Correcting for observer bias behaviour: learning from a virtual observer approach.

Emy Guilbault[1], Ian Renner[2], and Panu Somervuo[3]

[1]University of Helsinki
[2]The University of Newcastle
[3]University of Helsinki Faculty of Biological and Environmental Sciences

November 22, 2023

Abstract

In recent years, the increase of data availability through citizen science campaigns has raised questions on the quality of this data. Species distribution models can be severely impacted by non-random spatial distributions of records. Multiple methods exist to correct for spatial bias and most of them imply that the sampling is uneven in space and determined by the observers' choices of where to search for observations. One common correction method is to include a covariate in the model as a proxy for sampling bias and correcting for this bias by setting this covariate equal to a common value upon prediction. However, this approach implies that each observer behaves in the same manner, which in practice may not be the case. Here, we differentiate two common observer behaviours: exploring and following. Under this paradigm, explorers seek to observe species in new places far away from other observations and away from common routes of transit. By contrast, followers search near already observed species locations and remain closer to common routes of transit. In this paper, we investiage whether the current approaches to correcting for observer bias hold under varying observer behaviours, or whether a data-driven approach based on modelled observer behaviour may lead to better predictions. To do so, we developed a new software platform, obsimulator, to simulate patterns of points driven by observer behaviour. We established two correction methods based on a bias incorporation approach using k-nearest neighbours and density calculation. Broadly, we found that the method of including a bias covariate and setting it to a common value for prediction yields the best results. We also found that the knn-based correction outperformed the density-based correction. Additionally, we provide guidance for setting model parameters based on the ratio of explorers versus followers in the observers' cohort.

Correcting for observer bias behaviour: learning from a virtual observer approach.

November 20, 2023

Abstract

In recent years, the increase of data availability through citizen science campaigns has raised questions on the quality of this data. Species distribution models can be severely impacted by non-random spatial distributions of records. Multiple methods exist to correct for spatial bias and most of them imply that the sampling is uneven in space and determined by the observers' choices of where to search for observations. One common correction method is to include a covariate in the model as a proxy and correcting for this bias by setting this covariate equal to a common value upon prediction. However, this correction implies that each observer behaves in the same manner, which in practice may not be the case. We can differentiate two common observer behaviors: exploring and following. Under this paradigm, explorers do not always follow the road network and will seek to observe species in new places far away from other observations. By contrast, followers will search close to already observed species locations and will stay closer to the road network. As such, it is worth investigating whether the current approaches to correcting for observer bias hold under varying observer behaviours, or whether a data-driven approach based on modelled observer behaviour may lead to better predictions. To do so, we developed a new software platform, obsimulator, to simulate patterns of points driven by observer behaviour. We established two correction methods based on a bias incorporation approach using k-nearest neighbours and density calculation. Broadly, we found that the method of including a bias covariate and setting it to a common value for prediction yields the best results. We also found that the knn-based correction outperformed the density-based correction. Additionally, the optimal number of neighbouring points and smoothing parameters depends on the ratio of explorers versus followers in the observers' cohort.

Keywords: Spatial point pattern - Citizen science - Ecologist simulator - Observer behaviour

1 Introduction

Citizen science data has become a common source of information in ecology Dickinson et al. [2010], but many challenges still exist to fully understand the strengths and weaknesses of such data Brown and Williams [2019]. Citizen science data has become popular for financial, practical and technological reasons Cohn [2008], Silvertown

[2009], Dickinson et al. [2010]. A major challenge for citizen science lies in the reliability of the data itself. Citizen science data may vary in quality by study area and by project Cohn [2008], Dickinson et al. [2010]. Even if multiple studies have shown that such projects are valuable for research, there are still some questions about the variability and accuracy of the data Kosmala et al. [2016], Aceves-Bueno et al. [2017]. Among other concerns, there may not be information about the data collection process and there is no guarantee of the validity of the observations nor the accuracy of the locations. With the increased use of presence-only data (PO) from opportunistic sources, researchers have devised statistical tools and filtering methods to cope with these concerns Dickinson et al. [2010], Freitag et al. [2016], Kosmala et al. [2016], Johnston et al. [2019].

During the observation process, observer behaviour and choices can greatly impact what is reported and where Arazy and Malkinson [2021], Bowler et al. [2022], Dimson and Gillespie [2023], Geldmann et al. [2016]. An observer's searching routine can be influenced by accessibility, such as the presence of transit lines (roads, railways or waterways) or by a particular environmental condition, resulting in less sampling effort in more remote locations. Moreover, some observers may choose to visit sites where they believe the species will be present due to previous records. The resulting data set of reported observations consequently represents a biased distribution of the true species pattern over the study area. Many methods have been developed over the years to account for this observer bias, including data modification (spatial filtering, the weighting of occurrences), background modification (target group background, presence-absence data, detectability), data integration (repeated data collection, combined datasets, ensemble or joint models) and incorporating bias (offset term, adding terms or covariates in a statistical model).

Data modification such as spatial filtering can be done using thinning methods or sub-sampling, but it is limited by the sample size because it reduces the number of records available and potentially the predictive performance Anderson and Raza [2010], Beck et al. [2014], Boria et al. [2014], Rose et al. [2019]. Another possibility is to apply a simple prior weighting term to the samples or occurrences Stolar and Nielsen [2015] or into the selection of pseudo-absences Zaniewski et al. [2002]. Background modification and target-group background approaches can generate presence-absence data (PA) with the same spatial bias Phillips and Dudík [2008], Higa et al. [2015], Phillips et al. [2009]. However, these have been criticised for reflecting the species' composition rather than distribution, and may overestimate bias in poorly sampled areas Elith and Leathwick [2007], Phillips et al. [2009], Mair et al. [2017]. The presence points of non-target species could be used as pseudo-absences Ranc et al. [2017] and can replace observer bias with species richness bias Warton et al. [2013]. More recently, Vollering et al. [2019] developed a "background thickening" method which increases the background density around point presences, showing promise for small sample sizes. Data integration can combine multiple data sources or models. Multiple collection repetitions can decrease the bias in the datasets but require more time and resources Tyre et al. [2003], Benoît and Allard [2009], Pollock et al. [2014]. One approach is pooling PO data with unbiased PA data, counts, or occupancy data, but this requires another unbiased dataset Fithian and Hastie [2013], Fithian et al. [2015], Renner et al. [2019]. An ensemble of outputs is an alternative which uses both species occurrences and remote sensing information; however, access and resolution are limited and ensemble element independence is rarely achieved Tang et al. [2020]. Finally, accounting both for data sampling processes through correlation structure or latent processes and ecological responses can overcome such bias Diggle et al. [2010], Conn et al. [2017], Johnston et al. [2020].

Here we focus on the latest bias correction category which happens during the modelling process. To account for such bias, an offset term in the linear predictors can be used, but this implies knowing the observer effort Chakraborty et al. [2011], Merow et al. [2016], Pacifici et al. [2017]. Some authors have introduced a spatially unstructured term Illian et al. [2013]; or a covariate that can inform about duration, length of search, expertise, ignorance score, or other information collected about observers Mair and Ruete [2016], Johnston et al. [2018], Kelling et al. [2019]. However, there is a possible confusion between sampling bias and autocorrelation among the environmental covariates Segurado et al. [2006]. Other modelling approaches offer flexibility with readily available tools, such as the quasi-linear Poisson point process in R to model environmental covariates and bias in separate clusters using harmonic Poisson point patterns Komori et al. [2020]. Finally, the bias can be corrected in the predictions using covariates as a proxy and thus factored out Chakraborty et al. [2011], Warton et al. [2013], El-Gabbas and Dormann [2018], Renner et al. [2019], Skroblin et al. [2019]. One common proxy is to calculate distances to the road network and correct for this bias by setting the modelled covariate equal to a common value Warton et al. [2013], Renner et al. [2019]. Still, this correction implies that each observer behaves in the same manner, which in practice may not be the case.

While a virtual species approach via simulations is of growing interest to test parameters and performances of modelling approaches [Meynard et al., 2019], the virtual ecologist approach with a focus on the sampling and observation process is still scarce Zurell et al. [2010]. In this article, we present the `obsimulator` software that we developed to produce presence-only data sets with different observer behaviour; controlling for their movements and spatial distribution as well as their ability to make an observation (accuracy). From there, we focus on the ability of the Warton et al. [2013] method to account for sampling bias under differing observer behaviour profiles. We differentiate two common observer behaviours: *exploring* and *following*. Under this paradigm, explorers do not always follow the road network and will seek to observe species in new places far away from other observations. By contrast, followers will search near already observed species locations and will stay closer to the road network. Using second order effects of point pattern analysis methods, we study the spatial patterns of observations and then, correct the sampling bias in spatial predictions. We investigate whether the Warton et al. [2013] approach to correcting for observer bias holds under varying observer behaviours, or whether a data-driven approach based on modelled observer behaviour may lead to better predictions.

2 Material and Methods

To investigate the impact of observers' behaviour on the resulting pattern of species they observe, we developed a virtual ecologist simulator. The software defines how the observers move and which targets they reach in space and time to mimic the sampling process of opportunistic observations. Following the sampled observation process, we study the spatial distribution of the observed records and test various bias correction methods derived from the Warton et al. [2013] method, as illustrated in Figure 1.

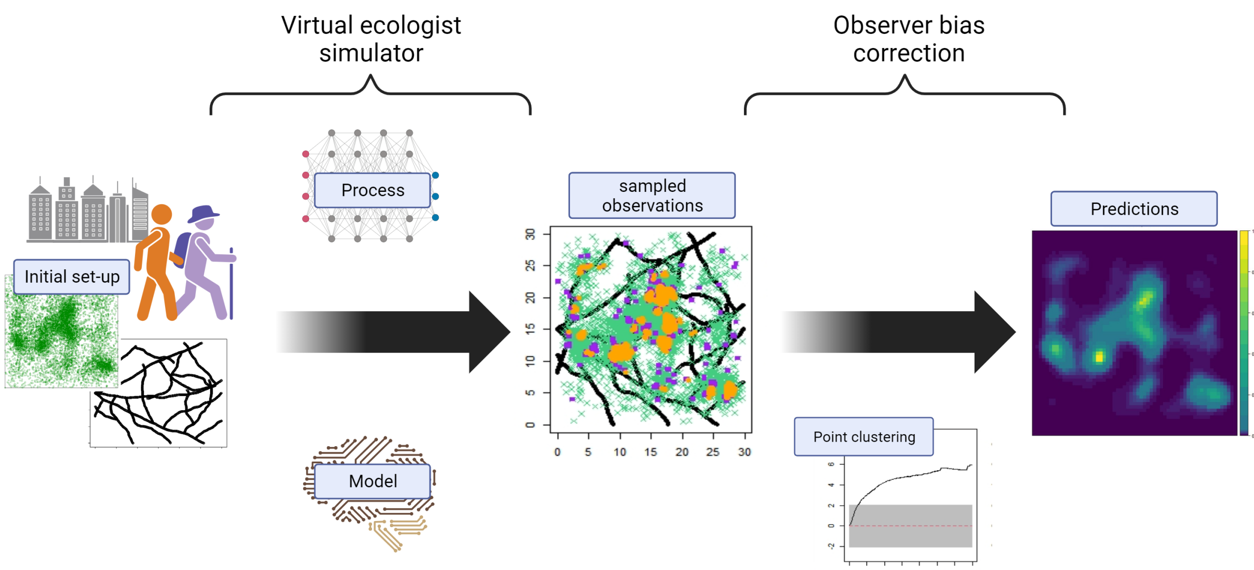

Figure 1: *Method display: virtual ecologist simulator and observer bias correction. Created with BioRender.com*

2.1 Virtual observer simulation

We developed a `C` program for simulating point processes in continuous time and space called `obsimulator` which is run via a computer terminal. The output can be imported into `R` R Core Team [2017] for summary analysis and visualisation. An example of how the software can be used and a description of the processes appear in the Appendices.

`Obsimulator` is defined by a process file and a model file. The process file contains the syntactic descriptions of the processes and their parameters (selection of targets, movement of observers and observation of species). The model file defines the identities and parameter values of the processes by which observers emerge, select their targets, move toward their targets, and make observations, as shown in Appendix **??**.

To simulate point patterns of observed targets using `Obsimulator`, we first define in `R` the initial states of our area of interest. The initial setup defines the city coordinates and the number of observers (both explorers and followers) as well as the distribution of the species (target points) that observers can potentially reach. In our case, we simulated targets in `R` following the methods outlined in Renner et al. [2019] without any sampling bias to model the true realization of the species in space according to their environmental preferences. We defined a set of four simulated covariates $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$ to represent the species' habitat preferences and simulated 5000 individuals to serve as target points for the observation process. We also created a road network along which observers travel to move towards targets using functions from Renner et al. [2019].

Having generated the road network and species distribution, we simulated the observation process using 20 observers in total. See Appendix **??** for details.

To explore the differences in the observed patterns through differing observer behaviour, we varied:

- The ratio of the number of explorers (E) versus the number of followers (F): 1:19, 5:15, 10:10, 15:5, or 19:1;
- The proportion of target points that end up being observed: 5%, 25%, or 45%.

We repeated the simulation 20 times using different seeds.

2.2 Measuring spatial clustering

To measure the spatial clustering of a point pattern, one can use Ripley's $K(r)$ function, the pair correlation function, or various extensions Renner [2013], Wiegand and Moloney [2013], Baddeley et al. [2016]. These measures can identify whether the point pattern is regular, independent or clustered using distances measures:

- K-function:

$$\hat{K}(r) = \frac{1}{|\mathcal{A}|} \sum \sum_{i \neq j} \frac{I_r(d_{ij})}{\hat{\mu_i}\hat{\mu_j} \times w_{ij} \times |\mathcal{A}|} \tag{1}$$

  Ripley's K-function counts the number of points within a buffer of radius r around each point location. In the numerator, d_{ij} is the distance between points i and j, I_r is an indicator of whether the input distance is less than or equal to r, and w_{ij} is a weight function that provides an edge correction. In the denominator, $\hat{\mu_i}$ and $\hat{\mu_j}$ are the intensities estimates at points i and j while $\mathcal{A}$ is the area of the spatial domain.

- The L-function is a rescaling of the K-function, defined as follows:

$$\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} \tag{2}$$

For information on the temporal and spatio-temporal clustering evaluation see Guilbault [2022]. We examined the L-function to assess the degree of spatial clustering in the simulated patterns of observed points. Specifically, we used the `R` functions `envelope` and `Linhom` in the package `spatstat` to plot $\hat{L}(r) - r$ along with 95% confidence envelopes as in Baddeley et al. [2016].

2.3 Model-based observer bias correction

To correct for observer bias, we have extended the Warton et al. [2013] method of including covariates to model observer bias and setting these covariates to a common value for prediction. We fit a Poisson point process model with both the four simulated environmental variables $\mathbf{x}_1, \ldots, \mathbf{x}_4$ used to generate the targets points and a proxy variable $\mathbf{z}^c$ for the observer bias. Thus we maximize the following log-likelihood function:

$$\log \mu(s) = \beta_0 + \sum_{i=1}^{4} x_i(s) \times \beta_i + \mathbf{z}^c \times \beta_z \tag{3}$$

where $\mu(s)$ is the intensity at a location s, β_i is the coefficient associated with the environmental variables $\mathbf{x}_i$, β_0 is the intercept, and β_z is the coefficient associated with the bias covariate $\mathbf{z}^c$.

Our proxy covariate for the observer bias is denoted $\mathbf{z}$, and is a measure of the distances between the points in space using one of two approaches. The first approach defines the bias according to a knn algorithm at different k nearest neighbor values: either single values 1, 2, 3, or 5, or a combination of values 1:k for $k = 2$, 3, or 5. These distances are calculated using the `nndist` function in `spatstat`. The second approach is a measure of the density of points with edge corrections, using different standard deviations of the isotropic smoothing kernel value: 0.1, 0.5, 1, 1.5, 2, or 5. These densities are computed using the `density.ppp` function in `spatstat`. The proxy covariate to correct for observer bias is created as follows:

$$\mathbf{z}^c = \alpha \times c + (1 - \alpha) \times \mathbf{z} \tag{4}$$

where $\alpha \in [0, 1]$ is a coefficient to adjust the bias correction, with values closer to 1 resulting in a stronger correction. By setting $\alpha = 1$, this correction method is equivalent to that of Warton et al. [2013]. Here, c is a chosen constant, commonly 0 or either the minimum or mean of $\mathbf{z}$. We only focus on the minimum value of $\mathbf{z}$. The bias covariate calculated reflects either the road network distribution, the point clustering or both. Our hypothesis was that the optimal choice of α may depend on the behaviour of the observers. In particular, we believed that a value of α closer to 1 would be optimal in settings where the relative number of followers was high and a value closer to 0 would be optimal in settings where the relative number of explorers was high.

2.4 Model evaluation

To evaluate the performances of the different models, we measure the agreement between the true species intensity and the predicted intensity using both Pearson's correlation and Integrated Mean Square Error (IMSE) [Swanepoel, 1988, Wand and Jones, 1994]. Because the scale of the IMSE depends on the magnitude of the true intensity, we rescaled both the true and predicted intensities to have a common mean to make for an equitable comparison. In practice, we measure the intensity at the quadrature points used in fitting the models, which simplifies the calculation. Because IMSE can give considerable variation among methods, we will use a normalized IMSE (NIMSE), defined by:

$$NIMSE = \frac{\sum(\hat{\mu}(s) - \mu(s))^2}{\hat{\sigma}_\mu^2} \tag{5}$$

Where, $\hat{\sigma}_{\mu}^2$ is the variance of the predicted intensity from the model. We evaluate the differences in NIMSE and correlation between the corrected model (3) and the non-corrected model (model fitted without the bias proxy covariate in (3) for the same percentage of observed targets to investigate the use of the clustering bias correction. Superior performance for the corrected models would be reflected in a negative difference in NIMSE and a positive difference in correlation.

In Section 3, we evaluate the performance of the different `obsimulator` parameters outlined in Section 2.1, the two correction methods (knn and density-based) as well as their parameters α, k, and standard deviations of the isotropic smoothing kernel value as defined in Section 2.3. We also examine the predicted intensity maps of these models.

3 Results

3.1 Differences in patterns of observed points

First, we investigate the spatial distribution of observed targets by the simulated followers and explorers. In Figure 2 the point clustering is more noticeable with fewer explorers (towards the right of the figure) and concentrated around nodes and road sections. Increasing the percentage of observed targets (second and third rows) amplifies this clustering.

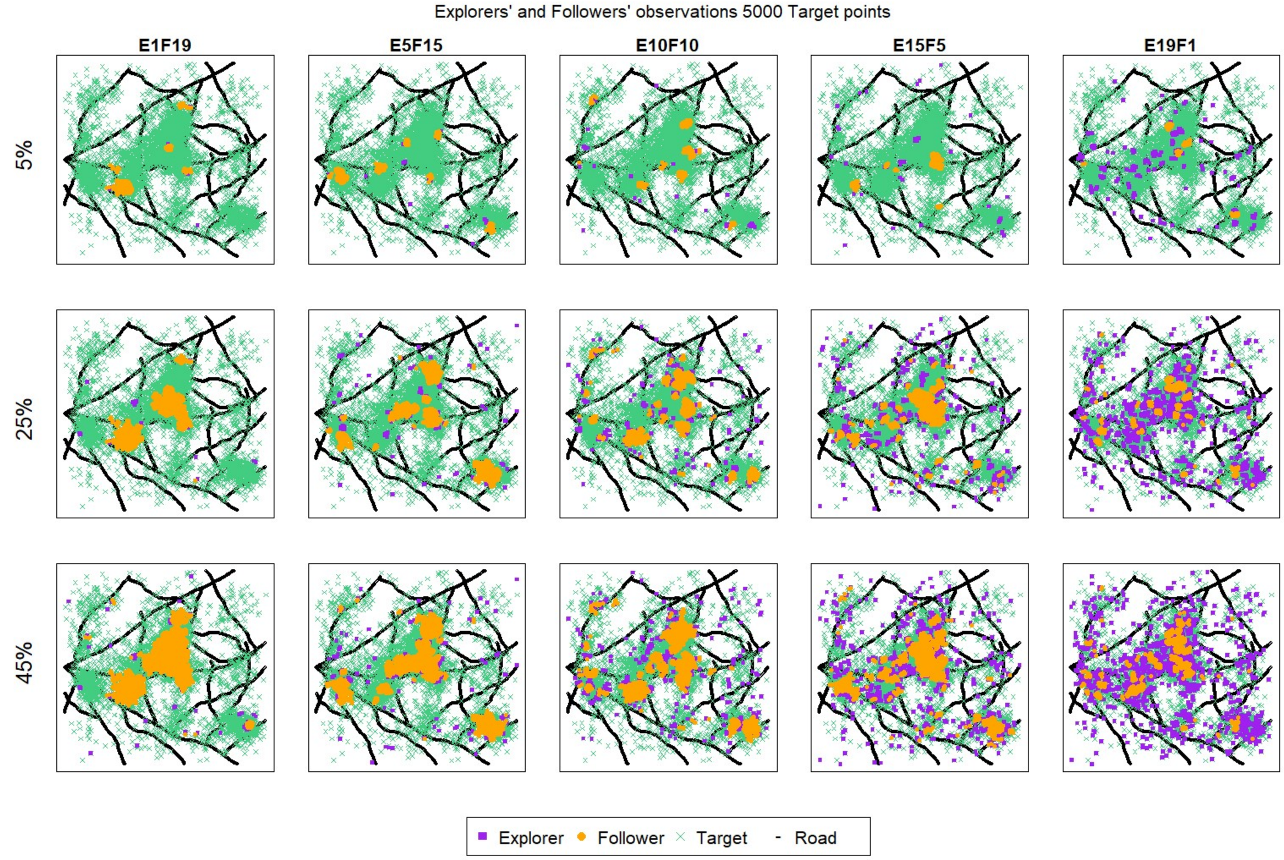

Figure 2: *Patterns of observed points by explorers (purple) and followers (orange) from among the 5000 target points (green). The road network is in black. Each row represents the proportion of observed points: 5%, 25% and 45%. Each column represents a different ratio of explorers and followers: 1:19, 5:15, 10;10, 15:5, and 19:1.*

The degree of spatial clustering as measured by $\hat{L}(r) - r$ and shown in Figure 3 appears higher and becomes significant (above the simulation envelope) at shorter distances u when the number of followers is higher than the number of explorers, with the possible exception of the ratios 5:15, 10:10, and 15:5 for 5% of observed target points.

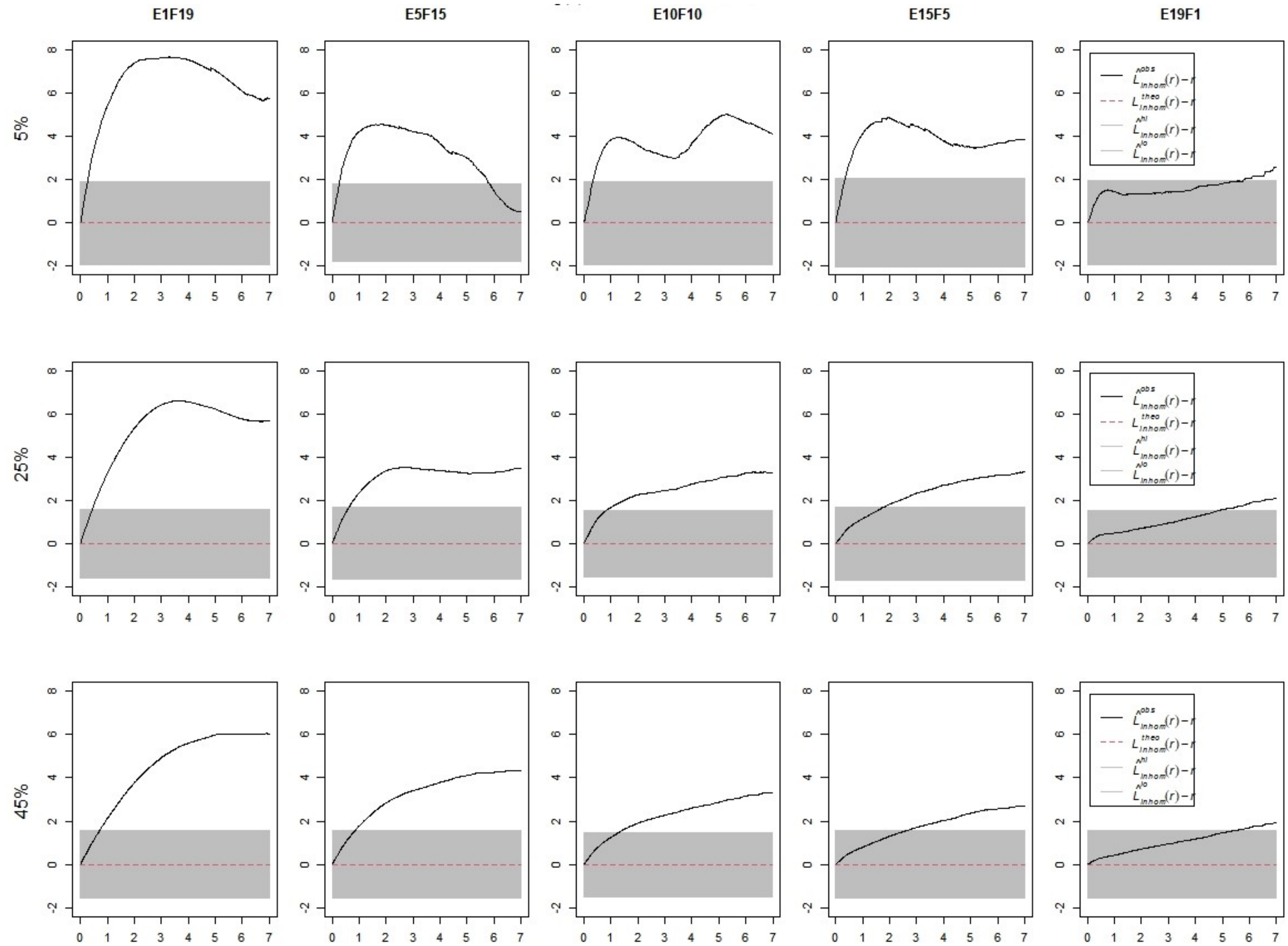

Figure 3: *Estimates of $\hat{L}(r) - r$ for observed patterns of points from a set of 5000 target points appear as the solid black line. Each row represents the proportion of observed points: 5%, 25% and 45%. Each column represents a different ratio of explorers and followers: 1:19, 5:15, 10;10, 15:5, and 19:1. The red line represents the theoretical clustering of an inhomogeneous Poisson process. 95% confidence bounds are shaded in gray.*

3.2 Correcting for observer bias

3.2.1 Comparison of the correction methods under the Warton et al. paradigm

First we evaluate how the different correction methods (density-based and knn-based) as well as the proxy (point clusters, roads network or both) influence the predictive performances for each ratio of explorers to followers. In Figure 4 we set the value of α to 1 as in Warton et al. [2013] and our method parameters such that k=1 (when the knn-based correction is used) and σ=1 (when the density based correction is used). As better predictions lead to lower NIMSE values, a negative difference indicates that the bias-corrected model outperforms the non-corrected model, while a positive difference indicates that the bias-corrected model underperforms the non-corrected model as explored in the appendix.

As the proportion of observed targets increases, the models which incorporate bias correction perform increasingly well in comparison to the uncorrected models. This is evident from the increasing negative differences in NIMSE shown in Figure 4 as well as the increasing positive differences in correlation shown in Appendix **??**. The corrections which used only the road network bias as a proxy performed the worst, while the methods which modelled observer bias using point density or a combination of point density and distance from the road network performed the best.

The benefit of the knn-based corrected models is most notable with a higher proportion of followers. More generally, with a higher proportion of explorers, the benefit of correcting for bias is small or non-existent. This is to be expected, as a higher proportion of explorers means fewer observers are searching near already observed points.

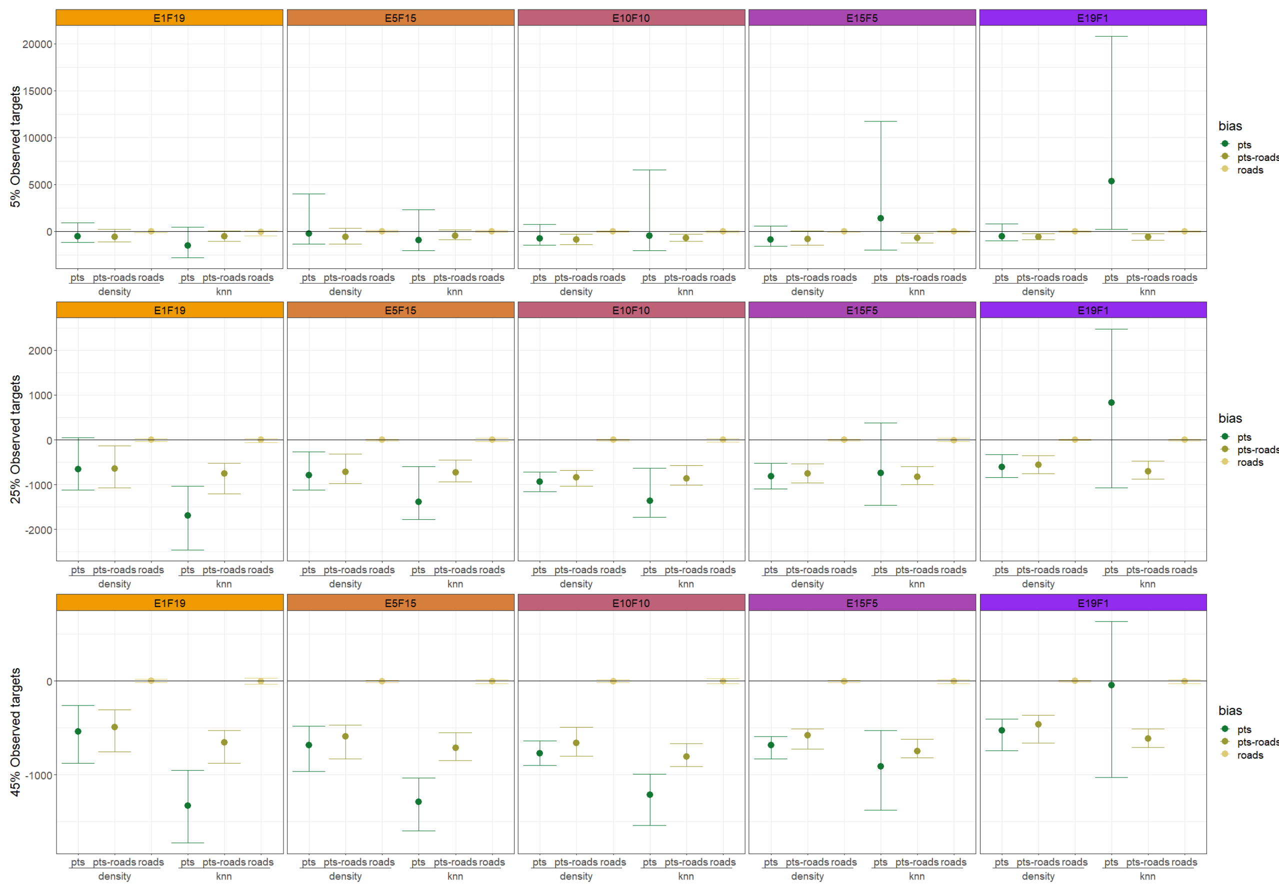

Figure 4: *Difference in NIMSE between bias-corrected predictions and non-corrected predictions for observed patterns from a set of 5000 target points. Here, bias correction is performed using a density-based or knn distance-based proxy variable. Lower values indicate better performance for the bias-corrected method. Each row represents the proportion of observed points: 5%, 25% and 45%. Each column represents a different ratio of explorers and followers: 1:19, 5:15, 10;10, 15:5, and 19:1. Each colored item differentiate the proxy variables (points only, point and roads and roads only). Each shape differentiate the set value for correction (minimum and null).*

3.2.2 Nearest neighbour distances-based correction

The best performing method with high clustering utilised a knn distance-based measure as a proxy for observer bias in Figure 4. Here, we investigate the performance for a range of different values of k and α. In Figure 5, the bias-corrected models tend to do best with values of α between 0.8 and 1. This is particularly true in the case of numerous followers. When the proportion of observed points decreases (5% and 25%), the benefit of bias correction shrinks with a higher ratio of explorers to followers, and even disappears with only 5% of target points observed. The choice of k also appears to have a greater effect with a smaller proportion of observed points and a higher ratio of explorers to followers. In particular, the performance of the bias-corrected model is worse when $k = 5$ or aggregates different numbers of nearest neighbors (1:2, 1:3, 1:5), particularly with 5% of observed targets. When the ratio of observers is dominated by explorers, larger values of k (3, 5) showed the best performance.

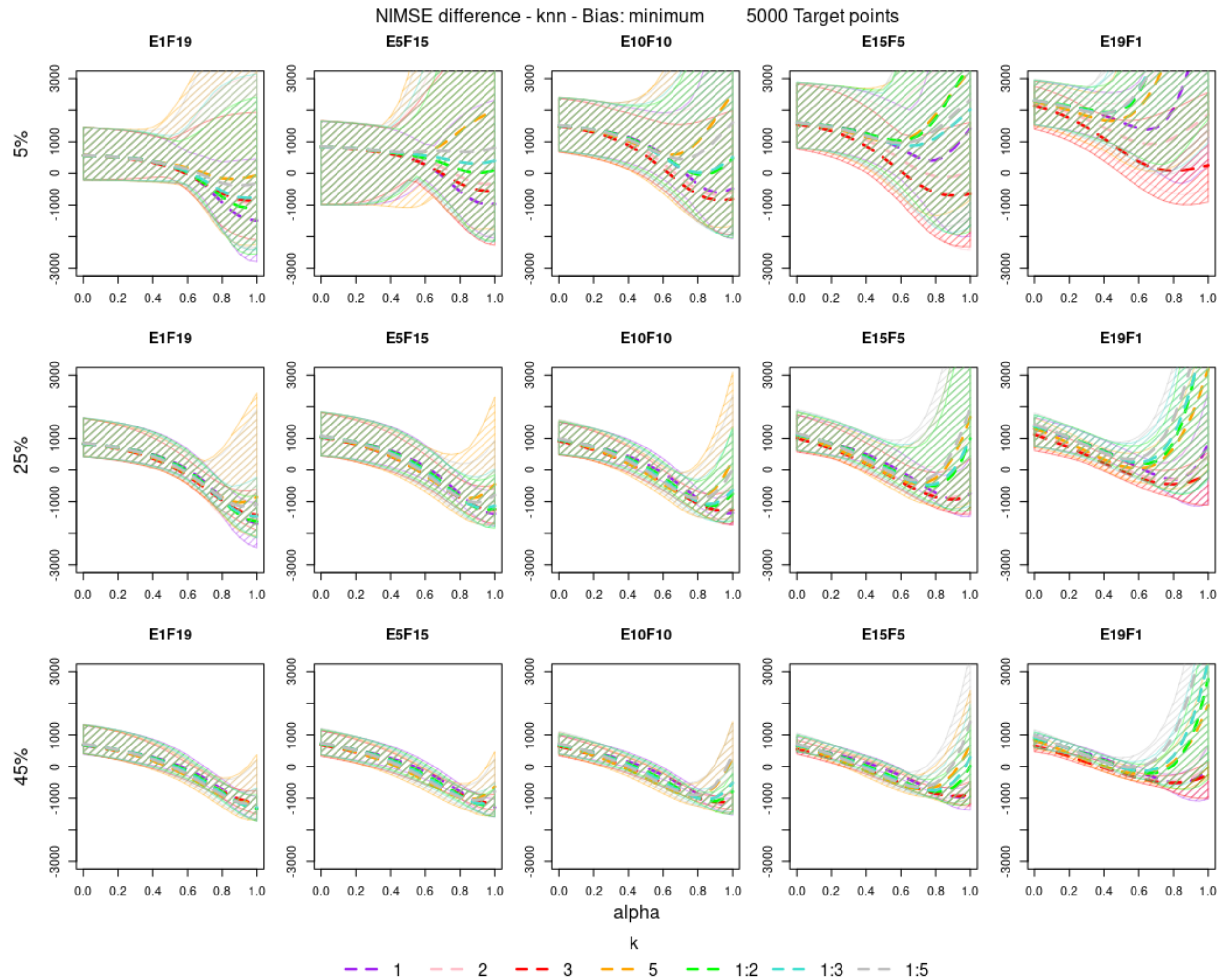

Figure 5: *Difference in NIMSE between bias-corrected predictions and non-corrected predictions for observed patterns from a set of 5000 target points. Here, bias correction is performed using a knn distance-based proxy variable. Lower values indicate better performance for the bias-corrected method. Each row represents the proportion of observed points: 5%, 25% and 45%. Each column represents a different ratio of explorers and followers: 1:19, 5:15, 10;10, 15:5, and 19:1. Each colored line represents a different number of nearest neighbours as presented in the plot legend.*

The analogous plot using correlation as a measure of performance is shown in the Appendix in Figure **??**. Because better predictions lead to higher correlations between the true and predicted intensity surfaces, a positive difference indicates that the bias-corrected model outperforms the non-corrected model whereas a negative difference indicates that the bias-corrected model underperforms the non-corrected model. The results are largely similar to those based on NIMSE. The bias-corrected model performs relatively best with higher values of α and when there are more followers, and there is greater variation in performance for different values of k with only 5% of the target points observed.

These conclusions are also apparent from the plots of predicted intensities in Figure 6. When $\alpha = 0$ as displayed on the left side of the figure, the bias-corrected models (first three rows) usually perform worse than the non-corrected models (fourth row) in comparison with the true intensity surface (fifth row). With only 5% of observed target points (first row), the signal is nearly imperceptible. When 25% or 45% of target points are observed (second and third rows), the signal becomes stronger, but still appears to lag behind the non-corrected model. We also note that the predicted intensity of the bias-corrected models appears closer to the true intensity when the number of explorers is higher.

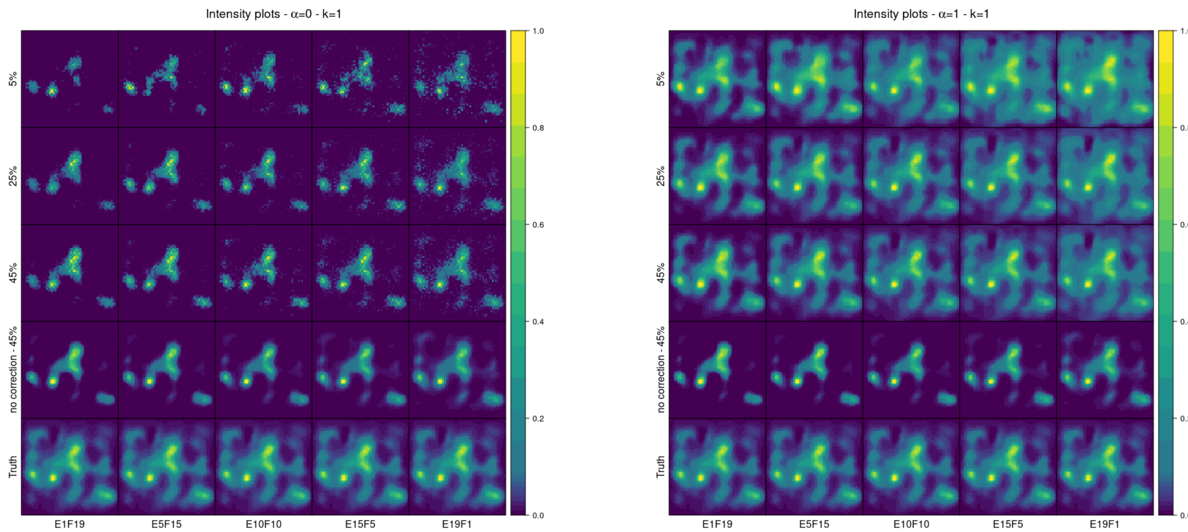

Figure 6: *Average predicted intensity maps for bias-corrected and non-corrected models based on 20 patterns observed from a set of 5000 target points. Here, the bias-correction was based on a knn distance-based proxy variable with* $\alpha = 0$ *(left) and* α*=1 (right) and* k*=3. The first three rows represent bias-corrected models with 5%, 25%, and 45% of target points observed. The fourth row represents the non-corrected model for 45% of target points observed, and the last row is the true species intensity. Each column represents a different ratio of explorers to followers: 1:19, 5:15, 10:10, 15:5, and 19:1.*

When $\alpha = 1$, as shown in Figure 6, the predicted intensity maps from the bias-corrected models (first three rows) more closely resemble the true species intensity ('Truth' on the fifth row) than the non-corrected models (fourth row), which tend to only highlight the high intensity areas. With $k = 1$, the corrected models had better performance (as reflected by a negative NIMSE difference) only for ratios that are balanced (10:10) or in favour of followers (1:19,5:15). For ratios that favour explorers, modelling the observer bias with $k = 1$ did not perform well, particularly with a smaller proportion of observed targets. In the predicted intensity maps, we indeed see that for these ratios, the correction over-represents areas of low intensity from the true species intensity.

4 Discussion

In this article, we implemented a virtual ecologist simulation tool `obsimulator` to study the impact of observer behaviour on the observed pattern of points. The simulator is designed to account for two types of observers' behaviour: explorers and followers. We have investigated differences in predictive performance with different bias correction approaches (knn distance-based, density-based, none) varying the ratio of explorers and followers (1:19, 5:15, 10:10, 15:5, 19:1), the methods' parameters (k, σ and α), and the proportion of target points observed (5%, 25%, 45%). We studied the spatial clustering of the patterns of points under these conditions using L-functions. The ratio of explorers and followers had a clear impact on spatial clustering, with greater clustering with a higher proportion of followers. This is expected, as followers select already-observed points as targets, leading to more clustered patterns. With more explorers, there are more observed targets for followers to select, leading to smaller clusters and larger distances.

To correct for observer bias, we extended the method of Warton et al. [2013]. This method uses proxy covariates to model observer bias and then corrects them by reducing the impact of these covariates. By setting $\alpha = 1$, the method is equivalent to that of Warton et al. [2013]. We chose to use two types of proxy covariates to model this bias — a knn distance-based measure and a density-based measure. The parameter α controls the degree of the correction, as shown in Equation (4). Regardless of the ratio of explorers and followers, the bias-corrected models perform best for very high levels of α between 0.8 and 1, when the correction is closest to that of Warton et al. [2013]. This suggests that the Warton et al. [2013] method of bias correction holds up well under various types of observer behaviour. The proxy variable is also an important factor to consider. Human infrastructure such as roads are a common bias proxy [Geldmann et al., 2016] but observer behaviour can reflect other choices such as moving towards known observations. This type of sampling bias is often not accounted for and can be corrected. We showed that accounting for observation distances to each other in context of high clustering is the best way to account for this observer behaviour.

Between the two correction methods (knn distance-based and density-based), the knn distance-based correction showed the best performance overall. The knn distance-based method of correction depends on parameters like the number of nearest neighbors considered and the metric used to calculate the distance between points. These parameters highly impact the algorithm's results [Guo et al., 2003, Wu et al., 2008, Weinberger and Saul, 2009]. In this context, it is clear that the value of k impacts the performance, particularly when measured with NIMSE. When the number of observed points is small, such as the case with 5% of observed target points, high values of k such as 3 had the best performance when there is less spatial clustering (i.e. more explorers than followers). This suggests that with smaller data sets with not much clustering, a better prediction of the amount of clustering is obtained with larger numbers of k. Consequently, when lacking data, using larger numbers of neighboring points provides a better estimate of bias. When the number of followers is greater than or equal to the number or explorers with only 5% of target points observed, smaller values of k perform best ($k = 1, 2$, or both 1 and 2). When we observe 25% or 45% of target points, values of $k = 1$ or $k = 3$ exhibited the best performance, particularly when there is less spatial clustering due to a higher proportion of explorers and thus the higher number of targets observed to inform future observers.

The density-based method of correction did not perform as well as the knn distance-based method overall, despite the fact that the knn distance-based method is based on a circular buffer area whereas the density-based method could allow for other clustering shapes. The performance of the density-based method depends on the type of smoothing kernel and the bandwidth choice. In this study, we have chosen Gaussian kernels and explored different values of the bandwidth parameter σ, but the `density.ppp` function allows for other kernel types such as Epanechnikov, quartic, or disc which could lead to improved performance. Differences in performance due to the choice of σ reflect the well-known problem of bandwidth selection between over and under smoothing Chen [2017]. Although the density-based correction method also suffered from scaling issues in the predicted intensity maps, the performance measures NIMSE and Pearson correlation are invariant to scale.

Simulations provide great tools to understand and study ecological processes. The `obsimulator` software allows users to vary observer behaviour under an explorer/follower paradigm. We have shown how clustering can be detected and how explorers and followers can change the pattern of observed points. Through this work, it is clear that identifying and correcting for observer bias leads to better predictions than not correcting for it, and that the best performance comes with a magnitude of correction akin to that of Warton et al. [2013], and that the benefit of this correction is greater with higher amounts of clustering.

Nonetheless, more complicated observer behaviour is possible and could lead to different conclusions. Indeed, other methods of bias correction, possibly tailored based on perceived observer behaviour, could perform best through a more in-depth study of differing observer behaviour and notably by including temporal information. The methods presented here offer a new way to correct for clustering in a pattern by smoothing the predictions according to density-based or knn distance-based proxy covariates. An improved method could include a combination of knn and kernel density methods to reflect the true clustering attributes of the distribution Tran et al. [2006]. Although not covered in this article, the `obsimulator` software also allows users to specify different rates of errors in reporting. In addition, a physical obstacle may be incorporated into the simulation design to replicate settings in which travelers are constrained in their movement.

Through the use of `obsimulator` to create different spatial patterns arising from differing observers' behaviour, this work demonstrates good practice for researcher using citizen science data. The pattern of point observations in such opportunistic data is the result of observer behaviour and can lead to high sampling biases. The observers' choices of where to search commonly result in clustered patterns biased toward roads, cities, or known target locations that we can account for using the methods presented in this manuscript.

Ethics statement

This does not apply to our research.

Data accessibility statement

Example codes from the analyses will be made available upon review and submission. The 'Obsimulator' is available on github but not provided here for the review process (double blind).

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Eréndira Aceves-Bueno, Adeyemi S. Adeleye, Marina Feraud, Yuxiong Huang, Mengya Tao, Yi Yang, and Sarah E. Anderson. The accuracy of citizen science data: A quantitative review. *The Bulletin of the Ecological Society of America*, 98(4):278–290, 2017. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/bes2.1336. https://doi.org/10.1002/bes2.1336.

Robert P Anderson and Ali Raza. The effect of the extent of the study region on gis models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus nephelomys) in venezuela. *Journal of Biogeography*, 37(7):1378–1393, 2010. https://doi.org/10.1111/j.1365-2699.2010.02290.x.

Ofer Arazy and Dan Malkinson. A framework of observer-based biases in citizen science biodiversity monitoring: Semi-structuring unstructured biodiversity monitoring protocols. *Frontiers in Ecology and Evolution*, 9, 2021. ISSN 2296-701X. doi: 10.3389/fevo.2021.693602. URL https://www.frontiersin.org/articles/10.3389/fevo.2021.693602.

A. Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns Methodology and Applications with R*. Tauylor and Francis group, LLC, 2016. https://book.spatstat.org/.

Jan Beck, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19:10–15, 2014. https://doi.org/10.1016/j.ecoinf.2013.11.002.

Hugues P Benoît and Jacques Allard. Can the data from at-sea observer surveys be used to make general inferences about catch composition and discards? *Canadian Journal of Fisheries and Aquatic Sciences*, 66 (12):2025–2039, 2009. https://doi.org/10.1139/F09-116.

Robert A Boria, Link E Olson, Steven M Goodman, and Robert P Anderson. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological modelling*, 275:73–77, 2014. https://doi.org/10.1016/j.ecolmodel.2013.12.012.

Diana E Bowler, Netra Bhandari, Lydia Repke, Christoph Beuthner, Corey T Callaghan, David Eichenberg, Klaus Henle, Reinhard Klenke, Anett Richter, Florian Jansen, et al. Decision-making of citizen scientists when recording species observations. *Scientific Reports*, 12(1):11069, 2022. doi: https://doi.org/10.1038/s41598-022-15218-2. URL https://doi.org/10.1038/s41598-022-15218-2.

Eleanor D. Brown and Byron K. Williams. The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology*, 33(3):561–569, 2019. doi: https://doi.org/10.1111/cobi.13223.

URL https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.13223.

Avishek Chakraborty, Alan E Gelfand, Adam M Wilson, Andrew M Latimer, and John A Silander. Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5):757–776, 2011. https://doi.org/10.1111/j.1467-9876.2011.00769.x.

Yen-Chi Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1(1): 161–187, 2017. https://doi.org/0.1080/24709360.2017.1396742.

Jeffrey P. Cohn. Citizen science: Can volunteers do real research? *BioScience*, 58(3):192–197, 03 2008. ISSN 0006-3568. doi: 10.1641/B580303. https://doi.org/10.1641/B580303.

Paul B. Conn, James T. Thorson, and Devin S. Johnson. Confronting preferential sampling when analysing population distributions: diagnosis and model-based triage. *Methods in Ecology and Evolution*, 8(11):1535–1546, 2017. https://doi.org/10.1111/2041-210X.12803.

Janis Dickinson, Benjamin Zuckerberg, and David Bonter. Citizen science as an ecological research tool: Challenges and benefits. *Annual Review of Ecology and Systematics*, 41:149–172, 12 2010. https://doi.org/10.1146/annurev-ecolsys-102209-144636.

Peter J Diggle, Raquel Menezes, and Ting-li Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2):191–232, 2010. https://doi.org/10.1111/j.1467-9876.2009.00701.x.

Monica Dimson and Thomas W. Gillespie. Who, where, when: Observer behavior influences spatial and temporal patterns of inaturalist participation. *Applied Geography*, 153:102916, 2023. ISSN 0143-6228. doi: https://doi.org/10.1016/j.apgeog.2023.102916. URL https://www.sciencedirect.com/science/article/pii/S0143622823000474.

Ahmed El-Gabbas and Carsten F Dormann. Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. *Ecology and Evolution*, 8(4):2196–2206, 2018. https://doi.org/10.1002/ece3.3834.

Jane Elith and John Leathwick. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and distributions*, 13(3): 265–275, 2007. https://doi.org/10.1111/j.1472-4642.2007.00340.x.

William Fithian and Trevor Hastie. Finite-sample equivalence in statistical models for presence-only data. *The annals of applied statistics*, 7(4):1917, 2013. https://doi.org/10.1214/13-AOAS667.

William Fithian, Jane Elith, Trevor Hastie, and David A Keith. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438, 2015. doi: 10.1111/2041-210X.12242. http://europepmc.org/articles/PMC5102514.

Amy Freitag, Ryan Meyer, and Liz Whiteman. Strategies employed by citizen science programs to increase the credibility of their data. *Citizen Science: Theory and Practice*, 1(1), 2016. http://doi.org/10.5334/cstp.6.

Jonas Geldmann, Jacob Heilmann-Clausen, Thomas E. Holm, Irina Levinsky, Bo Markussen, Kent Olsen, Carsten Rahbek, and Anders P. Tøttrup. What determines spatial bias in citizen science? exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 22(11):1139–1149, 2016. doi: https://doi.org/10.1111/ddi.12477. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.12477.

Emy Paulette Guilbault. *Statistical Modelling of Species Distributions: from Bridging the Gap between Statistics and Ecology to Conservation Stakeholders' Challenges*. PhD thesis, The University of Newcastle, 2022. URL http://hdl.handle.net/1959.13/1476372.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. *Lect Notes Comput Sci*, 2888:986–996, 01 2003. https://doi.org/10.1007/978-3-540-39964-3_62.

Motoki Higa, Yuichi Yamaura, Itsuro Koizumi, Yuki Yabuhara, Masayuki Senzaki, and Satoru Ono. Mapping large-scale bird distributions using occupancy models and citizen data with spatially biased sampling effort. *Diversity and Distributions*, 21(1):46–54, 2015. https://doi.org/10.1111/ddi.12255.

Janine B. Illian, Sara Martino, Sigrunn H. Sørbye, Juan B. Gallego-Fernández, María Zunzunegui, M. Paz Esquivias, and Justin M. J. Travis. Fitting complex ecological point process models with integrated nested laplace approximation. *Methods in Ecology and Evolution*, 4(4):305–315, 2013. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210x.12017.

Alison Johnston, Daniel Fink, Wesley M Hochachka, and Steve Kelling. Estimates of observer expertise improve species distributions from citizen science data. *Methods in Ecology and Evolution*, 9(1):88–97, 2018. https://doi.org/10.1111/2041-210X.12838.

Alison Johnston, WM Hochachka, ME Strimas-Mackey, V Ruiz Gutierrez, OJ Robinson, ET Miller, T Auer, ST Kelling, and D Fink. Best practices for making reliable inferences from citizen science data: case study using ebird to estimate species distributions. *bioRxiv*, page 574392, 2019. https://doi.org/10.1101/574392.

Alison Johnston, Nick Moran, Andy Musgrove, Daniel Fink, and Stephen R Baillie. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 422:108927, 2020. https://doi.org/10.1016/j.ecolmodel.2019.108927.

Steve Kelling, Alison Johnston, Aletta Bonn, Daniel Fink, Viviana Ruiz-Gutierrez, Rick Bonney, Miguel Fernandez, Wesley M Hochachka, Romain Julliard, Roland Kraemer, et al. Using semistructured surveys to improve citizen science data for monitoring biodiversity. *BioScience*, 69(3):170–179, 2019. https://doi.org/10.1093/biosci/biz010.

Osamu Komori, Shinto Eguchi, Yusuke Saigusa, Buntarou Kusumoto, and Yasuhiro Kubota. Sampling bias correction in species distribution models by quasi-linear poisson point process. *Ecological Informatics*, 55: 101015, 2020. https://doi.org/10.1016/j.ecoinf.2019.101015.

Margaret Kosmala, Andrea Wiggins, Alexandra Swanson, and Brooke Simmons. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560, 2016. https://doi.org/10.1002/fee.1436.

Louise Mair and Alejandro Ruete. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PloS one*, 11(1):e0147796, 2016. https://doi.org/10.1371/journal.pone.0147796.

Louise Mair, Philip J Harrison, Mari Jönsson, Swantje Löbel, Jenni Nordén, Juha Siitonen, Tomas Lämås, Anders Lundström, and Tord Snäll. Evaluating citizen science data for forecasting species responses to national forest management. *Ecology and evolution*, 7(1):368–378, 2017. https://doi.org/10.1002/ece3.2601.

Cory Merow, Jenica M Allen, Matthew Aiello-Lammens, and John A Silander Jr. Improving niche and range estimates with maxent and point process models by integrating spatially explicit information. *Global Ecology and Biogeography*, 25(8):1022–1036, 2016. https://doi.org/10.1111/geb.12453.

Christine N. Meynard, Boris Leroy, and David M. Kaplan. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? *Ecography*, 42(12):2021–2036, 2019. https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.04385.

Krishna Pacifici, Brian J Reich, David AW Miller, Beth Gardner, Glenn Stauffer, Susheela Singh, Alexa McKerrow, and Jaime A Collazo. Integrating multiple data sources in species distribution modeling: a framework for data fusion. *Ecology*, 98(3):840–850, 2017. https://doi.org/10.1002/ecy.1710.

Steven J Phillips and Miroslav Dudík. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175, 2008. https://doi.org/10.1111/j.0906-7590.2008.5203.x.

Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197, 2009. https://doi.org/10.1890/07-2153.1.

Laura J Pollock, Reid Tingley, William K Morris, Nick Golding, Robert B O'Hara, Kirsten M Parris, Peter A Vesk, and Michael A McCarthy. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm). *Methods in Ecology and Evolution*, 5(5):397–406, 2014. https://doi.org/10.1111/2041-210X.12180.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Nathan Ranc, Luca Santini, Carlo Rondinini, Luigi Boitani, Françoise Poitevin, Anders Angerbjörn, and Luigi Maiorano. Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, 40(9):1076–1087, 2017. https://doi.org/10.1111/ecog.02414.

Ian W Renner, Julie Louvrier, and Olivier Gimenez. Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization.

Methods in Ecology and Evolution, 10(12):2118–2128, 2019. https://doi.org/10.1111/2041-210X.13297.

Ian Walton Renner. *Advances in presence-only methods in ecology.* PhD thesis, UNSW Sydney, 2013. URL http://hdl.handle.net/1959.4/52837.

Kevin C Rose, Patrick J Neale, Maria Tzortziou, Charles L Gallegos, and Thomas E Jordan. Patterns of spectral, spatial, and long-term variability in light attenuation in an optically complex sub-estuary. *Limnology and Oceanography*, 64(S1):S257–S272, 2019. https://doi.org/10.1002/lno.11005.

Page Segurado, Miguel B Araújo, and We Kunin. Consequences of spatial autocorrelation for niche-based models. *Journal of Applied Ecology*, 43(3):433–444, 2006. https://doi.org/10.1111/j.1365-2664.2006.01162.x.

Jonathan Silvertown. A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9):467 – 471, 2009. ISSN 0169-5347. URL http://www.sciencedirect.com/science/article/pii/S016953470900175X. https://doi.org/10.1016/j.tree.2009.03.017.

Anja Skroblin, Tracy Carboon, Gladys Bidu, Nganjapayi Chapman, Minyawu Miller, Karnu Taylor, Waka Taylor, Edward T Game, and Brendan A Wintle. Including indigenous knowledge in species distribution modelling for increased ecological insights. *Conservation Biology*, 2019. https://doi.org/10.1111/cobi.13373.

Jessica Stolar and Scott E Nielsen. Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, 21(5):595–608, 2015. https://doi.org/10.1111/ddi.12279.

Jan W. H. Swanepoel. Mean intergrated squared error properties and optimal kernels when estimating a distribution function. *Communications in Statistics - Theory and Methods*, 17(11):3785–3799, 1988. ISSN 0361-0926 1532-415X. https://doi.org/10.1080/03610928808829835.

Ying Tang, Julie A Winkler, Andrés Viña, Fang Wang, Jindong Zhang, Zhiqiang Zhao, Thomas Connor, Hongbo Yang, Yuanbin Zhang, Xiaofeng Zhang, et al. Expanding ensembles of species present-day and future climatic suitability to consider the limitations of species occurrence data. *Ecological Indicators*, 110:105891, 2020. https://doi.org/10.1016/j.ecolind.2019.105891.

Thanh N. Tran, Ron Wehrens, and Lutgarde M.C. Buydens. Knn-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics & Data Analysis*, 51(2):513–525, 2006. ISSN 0167-9473. https://www.sciencedirect.com/science/article/pii/S0167947305002537.

Andrew J. Tyre, Brigitte Tenhumberg, Scott A. Field, Darren Niejalke, Kirsten Parris, and Hugh P. Possingham. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, 13(6):1790–1801, 2003. doi: https://doi.org/10.1890/02-5078. URL https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/02-5078.

Julien Vollering, Rune Halvorsen, Inger Auestad, and Knut Rydgren. Bunching up the background betters bias in species distribution models. *Ecography*, 42, 06 2019. https://doi.org/10.1111/ecog.04503.

Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994. doi: https://doi.org/10.1201/b14876. URL https://doi.org/10.1201/b14876.

David I Warton, Ian W Renner, and Daniel Ramp. Model-based control of observer bias for the analysis of presence-only data in ecology. *PloS one*, 8(11):e79168, 2013. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). https://doi.org/10.1371/journal.pone.0079168.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2):207–244, 2009. https://doi.org/10.1145/1577069.1577078.21,38.

Thorsten Wiegand and Kirk A Moloney. *Handbook of spatial point-pattern analysis in ecology*. CRC press, 2013. doi: https://doi.org/10.1201/b16195. URL https://doi.org/10.1201/b16195.

Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008. https://doi.org/10.1007/s10115-007-0114-2.

A.Elizabeth Zaniewski, Anthony Lehmann, and Jacob McC Overton. Predicting species spatial distributions using presence-only data: a case study of native new zealand ferns. *Ecological Modelling*, 157(2):261–280, 2002. ISSN 0304-3800. doi: https://doi.org/10.1016/S0304-3800(02)00199-0. URL https://www.sciencedirect.com/science/article/pii/S0304380002001990.

Damaris Zurell, Uta Berger, Juliano S. Cabral, Florian Jeltsch, Christine N. Meynard, Tamara Münkemüller, Nana Nehrbass, Jörn Pagel, Björn Reineking, Boris Schröder, and Volker Grimm. The virtual ecologist approach: simulating data and observers. *Oikos*, 119(4):622–635, 2010. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0706.2009.18284.x.