point process model, we estimate the intensity of species records as a function of the chosen environmental covariates. The stability of this intensity surface depends not only on the number of records, but also on the choice of covariates used to characterize it. Indeed, reducing the number of candidate variables helps to explain which biological factors are important in determining a species' distribution. For example, MAXENT software by default uses a Lasso penalty, which shrinks parameter estimates $\widehat{\boldsymbol{\beta}}$ toward zero. While the

Lasso penalty is known to improve predictive performance and give numerical stability, the default penalty chosen by MAXENT software is *ad hoc*; the choice of the penalty criterion can have consequences in model interpretation, as reducing the number of candidate variables helps to explain which biological factors are important in determining a species' distribution, but some criteria impose larger penalties than others (i.e. BIC, MSI) (Renner, 2013).

In the point process framework, the 'spatstat' package (Baddeley & Turner, 2005; Baddeley et al., 2015) offers a number of tools to test model reliability, including significance levels for implemented variables and standard deviations of the predicted intensity. However, in spatstat, regularization tools aimed at boosting predictive performance through reducing model complexity, such as Lasso penalties, are not available. In the PPM-lasso framework of the 'ppmlasso' package, a number of Lasso-type penalties are included in order to shrink coefficients of point process models in a data-driven way, which tends to provide superior predictive performance to MAXENT (Renner & Warton, 2013). Nevertheless, there are no tools to explore model stability within the PPM-lasso framework.

Our goal was to develop a toolbox analogous to that of the 'spatstat' package, therefore writing new functions to explore model stability for models fitted with the ppmlasso package that would expand the toolkit for practitioners and researchers who want to have more control over their models. Based on the French citizen science program "En quête d'insectes ! Lucane cerf-volant" or "Stag beetle Quest", we explored different methods to assess model stability (or the capacity to predict correctly all presence data) within the PPM perspective fitted with a lasso penalty and observer bias corrections. Thanks to the extensive dataset offered by this dynamic program, we used random subsets of increasing size to test the

stability of models fitted with varying numbers of points in order to determine whether the model fitted with all points could be considered to have stabilized. Such methods will contribute to an increase in the usage of SDM for a wider audience of practitioners as we provide a toolbox of different R functions which may be used to explore stability of models fitted with the ppmlasso package. We present a detailed tutorial as supplementary material demonstrating usage of these functions and interpretation of their output. By doing so, we also conducted an ecological analysis of the distribution of *Lucanus cervus* in France.

## 2. Materials & Methods

### 2.1 Data

Species records were obtained from the Stag beetle Quest citizen science program launched in 2011 and managed by the Office for the Insects and their Environments (Opie) (Meriguet *et al.*, 2012). The program is focused on the French distribution of *Lucanus cervus* (Linnaeus, 1758) (Insecta, Coleoptera) and contains more than 16,000 records from 1905 onward. Data from before 2011 come from contributors who entered old records through the Stag beetle Quest online form. The database is composed of ~90 % presence-only data of which ~ 82 % of the records have a precise location.

The data retained for the study correspond to a recent and highly active period of observation (from 2007 to 2017) (Fig. 1a), thereby reducing the temporal heterogeneity of the dataset. Only verified observations (photography-based validation made by experts) were used, leaving a total of 2576 point locations.

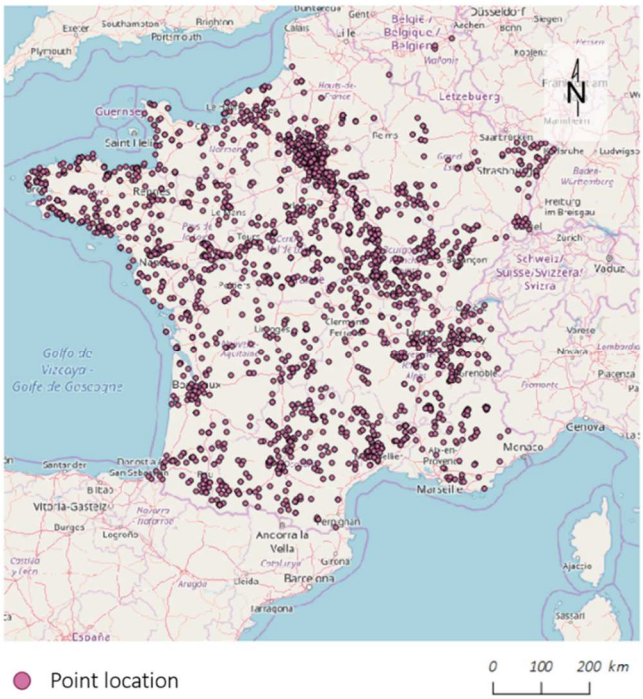



Figure 1. Point locations of the data used for the analysis. Map made by © OpenStreetMap contributors

Saproxilic beetles are species which are involved in or dependent on wood decay; in some European forests, the Lucanidae family presents the highest percentage of indicator species for dead-wood amount and temperature (Lachat et al. 2012). Indeed, like other exothermic insects, their life traits and abundance is related to climatic variables; additionally, as adult activity has been considered as weather-dependent (Fremlin & Fremlin 2010), we hypothesized that their sightings (observations) would be as well. Therefore, to model the distribution of *Lucanus cervus*, we used six environmental variables : 2 climate variables from WorldClim (Hijmans et al. 2005; Fick & Hijmans 2017) and four land use variables from the Corine Land Cover (2012) and Hilda databases (Fuch et al. 2013-2014-2015) (Table 1); Climatic variables were modelled with linear, interaction and quadratic terms, while land use (defined as proportion of the landscape cover within grid cells) and observer bias variables were entered as linear terms leading to a total of 10 covariates.

Variables were chosen based on the literature (Thomaes et al. 2008; Hawes 2008; Irmler *et al.* 2010; Frank et al. 2017) and our expertise, and verified if in the suite of variables no two variables have a Pearson correlation R >= 0.7. Because presence-only data are prone to observer bias, in which the observed pattern of points reflects not only the distribution of the species but also the distribution of the observers, we added an "observer bias" variable. It is common to use distances to roads or to natural areas for this purpose (Renner et al. 2015; Fisher-Phelps et al. 2017), nevertheless in our particular case, points seemed to be clustered around cities, which led us to include the natural logarithm of human population density as an observer bias variable, assuming that the species was reported more when human population density was higher. Therefore we included the human population variable from the SEDAC dataset (2016). All variables were available at 1 km x 1 km resolution (Table 1).

Table 1. Complete information of the environmental and bias variables included in the model. The climatic and human population variables' resolutions are 30 arc-seconds (approximately 1 km at the equator).

| | Type | Model form | Covariate | Source |
|---|---|---|---|---|
| **Environmental covariates** | Climatic | Linear, Quadratic | Mean annual Temperature (Bio 1) *Unit: Celsius degrees (°C)* | Bioclimatic variables from Worldclim (2017) Resolution ~1 km² |
| | | Linear, Quadratic | Mean annual Precipitation rate (Bio12) *Unit: millimeter (mm)* | |
| | | Interaction term | Mean annual temperature * Mean annual precipitation rate | |
| | Land Use | Linear | Percentage of broad-leaved forest cover in a 1km radius | Corine Land Cover (2012) Resolution 1 km² |
| | | Linear | Percentage of coniferous forest cover in a 1km radius | |
| | | Linear | Percentage of arable land cover in a 1km radius | |
| | | Linear | Percentage of forest cover in the past (1910 and 1960) at 1km radius | Hilda database (2013) Resolution 1 km² |

| | | | | |
|---|---|---|---|---|
| **Observer bias covariate** | Uneven sampling effort | Linear | Natural logarithm of the Human Population data | CIESIN gridded population of the world<br>SEDAC dataset (2016)<br>Resolution ~1 km² |

## 2.2 SDM Framework

The $m$ = 2576 stag beetle locations, denoted by $\boldsymbol{s}$, were modelled with a Poisson point process model. Under this model, we assume that the expected number of stag beetle presence reportings per unit area, called the intensity $\mu(s)$, varies spatially (therefore indexed by location $s$), according to environmental conditions $\boldsymbol{x}(s)$ and a term related to the observation process $z(s)$. Ecologically speaking, this intensity is not a probability of occurrence but a measure proportional to the abundance per unit area for the considered species (Renner *et al.* 2015) throughout an area $\mathcal{A}$. In our case, the intensity of points was fitted as a log-linear model of the predictors (Warton *et al.* 2013; Renner *et al.* 2015). Such predictors were split into two categories: environmental variables $\boldsymbol{x}(s)$ parameterized by $\boldsymbol{\beta}$ and the observer bias variable $z(s)$ parameterized by gamma ($\gamma$) (eq. 1).

Equation 1: $$\ln \mu(s) = \boldsymbol{x}(s)'\boldsymbol{\beta} + z(s)'\gamma$$

The parameters of the model are typically fitted via maximizing the log-likelihood expression below (eq. 2) (Cressie, 1993) which includes an intractable integral $\mu_{\mathcal{A}}$.

Equation 2: $$l(\boldsymbol{\beta}, \gamma; \boldsymbol{s}) = \sum_{i=i}^{m} ln\ \mu(s_i) - \mu_{\mathcal{A}}, \text{ where } \mu_{\mathcal{A}} = \int_{s\,\epsilon\,\mathcal{A}} \mu(s)\, ds$$

Because the integral $\mu_{\mathcal{A}}$ is intractable, it must be approximated via numerical quadrature (eq. 3).

Equation 3: $$\mu_{\mathcal{A}} \approx \sum_{i=m+1}^{m+n} w_i\, \mu(s_i)$$

This is done by introducing a set of $n$ quadrature points $\boldsymbol{s}_0 = \{s_{m+1}, \ldots, s_{m+n}\}$ throughout $\mathcal{A}$ along a regular grid and associating with the species locations $\boldsymbol{s}$ and the quadrature points

189 $\boldsymbol{s}_0$ a set of quadrature weights $\boldsymbol{w} = \{w_1, \dots, w_{m+n}\}$, leading to the approximate likelihood
190 below (eq. 4; Berman & Turner 1992).

191 Equation 4: $l(\boldsymbol{\beta}, \gamma; \boldsymbol{s}) \approx \sum_{i=i}^{m+n} w_i(y_i \ln(\mu(s_i)) - \mu(s_i))$

192 In equation 4, $y_i = \frac{\mathrm{I}(i \in \{i,\dots,m\})}{w_i}$; in other words, $y_i$ is equal to 1 over the quadrature
193 weight if $s_i$ is one of the presence points and 0 if $s_i$ is one of the quadrature points.

194 Quadrature points were initially placed on a regular 1 km x 1 km grid. However, initial
195 analysis of the data with the findres function of ppmlasso suggested that we did not need to
196 fit models at such a fine resolution, as the maximized log-likelihood appeared to stabilize at
197 a spatial resolution of 4 km x 4 km (see Appendix S1 in Supporting Information, Fig. S1.1),
198 which we hereby used in all of our models in order to reduce the time and computer power
199 needed to run the analysis, improving the efficiency of the analysis (Renner & Warton,
200 2013).

201 In our case, we used 10 covariates to model the observed pattern of stag beetle locations.
202 With so many covariates, we run the risk of overfitting the model as some may not be
203 informative of the distribution of the observed records. Therefore, we incorporated a lasso
204 penalty (Tibshirani 1996), which shrinks coefficients toward zero and in some cases may set
205 some coefficients to be exactly zero, effectively removing the associated covariates from the
206 model (Renner et al., in press). We fitted regularization paths of 200 Poisson PPMs with
207 increasing lasso penalties, and chose the model with the smallest model selection criterion,
208 here BIC.

209 Analyses were performed in R 4.0.2 (R Development Core Team 2020) using the 'ppmlasso'
210 package (Renner & Warton, 2013) and different R functions which were written to establish

intensity and coefficient measures. These functions are the stability assessment toolbox, hereafter referred to as “diagnostic tools”. Code, simulated data and a tutorial illustrating use of this code are provided in the supplementary material.

### 2.3 Diagnostic tools

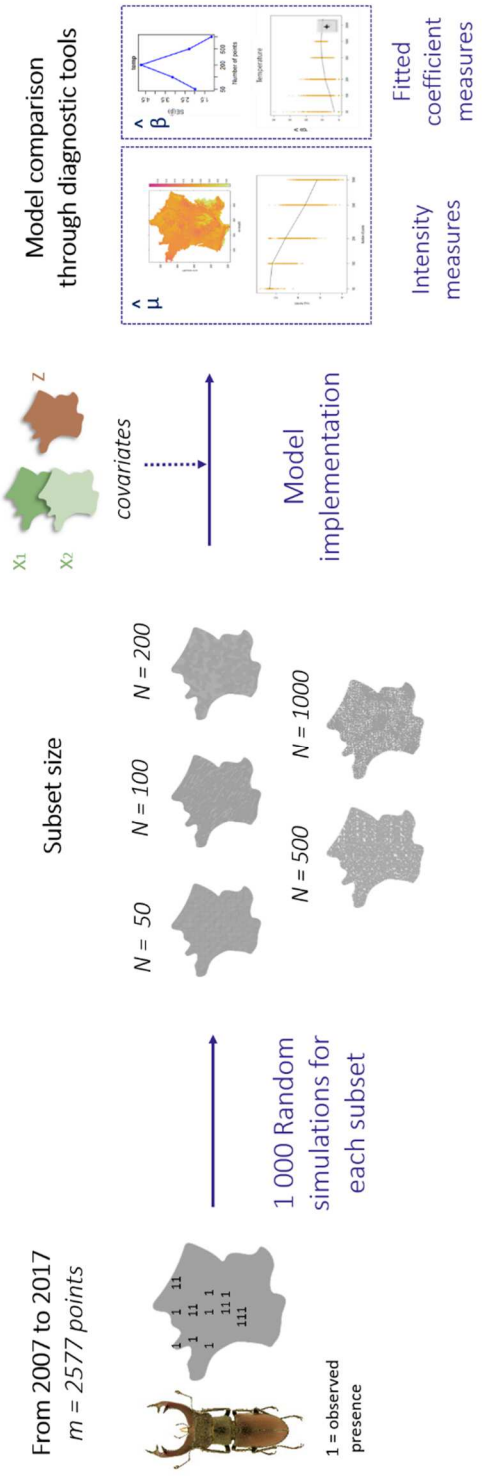
We evaluated the alignment of the fitted model using all available points with models fitted to random subsets of the available points with varying size using the aforementioned diagnostic tools. In this way, we can assess the number of points required to ensure reasonable trust in the fitted models. Our main idea was to assess model stability and congruence in the ecological information inferred from the models. Therefore, 1000 randomizations were run in R for each experiment for each number of subsampled points (N = 50, 100, 200, 500, 1000) (Fig. 2). By simulating a number of subsamples from the whole dataset available, we reproduced a general framework of ecological studies, where the observed dataset is a subset of the whole species pattern. The diagnostic tools we propose may be broadly divided into two categories: tools that measure stability of the fitted intensity surface $\hat{\mu}(s)$ and tools that measure alignment of the fitted coefficients $\widehat{\boldsymbol{\beta}}$ and $\hat{\gamma}$.

Table 2: A short description of the supplied R functions to explore model stability, which are contained in the DiagnosticFunctions.R file supplied in the supplementary material. Full details of these functions and a demonstration of their usage appears in the RMarkdown tutorial in the supplementary material.

| Function | Characteristic | Description |
|---|---|---|
| avg_mu_plot | Intensity surface | Produces a map of the average intensity for a given subset size |
| compute_intensity | Intensity surface | Computes the raw and rescaled intensities for a matrix of fitted model coefficients |
| Corr_plot | Intensity surface | Produces a trace plot of correlation coefficients between the intensity surfaces of the subset |

| | | |
|---|---|---|
| | | models compared to the model fitted with all available points |
| IMSE_plot | Intensity surface | Produces a trace plot of the integrated mean square error of the intensity surfaces of the subset models compared to the model fitted with all available points |
| makeraster | Intensity surface | Creates a raster object of a mapped measure from one of the other functions, with an option to export as a .tif file |
| quantilematch | Intensity surface | Produces a map of misalignment proportions of quantile-categorised intensity surfaces between the subset models of a given subset size and the model fitted with all available points |
| sd_plot | Intensity surface | Produces a map of standard deviations of the intensity surface for a given subset size |
| coef_plot | Fitted coefficients | Produces a trace plot of coefficient estimates across models of various subset sizes |
| coef_se_plot | Fitted coefficients | Produces a trace plot of the standard deviation of the coefficient estimates across models of various subset sizes |
| signcoefs | Fitted coefficients | Computes the number of positive, zero, and negative coefficient estimates for each covariate across all subset sizes |
| signplot | Fitted coefficients | Produces a barplot of the estimated coefficient signs for a given covariate across all subset sizes |
| ZeroEnvEffect | Fitted coefficients | Computes the number of fitted models of each subset size where all coefficients are shrunk to 0 |

230




231

232 Figure 2. The Lucanus PPM workflow for model simulation and diagnostic tools comparison.
233 Diagnostic tools are functions included in the ppmlasso package for R. Vectors from
234 freepik.com.

235 **2.3.1 Intensity measures**

The first set of diagnostic tools assess stability of the fitted intensity surface $\hat{\mu}(s)$. As this intensity surface is typically the primary output of a species distribution model, knowing whether the model which produced it can be assumed to have stabilized is an important consideration. We do this by exploring trends in the intensity surface as subset size changes. Let $\hat{\mu}_{i,N}(s)$ be the fitted intensity of the *i*th subset of size *N* at location *s*. As we will consider multiple subset sizes (*N* = {50, 100, 200, 500, 1000}), we would expect the range of these raw fitted intensities $\hat{\mu}_{i,N}(s)$ to expand as subset size *N* increases. Consequently, we rescale these fitted intensities to achieve a common scale. Because these tools are used to assess stability of a model fitted with the full set of *m* species points, we will define the rescaled fitted intensity $\hat{\mu}_{i,N,m}(s)$ to have the same scaling as the model which uses all *m* points as follows:

$$\text{Equation 5 : } \hat{\mu}_{i,N,m}(s) = \frac{m}{N}\hat{\mu}_{i,N}(s)$$

Here, we present five diagnostic measures of intensity surface stability which are in the DiagnosticFunctions.R file in the supplementary material:

- **Standard deviation of the Intensity.** The fitted model produces an estimate of intensity at each species location in $\boldsymbol{s}$ and each quadrature point in $\boldsymbol{s}_0$. As each subset is randomly sampled, we can examine trends in the variation of intensity as subset size changes. We can thus calculate the standard deviation of the rescaled intensities $\hat{\mu}_{i,N,m}(s)$ across all random subsets, and visualize them in a map produced by the function sd_plot. These standard deviations can be used to quantify the likely variation in intensity at each location *s* for a given subsample size.
- **Average rescaled intensity.** We can calculate the average rescaled intensity $\hat{\mu}_{avg,N,m}(s)$ across all random subsets of size N. Mapping these can indicate when

the fitted intensity has stabilized. We provide a function avg_mu_plot in order to map the rescaled intensity across subsets.

- **IMSE**. The integrated mean square error (IMSE) may be used to measure alignment between the fitted intensity surface using the full set of *m* points with the rescaled intensity of a subset. Because we expect the intensity surface to be right-skewed, we implement the IMSE as a sum of squared differences between the natural logarithm of the fitted rescaled intensities at the quadrature points, as follows:

Equation 6: $IMSE(\hat{\mu}_{i,N,m}) = \sum_{j=m+1}^{m+n}(\ln \hat{\mu}(s_j) - \ln \hat{\mu}_{i,N,m}(s_j))^2$

The higher the IMSE, the greater the dissimilarity between the fitted intensity surfaces. Across subsets, this tool can also be used to inform about model stability. We would expect IMSE to decrease as subset size increases. We provide a function IMSE_plot to visualize the IMSE for each simulated subset and a trace plot of the mean across subsets.

- **Correlation**. We can also measure alignment between the fitted intensity surface using the full set of *m* points with the rescaled intensity of a subset with a correlation measure, using either Pearson's correlation coefficient or non-parametric alternatives such as Spearman's rho or Kendall's tau. Unlike IMSE, correlation measures are bound between -1 and 1, and this scale-free property allows judgment to be made about the raw correlation value in addition to relative comparisons across subset sizes. We provide a function Corr_plot to visualize the chosen correlation measure for each simulated subset and a trace plot of the mean across subsets.

- **Quantile misalignment maps**. While the previous tools are useful summaries of the overall alignment between the fitted intensities of the model using all the points ($\hat{\mu}(s)$) with the rescaled intensities of the models fitted to the subsets ($\hat{\mu}_{i,N,m}(s)$), they do not indicate where the intensity surfaces differ (Pontius and Millones 2011). We provide a function quantilematch that produces a map of quantile misalignment between the models that use random subsets and the model that uses the full data. With this function, the user can supply the desired quantile cutoffs to determine the ordered categories. For example, if the quantiles argument is left at the default of (0.2, 0.4, 0.6, 0.8), locations are placed into one of five categories (corresponding to quantile ranges 0-0.2, 0.2-0.4, 0.4-0.6, 0.6-0.8, and 0.8-1) based on both the fitted intensities $\hat{\mu}(s)$ of the model using all available points as well as the fitted intensities $\hat{\mu}_{i,N,m}(s)$ of the models using random subsets. By quantifying the proportion of differences in categories for each location, the quantilematch function therefore highlights regions where the relative fitted intensities tend to differ between the models fitted to random subsets and the model fitted with all of the data points.

### 2.3.2 Fitted coefficient measures

Ecologists interested in exploring the effects of the environmental covariates ***x*** and the observer bias covariates ***z*** on the fitted model can explore all covariate effects included in the model. Consequently, we present two tools to measure the stability of the coefficient estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$ for the environmental parameters **β** and the observer parameters **γ**. Let $\hat{\beta}_{j,i,N}$ and $\hat{\gamma}_{k,i,N}$ be the *j*th environmental and *k*th observer bias coefficient estimates of the model fitted to the *i*th subset of size *N*. Tools that may be used to explore stability in the coefficient estimates are as follows:

- **Coefficient estimate variability**. As each subset is randomly sampled, we can examine trends in the variation of coefficient estimates as subset size changes. We can thus calculate the standard deviation of the coefficient estimates $\hat{\beta}_{j,i,N}$ and $\hat{\gamma}_{k,i,N}$ across all random subsets. We have provided a function coef_plot which constructs a scatterplot of the fitted estimates of a given coefficient across all simulated subsamples, along with a trace plot of the mean. In addition to the plot, it outputs the mean and standard deviation of the coefficient estimates for each subset size. This diagnostic tool can not only inform about model stability through its dispersion, but also highlight the effect of the different variables on the intensity. This second point can be of major importance for the ecological interpretation of results. Thus, we added the function coef_se_plot which displays empirical standard errors of coefficient estimates along with a trace plot of the standard deviation of the fitted parameter estimates $\hat{\beta}_{j,i,N}$ and $\hat{\gamma}_{k,i,N}$ for each subset size across all environmental and observer bias parameters.
- **Signs of coefficient estimates.** The sign of a coefficient estimate indicates whether it has a positive, neutral, or negative effect on the predicted species distribution, providing insight for ecologists. Consequently, exploring trends in the signs of the fitted coefficients can provide insight into the level of agreement in terms of ecological information. Across subsets, we can compute the proportion of fitted coefficients that have the same sign as the model which uses all *m* points and thus inform about model stability. We have provided a function signcoefs which outputs an array which counts the number of negative, zero, and positive signs for each subset size and coefficient and a function signplot to visualize bar plots of the sign of

the fitted parameter estimates $\hat{\beta}_{j,i,N}$ and $\hat{\gamma}_{k,i,N}$ for each subset size across all environmental and observer bias parameters.

## 3. Results

Full results of the Lucanus analysis are presented below and the supplementary material; we also provide a simulated dataset and stability screening results for this dataset in a separate tutorial.

### 3.1 Predicted intensity, IMSE and misalignment

As we modelled the intensity of the stag beetle across 1000 random subsets of points for different subset sizes, we compared the average intensity of each subset and the fitted intensity for the PPM which uses all 2576 points. As some fitted intensities are very low, we truncate intensities below $10^{-5}$ (1.5% of predicted intensities). By mapping the average rescaled intensity for each subset size (avg_mu_plot function), we note that a pattern appears to stabilize from N = 500 points (Fig. 3a). These maps provide point estimates of the intensity, but we can also assess variability by examining maps of standard deviations of the rescaled fitted intensities (sd_plot function; see Appendix S1 in Supporting information, Fig. S1.2). Moreover, taken one by one, visualization of intensities for models above 500 points were more consistent that those under 500. Indeed, for instance models based on subsets of 50 points appeared more variable between them than those using subsets of 500 points (Supporting information, Fig. S1.3).

Such differences were also visible by plotting the average Pearson correlation (Corr_plot function) between each subset's log intensity and the whole model with all available points was moderately good at 200 points (R ≥0.7) and nearly perfect (R ≥0.97) for models beyond 500 points (Fig. 3b). Furthermore, correlation between subsampled models and the whole

model was greater when the subset contained more than 100 points and consistently above 0.9 with 500 points or more (99.4% of correlations greater than 0.9 when N=500).

These results were confirmed by the log IMSE of each model across subsets (IMSE_plot function; Fig. 3c). Indeed, we can see how the average log IMSE by subset significantly decreases, from around 10.7 at 50 and 100 points to 6.1 at 1000 points. Indeed, pairwise comparisons of IMSE are all significantly different at the 0.1% level (t-test). However, we noticed that the lasso penalty shrank most of the coefficients to zero in models with 200 points or less, leading to greater differences which lead to high IMSE.

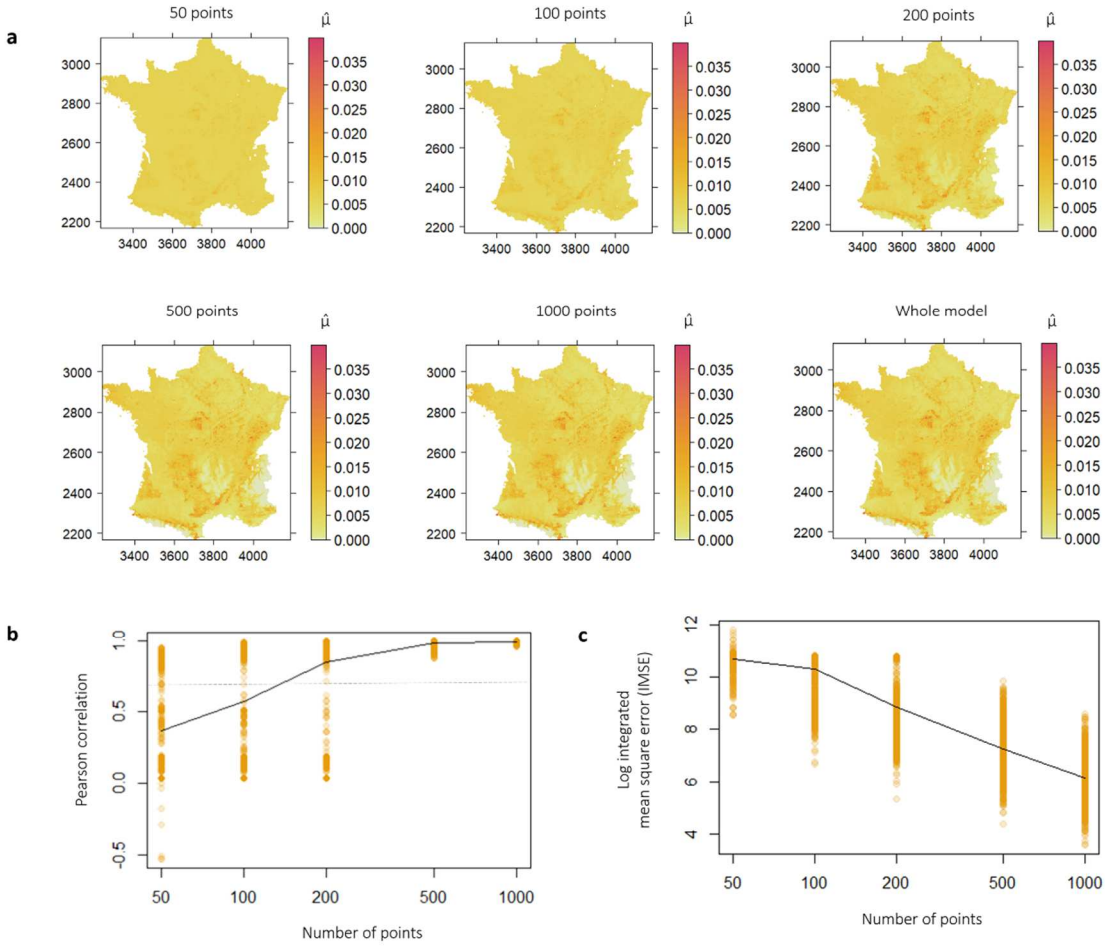



Figure 3. (a) Average rescaled intensity for each subset size (avg_mu_plot function; N = 50, 100, 200, 500 or 1000 points) and the intensity of the whole model with the 2576 stag beetle observations. Maps can be used for graphical comparison. (b) Pearson correlation between

the natural logarithm of the fitted intensity surface from the model using all 2576 points and the rescaled intensity surface from the models using random subsets; the dotted line shows the below which the correlation is considered as low (R >= 0.7) (Corr_plot function). (c) A logarithmic transformation of the integrated mean squared error (IMSE) for each simulation depending on the subset of given points (yellow) (IMSE_plot function).

Furthermore, regarding the quantile matching (quantilematch function), we observed that the level of misalignment is initially very high, because most of the models for subset sizes N = 50 and N = 100 set all coefficients to 0 (Fig. 4). Once we reach a subset size of N = 500, the level of misalignment is much lower. Even at N = 1000, however, there are certain regions where there is relatively high misalignment. Therefore, the interpretation of the intensity surfaces should be more prudent in such areas.

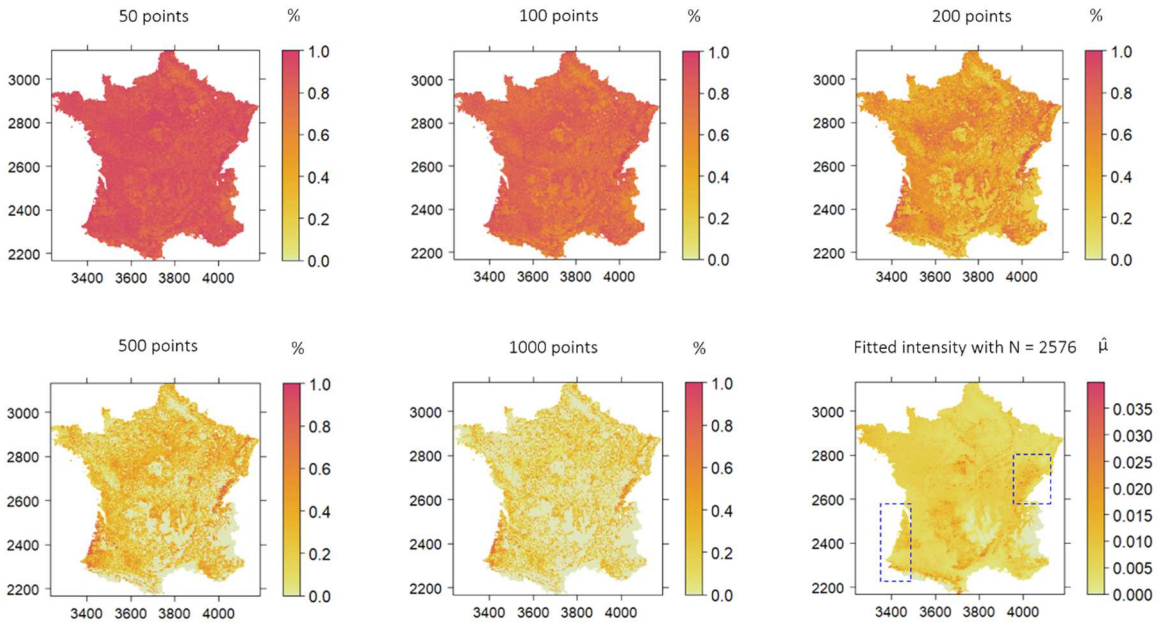
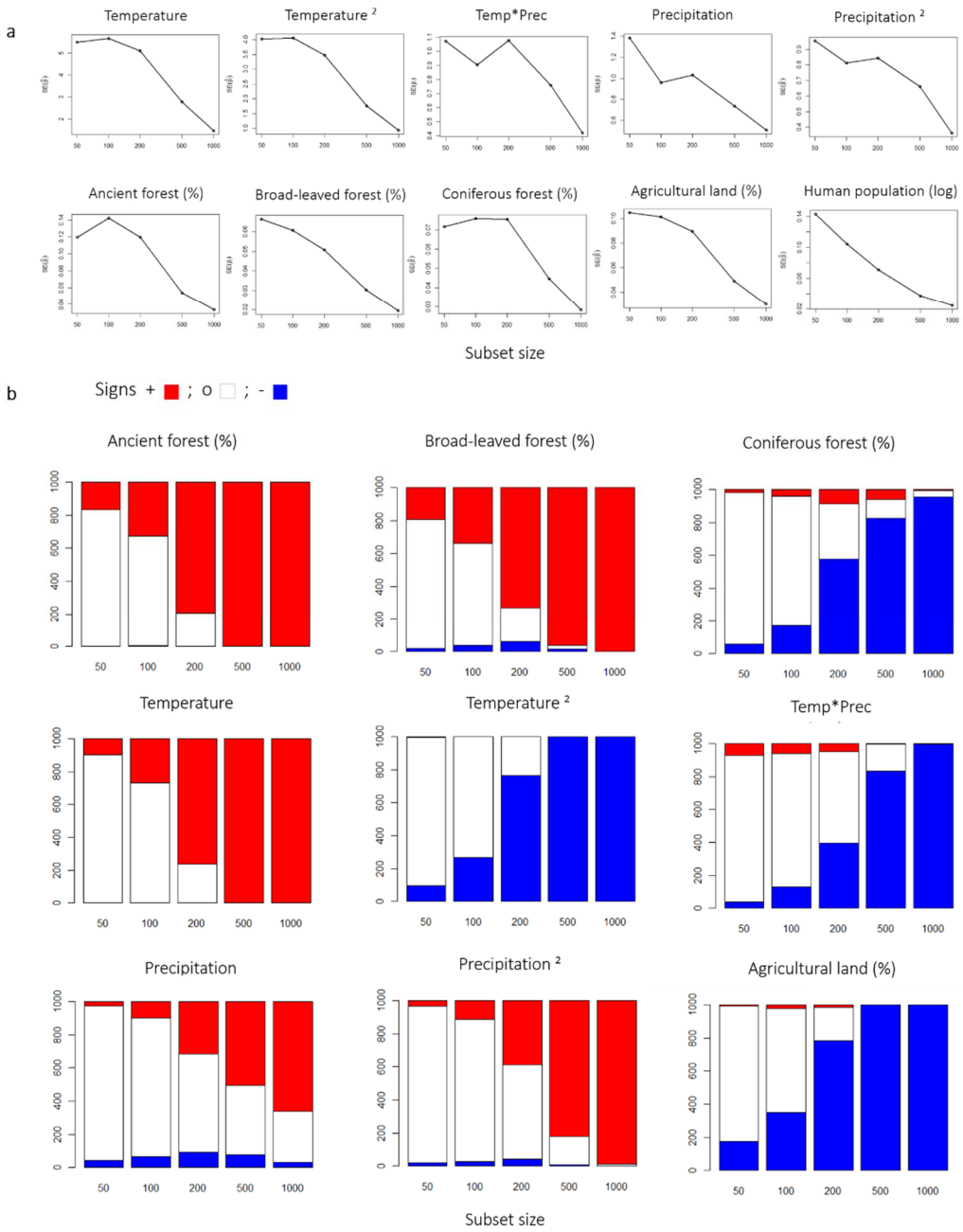



Figure 4. Maps of the proportion of subsets which place the locations into different categories defined by the quantiles 0.2, 0.4, 0.6, and 0.8 than the model using all available points (quantilematch function), for each subset size (50, 100, 200, 500 and 1000 points). The intensity surface (⍰) of the model using the all N = 2576 available points is graphically illustrated next to the 1000 points misalignment map in order to show the areas were the interpretation must be nuanced, particularly the southwest of France and in the east, near the Swiss border.

382

383 **3.2 Covariate effects (β̂ dispersion and ecological agreement)**

384 The plots of the fitted coefficients (coef_plot function) likewise suggest that the model
385 stabilizes when the subset size reaches 500, as the mean values of $\hat{\beta}_{j,i,N}$ appeared to
386 converge to the values $\hat{\beta}_j$ obtained from fitting a model to the full set of 2576 points and the
387 variation in $\hat{\beta}_{j,i,N}$ likewise appeared to decrease with increasing subset size (Supporting
388 information, Figure S1.4). Indeed, the standard deviation of coefficient estimates $\hat{\beta}$
389 consistently decreased when the number of points increased beyond 200 (coef_sd_plot
390 function; Fig. 5a). This result can be seen as congruent with the measures of intensity from
391 Section 3.1, which suggested the stability of models with a number of points of 500 or more.




392

393 Figure 5: (a) Trace plot of the standard deviation of the fitted parameter estimates $\hat{\beta}_{j,i,N}$ and $\hat{\gamma}_{k,i,N}$
394 for each subset size across all environmental and observer bias parameters (coef_sd_plot funtion. (b)
395 Bar plots of the sign of the fitted parameter estimates $\hat{\beta}_{j,i,N}$ for each subset size across all
396 environmental and observer bias parameters (signplot function).

397 It may seem counterintuitive that the average standard deviation, in some cases, increases
398 from N=50 to N=200 and then decreases (Fig. 5a). As many of the coefficient estimates are
399 set to 0 for models of subset size N= 50, this has the effect of decreasing the standard

400 deviation across all simulations. Those coefficients which are non-zero tend to have a large
401 spread, as shown in the graph with a large range of values for $\hat{\beta}_{j,i,50}$ (Supporting information,
402 Figure S1.3). In other words, there is a strong pull toward 0 for N= 50 (as shown in the
403 analysis of coefficient signs in the next subsection), but those coefficients which are not set
404 to 0 tend to be more variable. As subset size increases, the range of the fitted coefficients
405 tends to decrease, but as fewer coefficients are set to 0, the overall standard deviation may
406 be higher across all 1000 simulated subsets for N= 100 and N= 200. Once the subset size
407 reaches 500, however, the range becomes small enough that the overall standard deviation
408 starts to decrease, despite very few coefficients being set to 0.

409 The sign of the fitted coefficients $\hat{\beta}$ (signplot function) informed us about the contribution of
410 each environmental variable to the potential distribution of the Stag beetle in France (Fig.
411 5b; Supporting information, Figure S1.4). Mean annual temperature and mean annual
412 precipitation rate, the percentage of forest cover in the past and the current presence of
413 broad-leaved forest and the natural logarithm of the human population were positively
414 associated with the presence of the species. On the other hand, the percentage of arable
415 land and coniferous forest as the quadratic term of the temperature and the interaction
416 term of climatic variables, were negatively associated with Lucanus presence. The sign of the
417 fitted coefficient estimate $\hat{\gamma}$ (log of the human population variable) were always 100 %
418 positive, whatever the subset size (Supporting information, Figure S1.5).

419 Hence, we looked at the agreement between the signs of $\hat{\beta}$ and $\hat{\gamma}$ across subset models, as
420 shown in Figure 5b, with the signs of the coefficients from the model using all the available
421 data. For the temperature or ancient forest variables, for instance, even if the coefficients
422 were shrunk to zero in many cases (particularly models with less than 200 points) the sign

was always positive when it was non-zero. Coefficients of the other variables fluctuate from positive to negative signs, particularly for models with 200 points or less. For example, the estimated coefficients for precipitation or the interaction term were equally negative and positive until models with 500 points or more. Therefore, getting sign congruency is a sign of model stability. Nevertheless, we got a clear sign tendency with 200 points and more.

The tools are therefore congruent in their conclusions: to model the stag beetle distribution at the French mainland scale and with the given choice of variables and lasso penalty criterion, 500 points are needed to get stabilized models, and consequently in our point of view, also trustworthy conclusions. With a different choice of variables, the number of points necessary for reliable conclusions may differ – in general, the more variables included in the model, the more variation is expected in the fitted intensity surfaces, requiring larger numbers of points to stabilize. Regardless, the tools presented here can be tailored to different spatial scales and choices of variables to investigate model stability.

## 4. Discussion

### 4.1 Assessing PPM stabilization

By departing from our particular data and environmental context, we were able to explore the question of "at what point do my models stabilize?". Our suite of diagnostic tools provides a way to assess the stability of the model in its particular context. Hence, this methodology could be used in order to verify how stable a Poisson point process model fitted with a lasso-type penalty is. Moreover, if the models stabilize at a relatively low number of points, it might mean that the dataset could be divided into shorter periods and used for species distribution analysis across time. For instance, in our case, 500 points seem enough to have a reasonably trustworthy model of the stag beetle. We could have

potentially split the dataset in two and see the differences in the distribution between 2007-2012 and 2012-2017, but in our case temporal heterogeneity of the records did not permit this. Participation in the Stag beetle Quest significantly increased in 2015, and since the average number of records per year is 1000, this implies that a future comparison of models for different time periods could be possible.

As these diagnostic tools rely on exploring stability across different subset sizes, it is important to consider which subset sizes to specify in the simulations. In our context, we fitted models using 10 covariates with over 2500 point locations. We considered subset sizes ranging from N = 50 to N = 1000, thus representing between about 2% and 40% of the total number of points. Indeed, allowing the maximum subset size to be too large could give a false impression of stability due to the fact that there are fewer possible subsets and an increasing number of shared records across subsets. For instance, if we allow the subset size to be 80% of the number of available records, different subsets are guaranteed to share at least 60% of the records in common. As a general recommendation, we advise practitioners to consider subsets ranging in size from a minimum greater than the number of covariates and a maximum less than half of the total number of available records, though this may be quite limiting for data sets with few available records relative to the number of covariates.

While there are certainly other ways to create subsamples aside from sampling at uniform from the available points (i.e half split or block-crossed validation (Roberts et al. 2017), which are certainly preferable for validating models to independent data), such schemes do not seem appropriate for our work. In our model, we also include a term related to sampling bias, and incorporating different subsampling schemes could make it difficult to disentangle effects of the environment from effects of this sampling bias. In this work we want to call

attention to the fact that any given set of observed points represents some (likely biased) subsample of the true point pattern, and by sampling randomly, we thereby preserve any underlying bias patterns of the observed data set. Without direct information regarding sampling effort, creating random subsamples from the observed data set thus mimics the setting in which the observed point pattern is some random subsample of the true point pattern.

We also want to highlight that the criterion we used to select the optimal lasso penalty was the BIC. As we have seen that stability is greatly influenced by the proportion of models for which the coefficients are set to 0, the choice of the criterion for the lasso penalty will also impact the model complexity and hence the number of points necessary for the models to adequately stabilize. If we had instead chosen the AIC, which tends to choose lower penalties than the BIC, model stabilization might have been achieved with smaller subset sizes. Consequently, the effect of criterion choice for the lasso penalty on model stability is a potential area of future research.

If after using these tools, the model does not appear to have adequately stabilized, we recommend results be interpreted with corresponding caution, particularly when the model may be used to inform management or conservation actions. Greater model stability could be achieved by considering a smaller set of covariates, acknowledging that this would lead to less sophisticated ecological understanding of the species distribution and the environmental factors that drive it.

It is important to note that the tools presented in this paper require an adequate number of points in the original pattern to create reasonably-sized subsets. When presence records are rare (around the same number as the number of modelled covariates), it is impractical to

492 produce subsets to assess model stability as there is an increased risk of model convergence
493 problems when fitting models with small numbers of points. However, the spirit of this
494 paper is to explore questions related to the amount of trust that can be placed in a fitted
495 model, and a model fitted using a small number of records is unlikely to be very informative
496 or reliable.

497 Our approach exploits already existent tools in the ppmlasso package and can therefore be
498 already used. In principle, these tools could also be adapted for use with models fitted using
499 other software platforms, such as spatstat. However, these functions were specifically built
500 to extract information from objects with a ppmlasso class, so adaptation of the functions to
501 objects with other classes may be challenging. Moreover, spatstat provides its own functions
502 to assess model stability. Our functions explore model stability through subsetting largely
503 due to the fact that classical statistical estimators such as standard errors are not available
504 when fitting models with a lasso penalty, and the ppmlasso package is specifically designed
505 for the setting of our paper in fitting species distribution models with lasso penalties.

506 This data-driven scrutiny of sample size and model stability is more tailored to analysis of
507 different data sets than ad hoc rules for choosing the number of points to model a certain
508 species. Moreover, it helps us explore trust in the conclusions from the fitted model,
509 particularly for those who use SDMs to inform decisions for conservation.

510 **4.2 Ecological insight**

511 *Lucanus cervus* is a saproxylophagous beetle of conservation interest at the European scale
512 (cited in the 3$^{rd}$ appendix of the Berne convention of 1979 and the 2$^{nd}$ appendix of the
513 Habitat Directive of 1992), that is, subservient to dead or decaying wood; it is observable
514 near old trees, in forest but also in wooded and urban areas. It is a relatively common

515 species in France, and more largely in Europe (Paulian & Baraud, 1982; Bensettiti &
516 Gaudillat, 2004).

517 Using the SDM framework to have ecological insight about the distribution pattern of stag
518 beetles, we observed that climate variables dominated the spatial characterization of the
519 species, particularly the annual temperature (with $\hat{\beta}$ coefficients furthest from 0). This was
520 not a surprise as adults' activity is considered weather-dependent, particularly to conditions
521 of temperature and humidity (Fremlin & Fremlin 2010; Lachat et al. 2012). Indeed, the whole
522 model (Supporting information, Figure S1.2) shows that the species drastically rarifies in
523 mountainous regions where temperatures are lower and humidity higher (massif of the
524 Cevennes, the Pyrenees and the Alps). As the overall alignment between the fitted
525 intensities of the models above 500 points is high (less than 20 % misalignment; fig. 4) in the
526 mountainous areas, we can validate that *Lucanus cervus* sightings are weather related.

527 Land use plays a significant but secondary role. The extent of agriculture, an environmental
528 variable previously thought to be unfavorable for the species was useful. The influence of
529 broad-leaved vs coniferous forests became unambiguous (respectively positive and negative)
530 above 50 points, which may be due to the mixture of trees in forests and the way in which
531 Corine Land cover classifies landscape features (through a visual interpretation of satellite
532 images) at small scale.

533 The abundance of ancient forest was positively associated and plays a significant role in the
534 Lucanus distribution among the land use variables. Perhaps it is due to the selection of local
535 broad-leaved oaks and beeches (Bazire & Gadant, 1991) and availability of dead wood in
536 such plots of old forest. The influence of this variable confirms the influence of the landscape
537 matrix and its history in the current distribution of the stag beetle, as old-growth deciduous

538 forests favor the presence of this saproxylophagous species; it also underlines the inertia of
539 forest systems and should warn us about the potential consequences of the large
540 domination of coniferous plantations occurring for the last 70 years in France (Bazire &
541 Gadant, 1991; Boutefeu 2005).

542 In France, broad-leaved forests are mainly in the plains or at medium altitude. Coniferous
543 stands are mainly in mountainous areas, in the Landes highlands and in recent plantations in
544 western France (Garnier et al., 2018). It is known that coniferous forests are not favorable
545 for this species, even if some *Lucanus* can breed on *Pinus* spp and *Thuja* spp (Paulian &
546 Baraud, 1982; Bensettiti & Gaudillat, 2004). The bar plots of the sign of the fitted parameter
547 estimates showed us that above 500 points the sign is mainly negative. Ecologically, we
548 expected a negative sign, and we only consistently see it from N = 200 onward, so models
549 fitted with fewer than 100 points could have led to conclusions contradictory to ecological
550 knowledge. In interaction with other variables, such as the climatic ones, this can also
551 explain the absence of *Lucanus cervus* in mountainous areas and could be therefore
552 explored in future models.

553 Complementary variables, such as biotic interactions with fungi in decaying wood or other
554 invertebrate species, wood species selection for breeding or micro-climatic variables, which
555 are important for invertebrate development and suggested as important for saproxilic
556 beetles (Diniz-Filho *et al.*, 2010; Quinto *et al.*, 2015; Ulyshen *et al.*, 2017; Garrick *et al.*,
557 2019), were not included as they were not available.

558 The importance of the bias covariate (human population) was significant, showing once
559 again the importance of variables that can correct for sampling bias to accurately model
560 species distributions.

The conclusions inferred from a fitted SDM may be incomplete from an ecological point of view or even inaccurate at small scales. Here we want to underline the important role of experts of the target species and the fact that models approximate a complex reality and should be used with parsimony and caution, especially in conservation contexts. Furthermore, we encourage practitioners to always keep in mind the areas where categories of intensity are most likely to differ between the models fitted to random subsets and the model fitted with all available points (misalignment map). In our case, precautions must be taken before interpreting the whole model in the south-west of France (west part of the Landes highlands) and in particular in the east near the border with Switzerland (Massif du Jura), even though in this second area we had some observations (fig. 1).

## 5. Conclusion

PPMs not only offer a unifying framework to fit presence-only species distribution models with many advantages in model implementation and interpretation, but also possess a number of ready-to-use diagnostic tools that can inform about model consistency and stability. Without any rule of thumb or an obscure single metric, the number of needed points in a particular environmental and spatial context to achieve model stability can be explored from perspectives relating to both the fitted intensity surface and the fitted model coefficients. All of the diagnostic tools are congruent and can be used for any kind of point process model. Above all, we recommend collaboration between species experts and researchers in ecology and statistics to build realistic, field-informed, trustworthy models and test them before applying them. Thanks to the diagnostic tools offered by PPMs, a constructive step-by-step process may allow us to rapidly increase our knowledge of species distributions, even for the less studied ones.

## 6. Aknowledgements

We would like to thank all of the contributors of the Stag Beetle Quest for their enthusiasm for biodiversity citizen science programs and reviewers for the valuable suggestions that greatly improved the manuscript.

Funding – This study was funded by the invited professor grant from the Université Paul-Valéry Montpellier 3 (France).

## *II. References*

Aguirre-Gutiérrez, J., Carvalheiro, L.G., Polce, C., van Loon, E.E., Raes, N., Reemer, M., Biesmeijer, J.C. (2013) Fit-for-Purpose: Species Distribution Model Performance Depends on Evaluation Criteria - Dutch Hoverflies as a Case Study. PLoS One 8. doi:10.1371/journal.pone.0063708

Ahmed, S. E., McInerny, G., O'Hara, K., Harper, R., Salido, L., Emmott, S., & Joppa, L. N. (2015) Scientists and software–surveying the species distribution modelling community. Diversity and Distributions, 21(3), 258-267.

Alabri, A., Hunter, J. (2010) Enhancing the Quality and Trust of Citizen Science Data. 2010 IEEE Sixth International Conference on e-Science, 81–88. doi:10.1109/eScience.2010.33

Bazire, P. & Gadant, P. (1991) La forêt en France - Les études de la documentation française, Paris. 142 p.

Baddeley, A. & Turner, R. (2005) Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software, 12, 1–42.

605 Baddeley, A., Rubak, E., & Turner, R. (2015). Spatial point patterns: methodology and
606 applications with R. CRC press.

607 Bensettiti, F. & Gaudillat, V. (2004) Cahiers d'habitats Natura 2000. Connaissance et gestion
608 des habitats et des espèces d'intérêt communautaire. Tome 7. Espèces animales. La
609 Documentation française. 234-235.

610 Berman, M. & Turner, T. R. (1992) Approximating point process likelihoods with GLIM.
611 Journal of the Royal Statistics Society, Series C, Applied statistics, 41, 31–38.

612 Boutefeu, B. (2005) L'aménagement forestier en France : à la recherche d'une gestion
613 durable à travers l'histoire. VertigO, 6(2), 1-8. doi: 10.4000/vertigo.4446

614 Cardoso, P., Erwin, T.L., Borges, P.A. V, New, T.R. (2011) The seven impediments in
615 invertebrate conservation and how to overcome them. Biological Conservation. 144, 2647–
616 2655. doi:10.1016/j.biocon.2011.07.024

617 Cressie, N. A. C. (1993) Statistics for Spatial Data. John Wiley & Sons, New York.

618 Department of Data and Statistical Studies of the Ministry of Ecology (France) (2012) Corine
619 Land Cover. http://www.geocatalogue.fr/Detail.do?id=300875

620 Devictor, V., Whittaker, R. J., Beltrame, C. (2010) Beyond scarcity: citizen science
621 programmes as useful tools for conservation biogeography. *Diversity and distributions*,
622 *16*(3), 354-362.

623 Diniz-Filho, J.A.F., de Marco, P., Hawkins, B.A. (2010) Defying the curse of ignorance:
624 Perspectives in insect macroecology and conservation biogeography. Insect Conservation
625 and Diversity. 3, 172–179. doi:10.1111/j.1752-4598.2010.00091.x

626 Donaldson, M.R., Burnett, N.J., Braun, D.C., Suski, C.D., Hinch, S.G., Cooke, S.J., Kerr, J.T.
627 (2016) Taxonomic bias and international biodiversity conservation research. Facets 1, 105–
628 113. doi:10.1139/facets-2016-0011

629 Duque-Lazo, J., Van Gils, H. A. M. J., Groen, T. A., & Navarro-Cerrillo, R. M. (2016)
630 Transferability of species distribution models: The case of Phytophthora cinnamomi in
631 Southwest Spain and Southwest Australia. Ecological modelling, 320, 62-70.

632 Fick, S.E., Hijmans, R.J. (2017) Worldclim 2: New 1-km spatial resolution climate surfaces for
633 global land areas. International Journal of Climatology.

634 Fisher-Phelps, M., Cao, G., Wilson, R.M., Kingston, T. (2017) Protecting bias: Across time and
635 ecology, open-source bat locality data are heavily biased by distance to protected area.
636 Ecological Informatics. 40, 22–34. doi:10.1016/j.ecoinf.2017.05.003

637 Frank, K., Hülsmann, M., Assmann, T., Schmitt, T., Blüthgen, N. (2017) Land use affects dung
638 beetle communities and their ecosystem service in forests and grasslands. Agriculture,
639 Ecosystems and Environment. 243, 114–122. doi:10.1016/j.agee.2017.04.010

640 Fremlin, M., Fremlin, D. H. (2010) Weather-dependence of Lucanus cervus L.(Coleoptera:
641 Scarabaeoidea: Lucanidae) activity in a Colchester urban area. Essex Naturalist (New Series),
642 27, 214-230.

643 R. Fuchs, M. Herold, P.H. Verburg, J.G.P.W. Clevers (2013): A high-resolution and harmonized
644 model approach for reconstructing and analysing historic land changes in Europe,
645 Biogeosciences, 10(3), 1543–1559, doi:10.5194/bg-10-1543-2013

646 Fuchs, R., Herold, M., Verburg, P.H., Clevers, J.G.P.W., Eberle, J. (2014) Gross changes in
647 reconstructions of historic land cover/use for Europe between 1900-2010. Global Change
648 Biology, doi: 10.1111/gcb.12714

649 Fuchs, R., Verburg, P.H., Clevers, J.G.P.W., Herold, M. (2015) The potential of old maps and
650 encyclopaedias for reconstructing historic continental land cover/use change, Applied
651 Geography, 59, 43-55. doi:10.1016/j.apgeog.2015.02.013

652 Garnier, M., Bir, J., Du Puy, S., Derrière, N., Dalmasso, M., Wurpillot, S., Colin, A., Benest F.
653 (2018). La forêt française – État des lieux et évolutions récentes – Panorama des résultats de
654 l'inventaire forestier. IGN. Édition 2018, 56 pages. https://inventaire-
655 forestier.ign.fr/IMG/pdf/180906_publiff_bd.pdf

656 Garrick, R.C., Reppel, D.K., Morgan, J.T., Burgess, S., Hyseni, C., Worthington, R.J., Ulyshen,
657 M.D. (2019). Trophic interactions among dead-wood-dependent forest arthropods in the
658 southern Appalachian Mountains, USA. Food Webs, 18, e00112.
659 doi:10.1016/j.fooweb.2018.e00112

660 Guillera-Arroita, G. (2017) Modelling of species distributions, range dynamics and
661 communities under imperfect detection: advances, challenges and opportunities. Ecography,
662 40(2), 281-29.

663 Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E.,
664 Mccarthy, M.A., Tingley, R., Wintle, B.A. (2015) Is my species distribution model fit for
665 purpose? Matching data and models to applications. Global Ecology and Biogeography. 24,
666 276–292. doi:10.1111/geb.12268

667 Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I.,
668 ... Martin, T. G. (2013) Predicting species distributions for conservation decisions. *Ecology*
669 *letters*, *16*(12), 1424-1435.

670 Hawes C. J. (2008) The stag beetle Lucanus cervus (Linnaeus, 1758) (Coleoptera: Lucanidae):
671 a mark-release-recapture study undertaken in one United Kingdom residential garden *IN*
672 Revue d'écologie, SUP10" 4ème Colloque sur la Conservation des Coléoptères
673 Saproxyliques", Vivoin, Sarthe. Société nationale de protection de la nature et
674 d'acclimatation de France, Paris .

675 Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. (2005) Very high resolution interpolated
676 climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978.

677 Irmler U, Arp H, Nötzold R. (2010) Species richness of saproxylic beetles in woodlands is
678 affected by dispersion ability of species, age and stand size. Journal of Insect Conservation,
679 14(3), 227-235.

680 Isaac, N.J.B. & Pocock, M.J.O. (2015) Bias and information in biological records. Biological
681 Journal of the Linnean Society, 115(3), 522–531. https://doi.org/10.1111/bij.12532

682 Meriguet, B., Merlet, F., Houard, X. (2012) Enquête d'insecte : le Lucane cerf-volant - Bilan
683 2011 et perspectives 2012. Insectes, 24. Doi:10.13140/RG.2.2.26326.70721

684 Lachat, T., Wermelinger, B., Gossner, M. M., Bussler, H., Isacsson, G., & Müller, J. (2012).
685 Saproxylic beetles as indicator species for dead-wood amount and temperature in European
686 beech forests. Ecological Indicators, 23, 323-331.

687 Leandro, C., Jay-Robert, P., Vergnes, A. (2017) Bias and perspectives in insect conservation: A
688 European scale analysis. Biological Conservation. 215, 213–224.
689 doi:10.1016/j.biocon.2017.07.033

690 Paulian, R., Baraud, J. (1982) Faune des Coléoptères de France, vol. II: Lucanoidea et
691 Scarabaeoidea. Paris, Lechevalier.

692 Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017) Opening the
693 black box: An open-source release of Maxent. *Ecography*, *40*(7), 887-893.

694 Pontius Jr, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and
695 allocation disagreement for accuracy assessment. International Journal of Remote Sensing,
696 32(15), 4407-4429.

697 Powney, G. D., Isaac, N. J. (2015) Beyond maps: a review of the applications of biological
698 records. *Biological Journal of the Linnean Society*, *115*(3), 532-542.

699 Quinto, J., De Los Ángeles Marcos-García, M., Díaz-Castelazo, C., Rico-Gray, V., Galante, E.,
700 Micó, E. (2015) Association patterns in saproxylic insect networks in three Iberian
701 Mediterranean woodlands and their resistance to microhabitat loss. PLoS One 10, 1–14.

702 R Development Core Team. (2020) R: A language and environment for statistical computing,
703 reference index version 4.0.2. Viena, Austria. https://www.r-project.org/

704 Renner, I.W. (2013) Advances inPresence-Only Methodsin Ecology. Submitted for the degree
705 of Doctor of Philosophy. University of New South Wales, Australia.

706 Renner, I.W., Warton, D.I., Hui, F.K.C. (in press) What is the effective sample size of a spatial
707 point process? Australian and New Zealand Journal of Statistics.

708 Renner, I.W., Warton, D.I. (2013) Equivalence of MAXENT and Poisson Point Process Models
709 for Species Distribution Modeling in Ecology. Biometrics 69, 274–281. doi:10.1111/j.1541-
710 0420.2012.01824.x

711 Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton,
712 D.I. (2015) Point Process Models for Presence-Only Analysis. Methods in Ecology and
713 Evolution 6 (4): 366–79. doi:10.1111/2041-210X.12352.

714 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Warton, D. I.
715 (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or
716 phylogenetic structure. Ecography, 40(8), 913-929.

717 Ross, L.K., Ross, R.E., Stewart, H.A., Howell, K.L. (2015) The influence of data resolution on
718 predicted distribution and estimates of extent of current protection of three “listed” deep-
719 sea habitats. PLoS One 10, 1–19. doi:10.1371/journal.pone.0140061

720 SEDAC (2016) Gridded Population of the World, version 4 (GPWv4).
721 https://sedac.ciesin.columbia.edu/data/collection/gpw-v4

722 Stirling, D. A., Boulcott, P., Scott, B. E., & Wright, P. J. (2016) Using verified species
723 distribution models to inform the conservation of a rare marine species. Diversity and
724 Distributions, 22(7), 808-822.

725 Stockwell, D. R., & Peterson, A. T. (2002) Effects of sample size on accuracy of species
726 distribution models. Ecological modelling, 148(1), 1-13.

727 Thomaes, A., Kervyn, T., Maes, D. (2008) Applying species distribution modelling for the
728 conservation of the threatened saproxylic Stag Beetle (Lucanus cervus). Biological
729 Conservation. 141, 1400–1410. doi:10.1016/j.biocon.2008.03.018

730 Thuiller, W., Lafourcade, B., Engler, R., & Araujo, M. B. (2009) BIOMOD – A platform for
731 ensemble forecasting of species distributions. Ecography, 32, 369-373

732 Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. Journal of the Royal
733 Statistical Society, Series B, 58, 267–288.

734 Ulyshen, M.D., Zachos, L.G., Stireman, J.O., Sheehan, T.N., Garrick, R.C. (2017) Insights into
735 the ecology, genetics and distribution of Lucanus elaphus Fabricius (Coleoptera: Lucanidae),
736 North America's giant stag beetle. Insect Conservation and Diversity. 10, 331–340.
737 doi:10.1111/icad.12229

738 Virgili, A., Authier, M., Monestiez, P., Ridoux, V. (2018) How many sightings to model rare
739 marine species distributions. *PloS one*, *13*(3).

740 Warton, D.I., Renner, I.W., Ramp, D. (2013) Model-based control of observer bias for the
741 analysis of presence-only data in ecology. PLoS One 8. doi:10.1371/journal.pone.0079168

742 *Data accessibility statement*:All data (tutorial, Rdata files and scripts with functions) have
743 been uploaded as "supplementary material".

744