Classification of unlabeled observations in Species Distribution Modelling using Point Process Models.

Emy Guilbault[1], *Ian Renner*[1], *Michael Mahony*[2], *Eric Beh*[1].

Emy.Guilbault@uon.edu.au

April 16, 2019

[1] *School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia.*

[2] *School of Environmental and Life Sciences, University of Newcastle, Callaghan, NSW, Australia.*

1 Abstract

1. Species distribution modelling, which allows users to predict the spatial distribution of species with the use of environmental covariates, has become increasingly popular, with many software platforms providing tools to fit species distribution models. However, the species observations used in species distribution models can have varying levels of quality and can have incomplete information, such as uncertain species identity.

2. In this paper, we develop two algorithms to reclassify observations with unknown species identities which simultaneously predict different species distributions using spatial point processes. We compare the performance of the different algorithms using different initializations and parameters with models fitted using only the observations with known species identity through simulations.

3. We show that performance varies with differences in correlation among species distributions, species abundance, and the proportion of observations with unknown species identities. Additionally, some of the methods developed here outperformed the models that didn't use the misspecified data.

4. These models represent an helpful and promising tool for opportunistic surveys where misidentification happens or for the distribution of species newly separated in their taxonomy.

Keywords: Presence-only data - Ecological statistics - Misidentification - Classification - Mixture modelling - EM algorithm - Machine learning

2 Introduction and background

Species distribution modelling has been a popular topic in ecological statistics over the past decade. Many tools and methods have been developed to provide a means to explore the distributions of species through mapping of suitable environments (Jewell *et al.*, 2007; Peterman *et al.*, 2013; Nezer *et al.*, 2016; Inoue *et al.*, 2017; Schank *et al.*, 2017). Although there are a large number of algorithms and software platforms that can fit species distribution models (SDMs), generalization of these methods and specific applications to real data sets can be tricky (Burnham & Anderson, 2002; Aarts *et al.*, 2012; Guillera-Arroita *et al.*, 2015).

The most common sources of species information used in SDMs are presence-only (PO) and presence-absence (PA) data. PO data only contains information about species presence, in contrast to PA data which records both where species have been found present and where they have not been found (Warton & Shepherd, 2010; Renner *et al.*, 2015). Although PA data is generally of higher quality, it is also less common than PO data because it requires more rigorous planning to visit a set of pre-determined sites. On the other hand, PO data sets are very common, arising from surveys or opportunistic sightings, but they usually have lower quality (van Strien *et al.*, 2013; Ruete & Leynaud, 2015). Point process models (PPMs) are a common tool for fitting SDMs to analyze PO data (Warton & Shepherd, 2010; Mi *et al.*, 2014; Renner *et al.*, 2015) and have been used to fit models for real datasets and simulated data (Baddeley *et al.*, 2006; Illian *et al.*, 2012; Renner & Warton, 2013; Baddeley *et al.*, 2015).

Unreliable or unknown species observation identification is also a main concern in ecology. For example, species records can become confounded when species taxonomy changes (Mahony *et al.*, 2006). Conservation planning efforts depend on clear identification of species and understanding of their distributions and habitat requirements (Franklin, 2013; Guisan *et al.*, 2013). Such concerns are very rarely considered while building SDMs, as people usually clean the data or make some assumptions to avoid such identification problems.

Mixture modelling is a common tool used to represent complex distributions and aims to identify different groups within a dataset while modelling heterogeneity (Martinez, 2015). In communities or groups of individuals/species it is possible to classify or cluster them according to covariate information by using finite mixture modelling (McLachlan & Peel, 2000; Frame & Jammalamadaka, 2007; Dunstan *et al.*, 2013; Fernández-Michelli *et al.*, 2016). One particular application of this approach is to deal with over-dispersed data and to model the different ecological processes at the same time for a single species or for different species in order to classify them (Matthews *et al.*, 2001; Zhang *et al.*, 2004; Tracey *et al.*, 2013).

Machine learning algorithms are also becoming more common in statistical ecology because they can deal with unknown information and recognize some structure in the data (Hastie *et al.*, 2001; Thessen, 2016; Browning *et al.*, 2018). Some algorithms can group observations with similar characteristics (unsupervised learning) and some use separate labeled datasets (supervised learning) or partially labeled data within the studied dataset (semi-supervised learning) to classify the observations (Wendel *et al.*, 2015; Fernández-

Michelli *et al.*, 2016; Vo *et al.*, 2018; Zhou, 2018). Some recent publications have applied machine learning algorithms to fit PPMs in a Bayesian framework (Tran, 2017; Vo *et al.*, 2018), but the literature on using machine learning algorithms to fit PPMs is not yet well-developed. Additionally, several R packages have been developed to deal with machine learning procedures (Benaglia *et al.*, 2009; Iovleff, 2018), but none accommodate the intersection of point process modelling with mixture modelling or machine learning algorithms.

In this paper we develop new tools for fitting models to multi-species PO data with partial species identification by combining the PPM framework with mixture modelling and machine learning approaches to accommodate incomplete labelling. These tools implement two algorithms to reclassify the unreliable observations to belong to one of the existing species. The first tool fits mixtures of PPMs to all available data with an Expectation-Maximization (EM) algorithm and uses them to classify the unreliable points. This method will be called *Mixture method*. The second tool employs an iterative technique to fit separate PPMs to points with known labels augmented by some points with unknown labels depending on classification probabilities at each iteration. This method will be hereafter known as the *Loop method*. Using simulations, we compare the performance in classification and prediction for the proposed algorithms to the simple, standard approach of fitting individual PPMs to the points with known species labels only. We found that performance varied based on the choice of initialization and algorithm parameters but some of the methods can outperform the fitting of individual PPMs.

3 New modelling methods

3.1 Notation

The fitted point process models in our proposed methods make use of a total of $M + N + Q$ locations as follows:

Let $\mathbf{s}_1 = \{s_1, \ldots, s_{m_1}\}$, $\mathbf{s}_2 = \{s_{m_1+1}, \ldots, s_{m_1+m_2}\}$, $\ldots$, $\mathbf{s}_K = \{s_{M-m_K+1}, \ldots, s_M\}$ be vectors that contain all of the observed locations with known species identities $1, 2, \ldots, K$, respectively. These are represented by the orange, purple, and turquoise dots in Figure 1 for a hypothetical dataset. Let $|\mathbf{s}_1| = m_1, |\mathbf{s}_2| = m_2, \ldots, |\mathbf{s}_K| = m_K$ be the number of observed locations with known species identity for each of the K species. We collect the $M = m_1 + m_2 + \ldots + m_K$ total locations with known species identities of all K species in $\mathbf{s} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K\}$. Let $\mathbf{u} = \{s_{M+1}, \ldots, s_{M+N}\}$ contain the N observed locations with uncertain species identities. These are represented by the black question marks in Figure 1.

Let $\mathbf{q} = \{s_{M+N+1}, \ldots, s_{M+N+Q}\}$ contain the locations of Q quadrature points placed along a regular $c_1 \times c_2$ grid throughout the study region (Figure 1). Each quadrature point is placed at the center of one of Q unique rectangular grid cells throughout the study region. Let $c(s)$ be the grid cell in which location

s is contained.

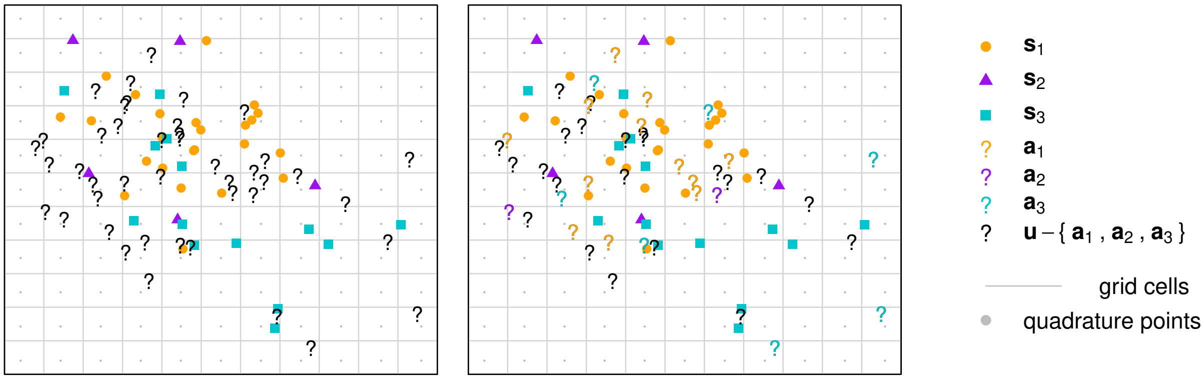

Figure 1: Three illustrative point patterns. The orange, purple, and turquoise colored dots represent locations with known species identity, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$. The gray dots represent quadrature points $\mathbf{q}$, which are spaced evenly along a regular grid such that one quadrature point is at the centre of each rectangular grid cell. The black question marks (left) represent observed locations $\mathbf{u}$ with uncertain species identity. The locations in $\mathbf{a}_1 \in \mathbf{u}$, $\mathbf{a}_2 \in \mathbf{u}$, and $\mathbf{a}_3 \in \mathbf{u}$ which are reclassified as belonging to one of the species are represented by coloured question marks (right).

3.2 Loop methods

The three loop algorithms proceed by iterating between steps that augment the vectors of locations with known species identities $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K$ with locations $\mathbf{a}_1 \subset \mathbf{u}, \mathbf{a}_2 \subset \mathbf{u}, \ldots, \mathbf{a}_K \subset \mathbf{u}$, update the quadrature weights, and fit point process models as follows:

1. Fit K initial point process models using the vectors of observed locations with known species identity $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_K$.

2. Compute the predicted intensities $\widehat{\mu}_i(s)$ for all $s \in \{\mathbf{s} \cup \mathbf{u}\}$ for $i \in \{1, \ldots, K\}$.

3. Derive an $(M + N) \times K$ matrix of membership probabilities $\boldsymbol{\omega}$, where

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_1(s_1) & \omega_2(s_1) & \ldots & \omega_K(s_1) \\ \omega_1(s_2) & \omega_2(s_2) & \ldots & \omega_K(s_2) \\ \vdots & \vdots & \ldots & \vdots \\ \omega_1(s_{M+N}) & \omega_2(s_{M+N}) & \ldots & \omega_K(s_{M+N}) \end{bmatrix}$$

The membership probability of location s for species i is defined as

$$\omega_i(s) = \begin{cases} \mathbb{1}(s \in \mathbf{s}_i) & : s \in \mathbf{s} \\ \frac{\widehat{\mu}_i(s)}{\sum_{j=1}^{K} \widehat{\mu}_j(s)} & : s \in \mathbf{u}. \end{cases} \tag{1}$$

That is, the membership probabilities for the locations with known species identity are 1 for the correct species and 0 otherwise, and for the locations with unknown species identity, they are proportional to the fitted intensities.

4. Define an augmented vector for species i as $\mathbf{y}_i = \mathbf{s}_i \cup \mathbf{a}_i$ for all $i \in \{1, \ldots, K\}$. We define $\mathbf{a}_i$ as follows:

    - For the **Normal** method, $\mathbf{a}_i = \mathbf{u}$ (left panel of Figure 2).
    - For the **Loop grW** method, $\mathbf{a}_i = \mathbf{u}_{[\omega_i(s) \geq \delta]}$, where δ is a minimum membership probability threshold that takes the following values successively at each iteration $\{\delta_{\max}, \delta_{\max} - \delta_{\text{step}}, \ldots, \delta_{\min}\}$. That is, the Loop grW method augments the locations with known species identity i with the locations with unknown species identity with membership probabilities for species i that are higher than the current threshold δ (middel panel of Figure 2).
    - For the **Loop hgW** method, $\mathbf{a}_i = \mathbf{u}_{[\omega_i(s) \geq \boldsymbol{\omega}_{i,(M+N-a+1)}]}$, where $\boldsymbol{\omega}_{i,(j)}$ represents the j^{th} smallest entry of vector $\boldsymbol{\omega}_i$, the i^{th} column of $\boldsymbol{\omega}$, and a represents the number of locations to be augmented. We set a to be the same integer for all K species for some a between 1 and $\lfloor \frac{N}{K} \rfloor$ then at each iteration a is increased by one (right panel of Figure 2).

5. Update the quadrature weights for each species. First, assign each location in $\{\mathbf{y}_1, \ldots, \mathbf{y}_K, \mathbf{q}\}$ to a grid cell. Then, compute the vector of quadrature weights $\mathbf{w}_i$ for all points $t \in \{\mathbf{y}_i \cup \mathbf{q}\}$ as follows:

$$w_i(t) = \frac{c_1 \times c_2 \times \omega_i(t)}{1 + \sum_{s \in \{\mathbf{y}_i \cup \mathbf{q}\}} \mathbb{1}(c(s) = c(t))\omega_i(s)}. \tag{2}$$

    This way of computing quadrature weights is an extension of standard quadrature weight schemes for point process models (Berman & Turner, 1992), in which the weight for location s is equal to the area of the grid cell $c(s)$ that contains s divided by the total number of quadrature and observed locations in $c(s)$. Here, we divide the area of the grid cell by the sum of the membership probabilities of the observed locations in the grid cell (both with and without known species identities) plus 1 (for the one quadrature point in the grid cell).

6. Fit point process models using the augmented vector $\mathbf{y}_i$, quadrature points $\mathbf{q}$ and quadrature weights $\mathbf{w}_i$ for all species $i \in \{1, \ldots, K\}$.

7. Return to step 2 and stop when we either reach likelihood convergence or we reach a maximum number of iterations that is different depending on the method chosen. Likelihood convergence is determined by:

$$\delta_l = \frac{\sum_{j=1}^{K} \left| \ell^j_{h+1}(\boldsymbol{\beta}) - \ell^j_h(\boldsymbol{\beta}) \right|}{\left(\sum_{j=1}^{K} \ell^j_h(\boldsymbol{\beta}) \right)} < \epsilon \tag{3}$$

    for some choice of ϵ, where $\ell(\boldsymbol{\beta})^j_h$ is the fitted log-likelihood for the j^{th} species at the h^{th} iteration. The maximum number of iterations varies for the different methods, as follows:

- For the **Normal** method, the maximum number of iterations is set by the user. We set the default number of iterations to be 50.
- For the **Loop grW** method, the maximum number of iterations is determined by the choice of δ_{max}, δ_{step}, and δ_{min}.
- For the **Loop hgW** method, the maximum number of iterations is $\lfloor \frac{N}{K} \rfloor - a_1$, where $\lfloor c \rfloor$ rounds the number c down to the nearest integer, and a_1 is the first value of a chosen by the user. In the case of decimals numbers, only the floor is considered as the we can't add more points than available per species.

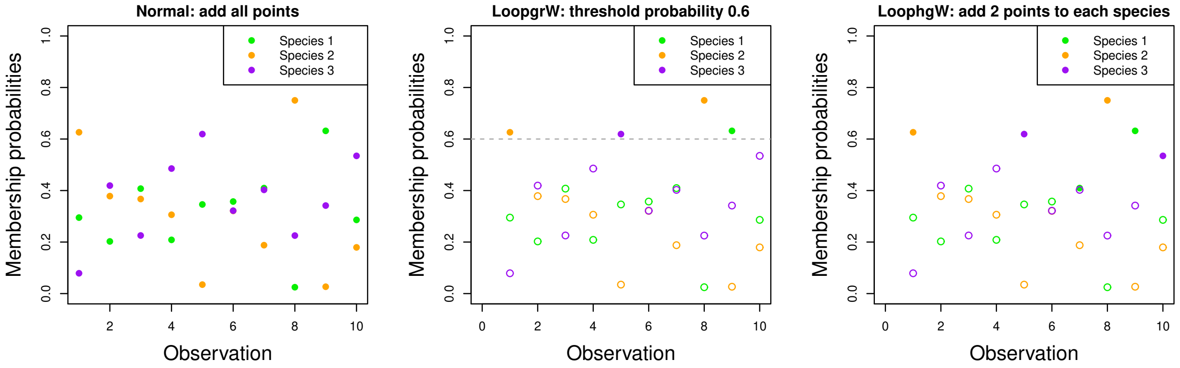

Figure 2: (Left) Normal Loop function. We add all points with unknown species labels to each species, using membership weights that are proportional to the fitted intensities. (Middle) Method Loop grW function. We add all points with membership probabilities greater than a threshold δ_{max}, then we decreases from that value to a minimum of δ_{min} by increments of δ_{step}. (Right) Method Loop hgW function. We add the a points with highest membership probabilities to each species, increasing the number a from 1 to $\lfloor \frac{N}{K} \rfloor$.

3.3 Mixture of PPMs method

The four mixture algorithms can be fitted by maximizing a log-likelihood function and reclassifying the locations with uncertain identity using an EM algorithm framework. The algorithm proceeds as follows:

1. We initialize the membership probabilities $\boldsymbol{\omega}$ for each location s for each species i in one of the following ways:

   - For the **knn method**, we calculate the distance $d_i(s)$ of each location s to the k^{th} nearest neighbor of species i, for all K species. We calculate the membership probability of location s for species i using:

$$\omega_i(s) = \begin{cases} \mathbb{1}(s \in \mathbf{s}_i) & : s \in \mathbf{s} \\ \frac{z_i(s)}{\sum_{j=1}^{K} z_j(s)} & : s \in \mathbf{u}. \end{cases} \tag{4}$$

where

$$z_i(s) = \frac{\min_{1 \le j \le K} d_j(s)}{d_i(s)} \tag{5}$$

- For the **kmeans method**, we define $\omega_i(s)$ as in (4) but define $z_i(s)$ as

$$z_i(s) = \frac{\min_{1 \le j \le K} d_j^{\mathrm{C}}(s)}{d_i^{\mathrm{C}}(s)}, \tag{6}$$

  where $d_i^{\mathrm{C}}(s)$ is the distance to the i^{th} centroid of the i^{th} cluster.

- For the **random method**, we define $\omega_i(s)$ as in (4) and $z_i(s)$ is drawn randomly from a uniform distribution:

$$z_i(s) \sim U[0, 1] \tag{7}$$

- For the **equal method**, we assign equal membership probabilities for the locations with uncertain identity:

$$\omega_i(s) = \begin{cases} \mathbb{1}(s \in \mathbf{s}_i) & : s \in \mathbf{s} \\ \frac{1}{K} & : s \in \mathbf{u}. \end{cases} \tag{8}$$

Regardless of the initialization method, the sum of membership probabilities across the all species is equal to 1 for all points.

2. Classify the locations in $\mathbf{u}$ to belong to one of the K species based on the membership probabilities $\boldsymbol{\omega}$.

3. Fit a point process model using a marked point pattern, where each observation s has a mark defined by the known or classified identity among the K species.

4. Compute the predicted intensities $\widehat{\mu}_i(s)$ for all $s \in \{\mathbf{s} \cup \mathbf{u}\}$ for $i \in \{1, \ldots, K\}$.

5. E step: We first get the predicted values of each species at the locations $s \in \{\mathbf{s} \cup \mathbf{u}\}$ and calculate the predicted intensity of the mixture of K densities using:

$$f(s) = \sum_{i=1}^{K} \pi_i \times f_i(s), \tag{9}$$

   where $f_i(s)$ is the density at location s for the i^{th} component and π_i is the mixing proportion or weight of the i^{th} species in the mixture.

6. We calculate new membership probabilities for each unknown point of $\mathbf{u}$ using:

$$\boldsymbol{\omega}_i(s) = \frac{\hat{\mu}_i(s)}{\sum_{i=1}^{k} \hat{\mu}_i(s)}, \tag{10}$$

where $\mu_i(s)$ is the intensity of the i^{th} species at location $s \in \mathbf{s}$. For the observations $\mathbf{s}$ with known labels, the membership probabilities are set to 1 for the correct species label and 0 otherwise.

7. M step: Classify the locations in $\mathbf{u}$ to belong to one of the K species. The classification for each point s corresponds to the highest membership probability $\omega_i(s)$ for $i \in \{1, \ldots, K\}$. We compute each species' proportion of the whole by summing the membership probabilities for each species across both $\mathbf{s}$ and $\mathbf{u}$.

8. Compute a marked PPM based on the updated classifications and membership probabilities.

9. Calculate the model log likelihood using:

$$\ell(\boldsymbol{\beta}) = \sum_{s \in \mathbf{s} \cup \mathbf{u}} f(s, \boldsymbol{\beta}) = \sum_{s \in \mathbf{s} \cup \mathbf{u}} \log \sum_{i=1}^{K} \pi_i \times f(s, \beta_i) \tag{11}$$

10. Repeat steps 4-9 until we achieve likelihood convergence, defined as follows:

$$\frac{|\ell_{h+1}(\boldsymbol{\beta}) - \ell_h(\boldsymbol{\beta})|}{(1 + |\ell_{h+1}(\boldsymbol{\beta})|)} < \epsilon \tag{12}$$

where $\ell_h(\boldsymbol{\beta})$ is the log-likelihood at the h^{th} iteration and ϵ is a pre-specified tolerance level.

4 Simulation framework

4.1 Simulation data

To compare the performance of the different algorithms, we simulated patterns $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$ of individuals for three species based on "true" distributions defined by four different predictors. Because performance could varied based on sample size, the correlations $\rho_{i,j}$ among the species distributions, and the proportion of observations with unknown labels, we consider similar and different low abundances by randomly simulating numbers of points between 20 and 50 for the species as well as the correlation between the true species distributions:

- Case 1: at least two species i and j have distributions that are highly correlated ($|\rho_{i,j}| \geq 0.85$ for some $i, j \in \{1, 2, 3\}$)
- Case 2: no two species have highly correlated distributions ($|\rho_{i,j}| < 0.45$ for all $i, j \in \{1, 2, 3\}$)

We chose these values for abundances as they would be small enough such that potential value of adding points with unknown species identities could be investigated, and we chose these cutoffs for correlation to create clearly distinguishable contexts.

We then created locations with unknown labels $\mathbf{u}$ by hiding uniformly at random a certain proportion of the total observations (20%, 50% and 80%). The locations in $\mathbf{t}_1$, $\mathbf{t}_2$, and $\mathbf{t}_3$ that retained their true species identities therefore became the simulated point patterns $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ with known species identities.

Simulations were conducted using the version 3.4.2 of R (R Core Team, 2017) and used high performance computing to implement 1000 simulations each for different combinations of abundances, correlation among species distributions, and proportions of observations with unknown labels. We also tested different parameters for the knn initialization of the mixture algorithm (the value of k neighbors), the Loop grW function (the maximum threshold δ_{max}, minimum threshold δ_{min} and the step size δ_{step}) and the Loop hgW function (initial number of points added to the point pattern a).

4.2 Suite of Evaluation tools

We consider various measures of performance for comparing the distributions. For classification methods, misclassification/accuracy analysis is a common measure of performance (Wendel *et al.*, 2015).We choose the highest mixing weight for each observation to determine the labeling when computing accuracy. We also compared the final membership probabilities of the correct labels of each point to 1 (the true weight) with a residual sum of squares (RSS).

$$\text{RSS} = \sum_{i=1}^{K} \sum_{s \in \mathbf{t}_i} (\omega_i(s) - 1)^2, \tag{13}$$

where $\omega_i(s)$ is the final membership probability for location s for the correct species i computed using the methods outlined in sections 3.2 and 3.3. Considering residual sum of squares (RSS) alone does not provide a reliable comparison because the number of unknown observations can vary, so we consider meanRSS instead to standardize the measure for all fitted models:

$$\text{meanRSS} = \frac{\text{RSS}}{N}, \tag{14}$$

where N is the number of observations with uncertain species identities.

We also considered measures that compare the true distribution from which we generate the points to the predicted distributions of the model. We use a sum of correlations between the true and predicted distributions across all species (hereafter referred to as 'sumcor') to assess how well the predicted distributions align with the true distributions. We can use various correlation measures such as Pearson's correlation coefficient, Kendall's τ or Spearman's ρ when computing sumcor.

Another global measure of predictive performance of the intensity estimates is the Integrated Mean Square Error (IMSE) (Swanepoel, 1988; Es, 1997). The function is defined as:

$$\text{IMSE} = E\left(\int_{-\infty}^{+\infty} ((\hat{f}_n(x) - f(x))^2) dx\right), \tag{15}$$

where $f_n(x)$ is an estimator of the density function $f(x)$. We standardized this value by rescaling the intensities to be able to compare each methods even if different number of points are considered and compute the IMSE using the values of the true and predicted intensities at the quadrature points $\mathbf{q}$, and sum across the 3 species.

5 Results

Here we present the results of the simulations, with more detailed results appearing in the Appendix. In this section, we only present the results from the knn, Lopp grW, Loop hgW and individual PPM methods that displayed the best performances. First, we present the model performances from varying data parameters (abundance, correlation and percentage of hidden labeled data). The individual PPM results will be used as a point of comparison with the other methods as the individual method does not include any of the points with unknown labels. We, then, focus on varying model parameters in the different methods (the value of k for knn, the values of δ_{max}, δ_{min} and δ_{step} for Loop grW and the value of a for Loop hgW). For these results, we set $k = 1$, $\delta_{\text{max}} = 0.5$, $\delta_{\text{min}} = 0.1$, $\delta_{\text{step}} = 0.1$ and $a = 5$ according to the algorithm parameters tests presented in section 5.2. For the performance results, the sumcor methods displayed the result using the Pearson correlation coefficient.

5.1 Varying species distributions

5.1.1 Different abundances and correlated distributions

In Figure 3, we consider different low abundances ($m_1 = 32$, $m_2 = 42$ and $m_3 = 23$) and where two distributions are highly correlated. With regard to classification performance, the different modelling methods have similar levels of accuracy, although when comparing meanRSS, the individual and Loop grW methods seem to outperform the other methods, especially as we increase the proportion of hidden observations. With regard to predictive performance, the Loop grW method appears to have the greatest performance when measured by IMSE and sumcor, particularly for 50% and 80% of hidden observations. The Loop hgW method performs comparably to the individual PPM method, although its preformance gets relatively better as we increase the proportion of hidden observations. The knn method has the highest IMSE for 50% and 80% of hidden observations, but it is competitive with the individual PPM and loop hgW method when comparing sumcor. See Tables 1 and 2 in the Appendix for a comparison of means and medians across all of these measures.

When examining the predicted intensities with 80% of the observations with hidden species identities, the true pattern appears best captured by the Loop grW method (Figure 4), consistent with sumcor. The Loop hgW method tends to overpredict the intensities.

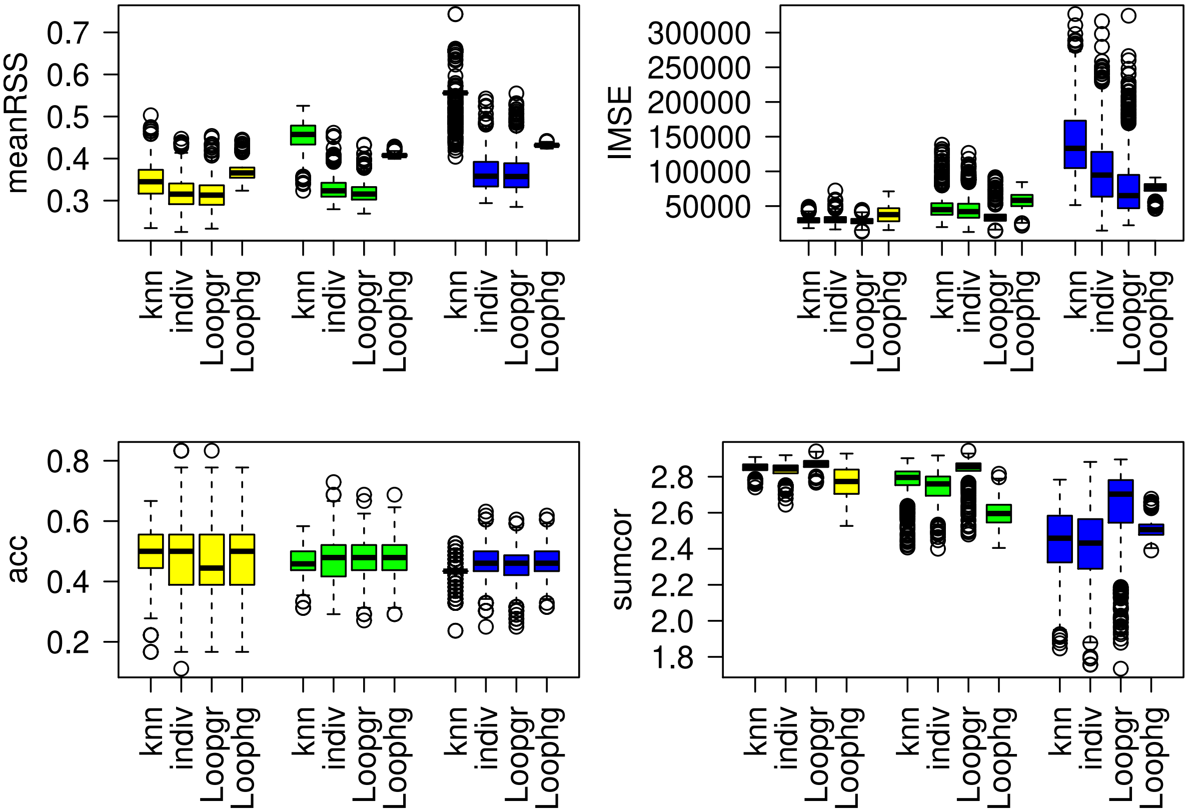

Figure 3: Measures of performance for the knn, individual, Loop grW and Loop hgW methods. Each color boxplot represents a different percentage of hidden observation: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$.

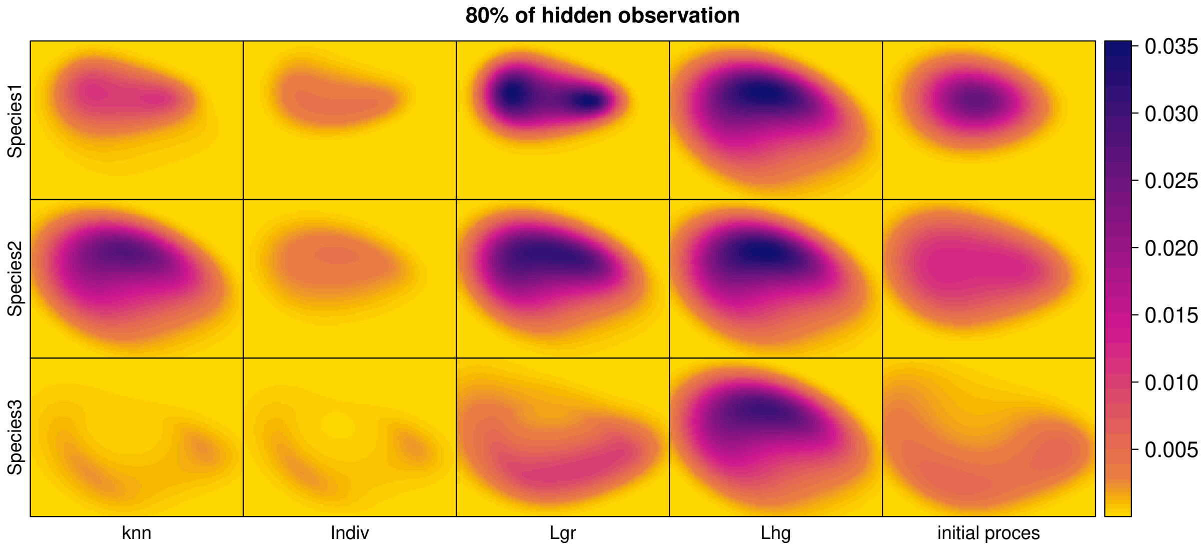

Figure 4: Predicted intensities obtained for the knn, individual, Loop grW and Loop grW methods and the initial intensities from the process with 80% of hidden observations. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$.

5.1.2 Similar abundances and correlated distributions

In Figure 5, we consider similar abundances ($m_1 = 33$, $m_2 = 34$ and $m_3 = 35$) and where two distributions are highly correlated. With regard to classification performance, the different modelling methods have similar levels of accuracy, except the knn method does relatively poorly with 80% of the observations hidden. The knn method also suffers worse performance as measured by meanRSS at 50% and 80% of hidden observations. Measures of predictive performance are similar to the case with different abundances and correlated distributions. The Loop grW method appears to outperform the others as the proportion of hidden observations increases, with the Loop hgW method competitive with the individual PPM method. The knn method appears to do worse with 80% hidden observations when measured by IMSE. See Tables **??** and **??** in the Appendix for comparisons of means and medians across all of these measures.

With 80% hidden observations, the Loop Loop grW method appears to be best aligned with the true intensities, as shown in Figure 6.

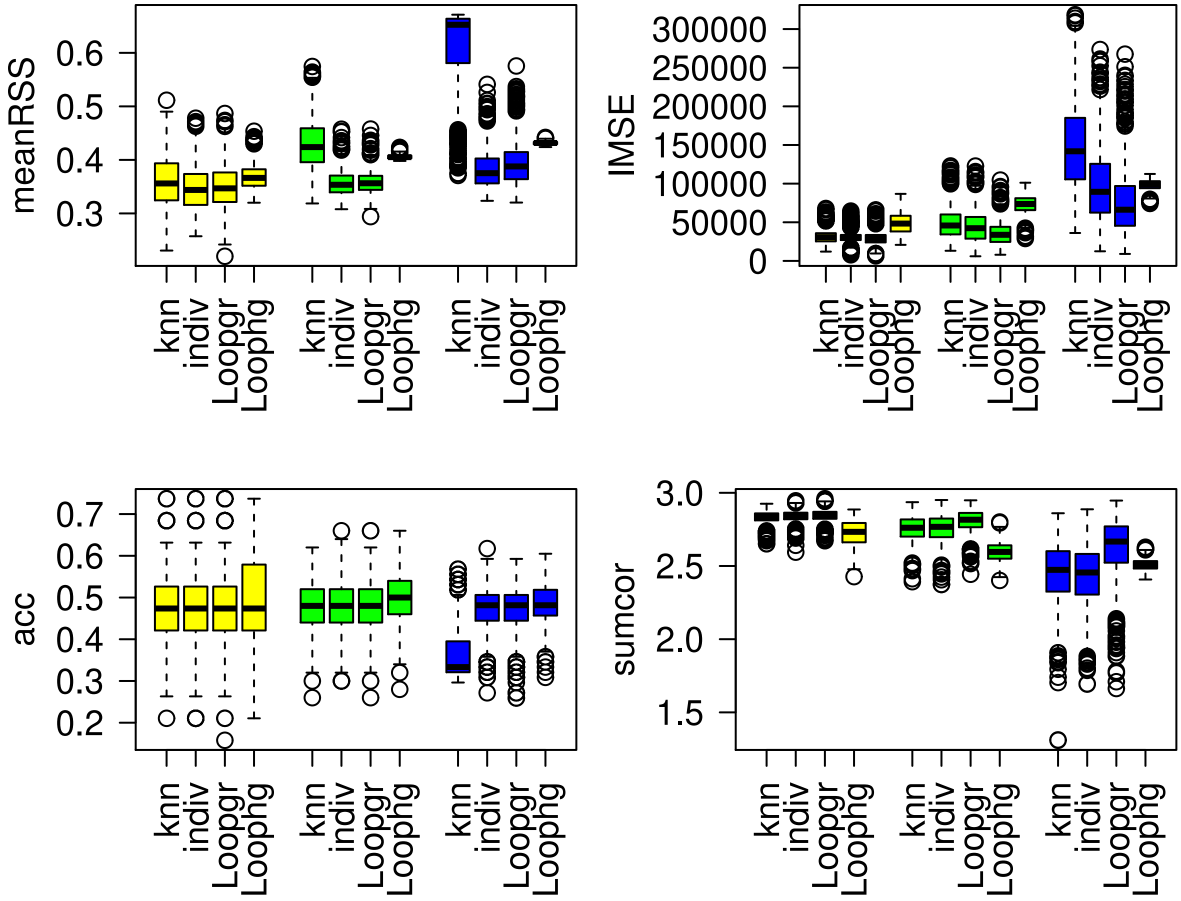

Figure 5: Measures of performance for the knn, individual, Loop grW and Loop hgW methods. Each color represents a different percentage of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 33$, $m_2 = 34$, $m_3 = 35$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$.

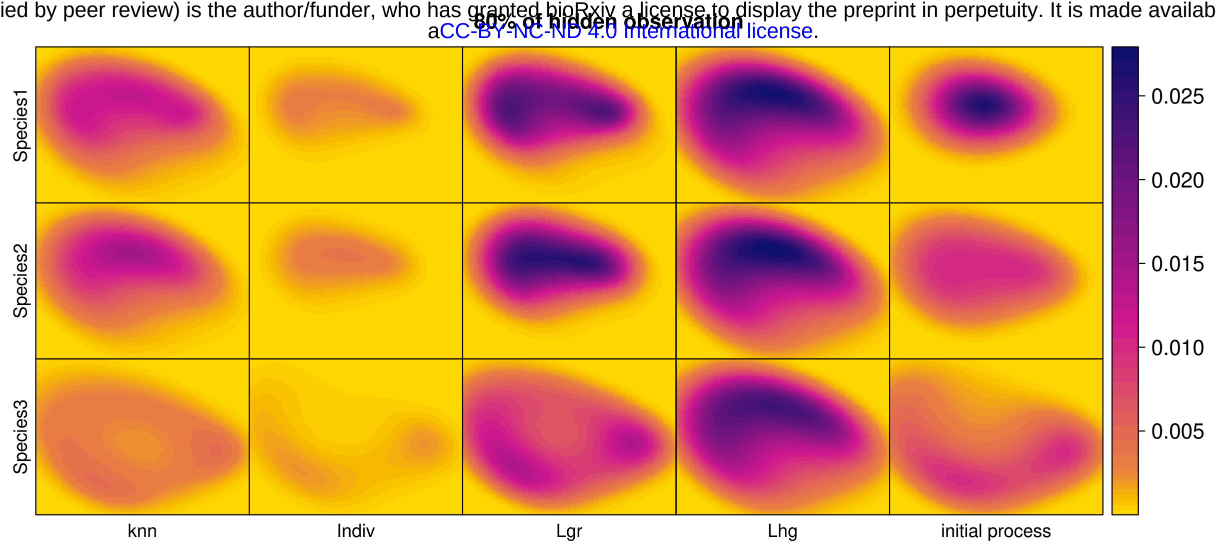

Figure 6: Predicted intensities obtained for the knn, individual, Loop grW and Loop hgW methods and the initial intensities from the process with 80% of hidden observations. The parameters of abundances and correlation are: $m_1 = 33$, $m_2 = 34$, $m_3 = 35$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$.

5.1.3 Different abundances and non correlated distributions

In Figure 7, we consider different abundances ($m_1 = 42$, $m_2 = 31$ and $m_3 = 25$) and where none of the distributions have high correlations. The classification performance and predictive performance comparisons look similar to the case of similar abundances and correlated distributions as shown in Figure 5, with the knn method having the worst classification performance described here at 50% and 80% of hidden observations and the Loop grW method outperforming the others in predictive performance, while the Loop hgW method is competitive with the individual PPM method and the knn method lags behind with IMSE at 80% of hidden observations. Tables 5 and 6 in the Appendix contains the means and medians across all performance measures for this context.

With 80% of hidden observation as shown in Figure 8, the Loop hgW method for species 1 and 3 and the Loop grW method for species 2 and 3 are the closest to the initial process.

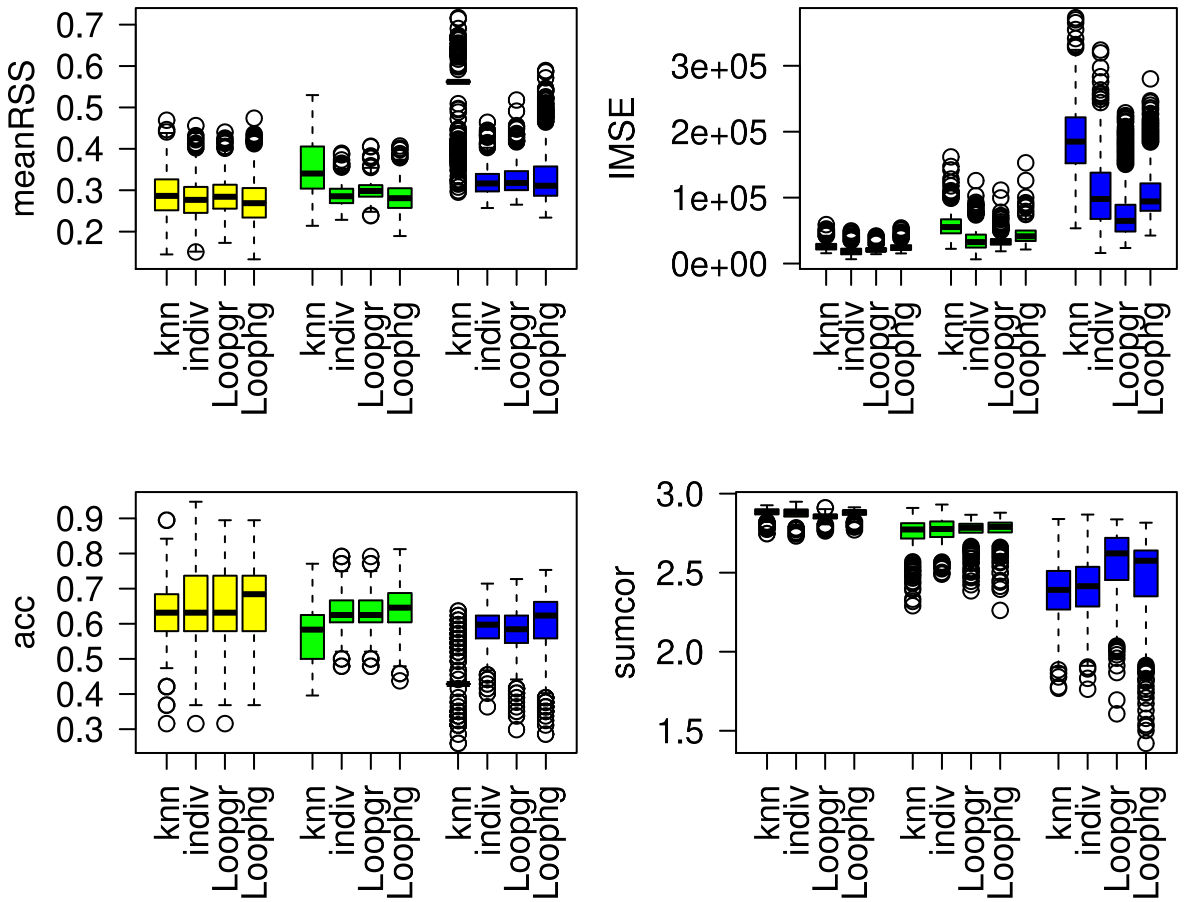

Figure 7: Measures of performance for the knn, individual, Loop grW and Loop hgW methods. Each color represents a different percentage of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 42$, $m_2 = 31$, $m_3 = 25$; $\rho_{1,2} = 0.09$, $\rho_{1,3} = -0.42$, $\rho_{2,3} = 0.20$.

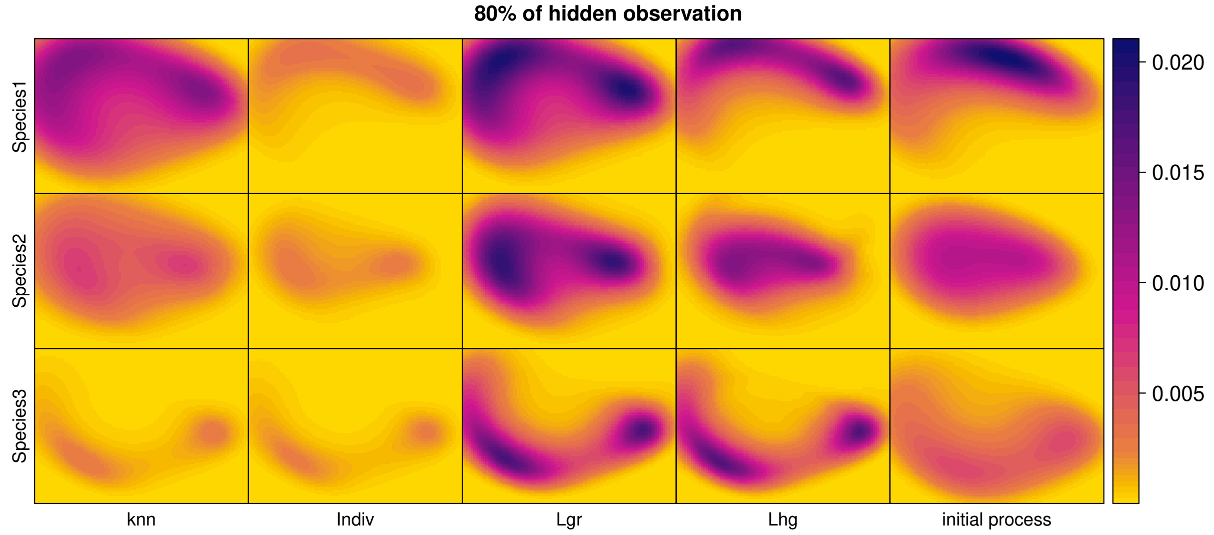

Figure 8: Predicted intensities obtained for the knn, individual, Loop grW and Loop hgW methods and the initial intensities from the process with 80% of hidden observations. The parameters of abundances and correlation are: $m_1 = 42$, $m_2 = 31$, $m_3 = 25$; $\rho_{1,2} = 0.09$, $\rho_{1,3} = -0.42$, $\rho_{2,3} = 0.20$.

5.1.4 Similar abundances and non correlated distribution

For similar abundances ($m_1 = 39$, $m_2 = 37$, $m_3 = 38$) and non correlated distributions, we again observe the same trends, as shown in Figure 9: the knn method is the worst method for relabeling performances and the only one not doing as well as the individual method for 50% and 80% of hidden observations. As in previous contexts, the Loop grW method shows the best predictive performance, with the Loop hgW method being competitive with the individual PPM method, and the knn method having higher IMSE than the other methods when 80% of the observations are hidden. Tables 7 and 8 in the Appendix contain the mean and median value for all performance measures.

The predicted intensities show the methods LgrW and knn being the closest to the initial process, as shown in Figure 10.

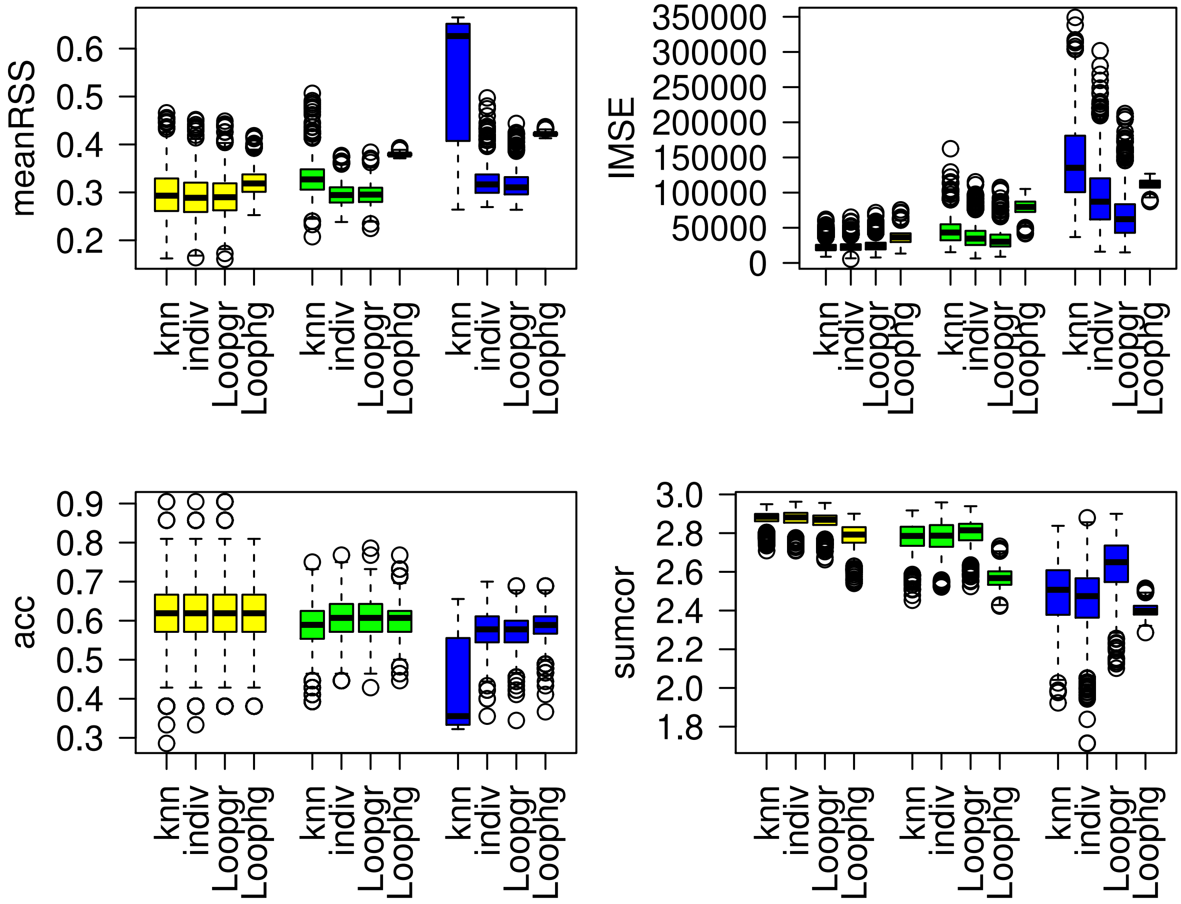

Figure 9: Measures of performance for the knn, individual, Loop grW and Loop grW methods. Each color represents a different proportion of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 39$, $m_2 = 37$, $m_3 = 38$; $\rho_{1,2} = 0.09$, $\rho_{1,3} = -0.42$, $\rho_{2,3} = 0.20$.

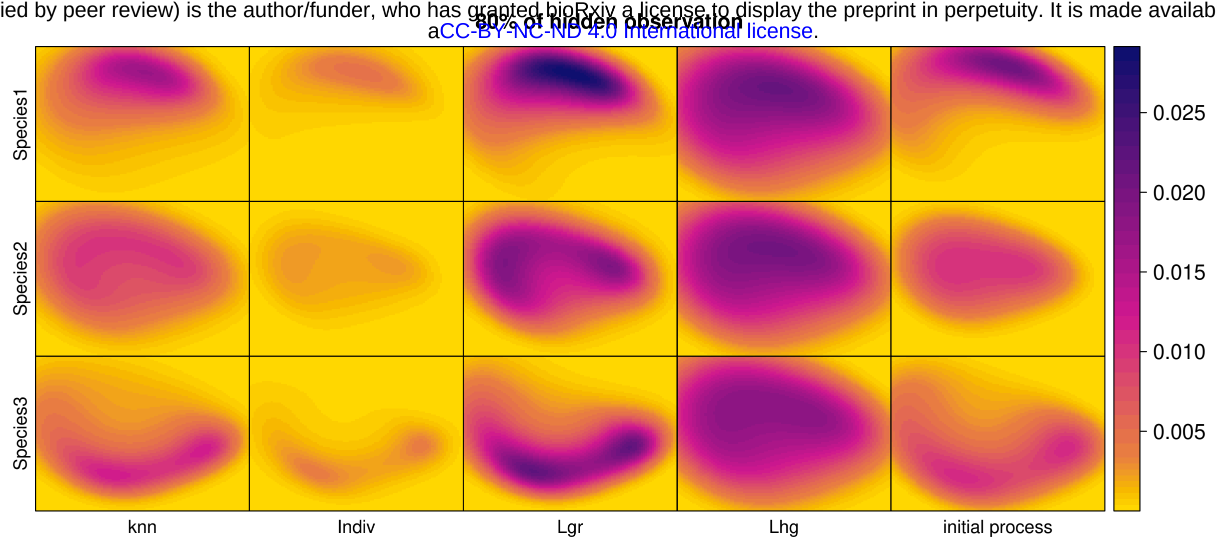

Figure 10: Predicted intensities obtained for the knn, individual, Loopg rW and Loop grW initialization methods and the initial intensities from the process at 80% of hidden observations. The parameters of abundances and correlation are:$m1$=39, $m2$=37, $m3$=38; ρ_{1-2}=0.09, ρ_{1-3}=-0.42, ρ_{2-3}=0.20

5.2 Testing algorithm parameters

5.2.1 knn method

We note that when the k nearest neighbor value increases (from 1 up to 20), the model performances decrease; Figure 11. It is particularly notable for the performances in prediction where sumcor performances decrease and IMSE performances increase. Also, there is an expected drop in performances as we increase the proportion of observations with unknown species labels.

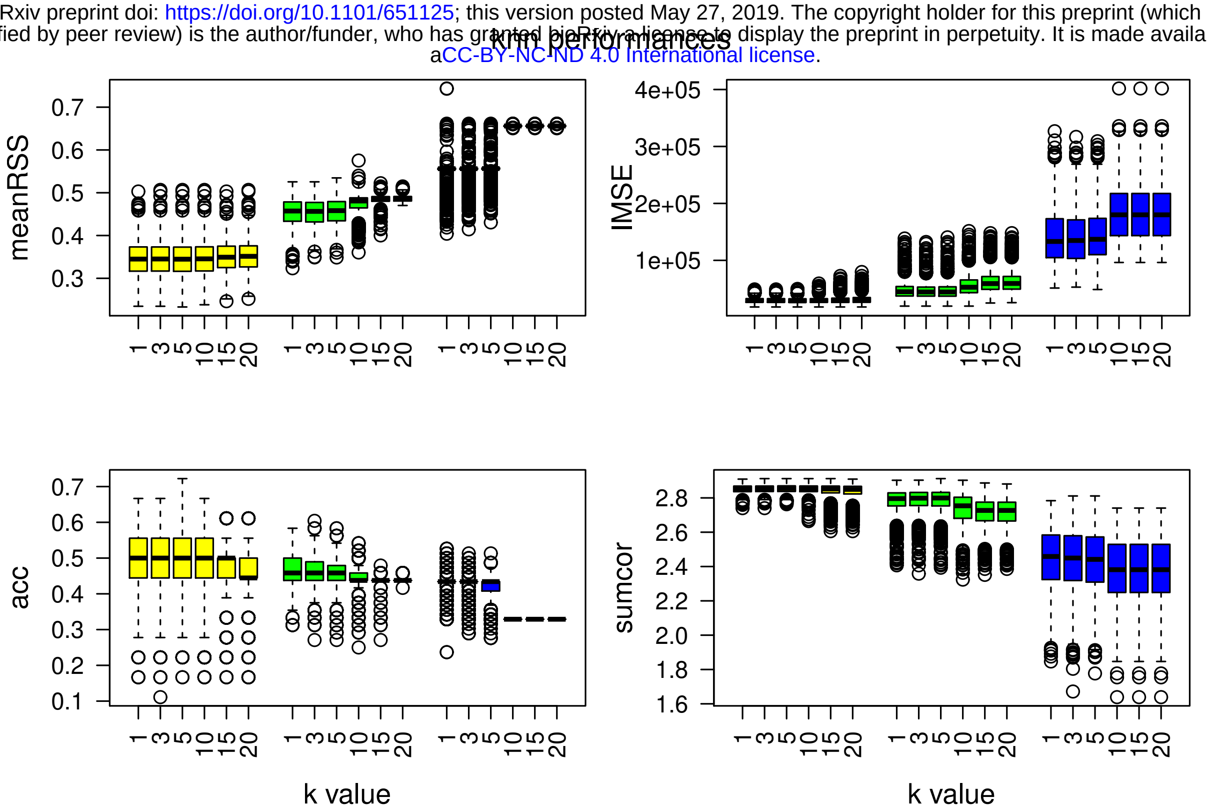

Figure 11: Model performances for the knn method. Each color represents a different percentage of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$

5.2.2 Loop grW method

For the Loop grW method we tested different parameters:

1. The initial membership probability threshold $\delta_{\max}$: while this parameter varies from 0.8 to 0.5 in increments of 0.1, the other Loop grW parameters are as follows: $\delta_{\min} = 0.1$ and $\delta_{\text{step}} = 0.1$.
2. The final membership probability threshold $\delta_{\min}$: while this parameter varies from 0.1 to 0.7 in increments of 0.2, the other Loop grW parameters are as follows: $\delta_{\max} = 0.8$ and $\delta_{\text{step}} = 0.1$.
3. The step size δ_{step}: while this parameter varies from a minimum of 0.01 to a maximum of 0.2, the other Loop grW parameters are as follows: $\delta_{\max} = 0.8$ and $\delta_{\min} = 0.1$.

When we change the value of $\delta_{\max}$, there is very little difference in performance within each proportion of observations with hidden labels, although $\delta_{\max} = 0.5$ appears to be slightly superior to the other choices for high percentage of hidden observation (Figure 12).

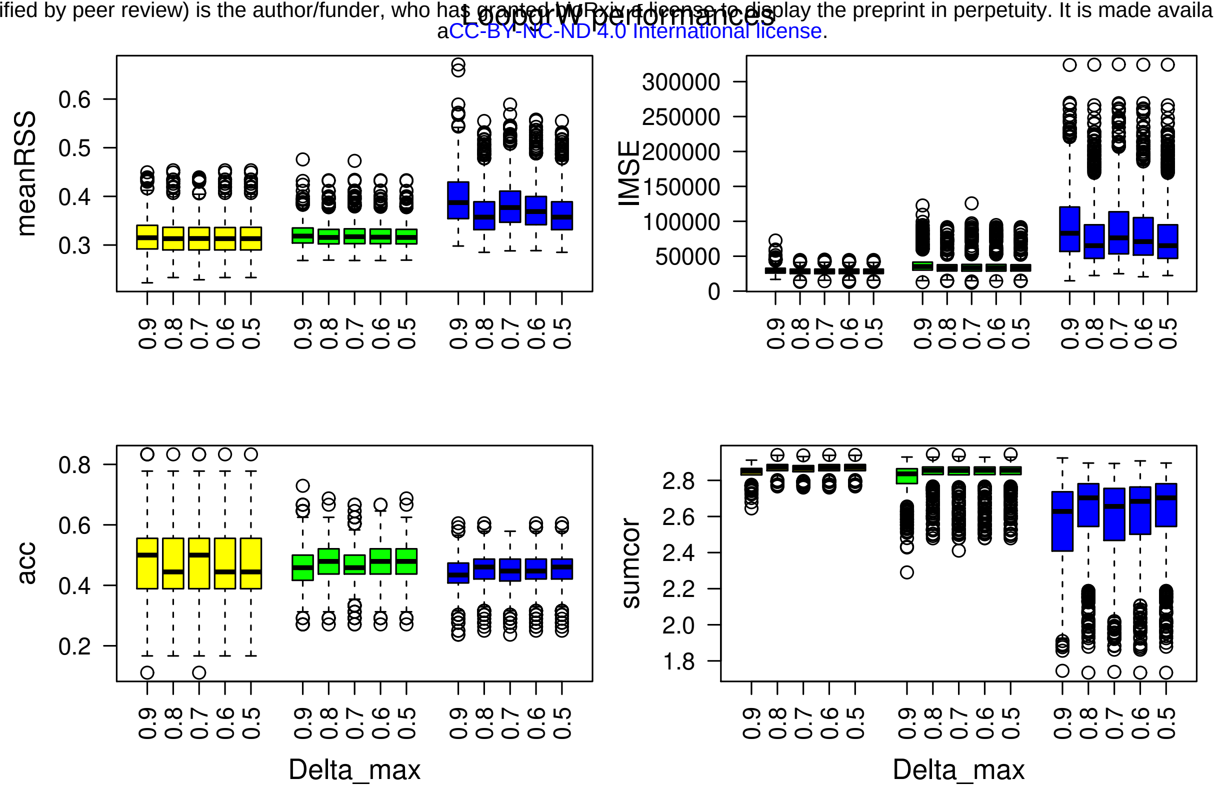

Figure 12: Model performances for the Loop grW method and for different values of $\delta_{\max}$. Each color represents a different proportion of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$

When changing $\delta_{\min}$, the classification accuracy is relatively the same (Figure 13). For MeanRSS, IMSE and sumcor, we can observe a curved pattern of performances, where the performances decrease (MeanRSS increases, IMSE increases and sumcor decreases) from $\delta_{\min}$ from 0.1 to 0.5 and then the performances get slightly better (MeanRSS decreases, IMSE decreases and sumcor increases) for $\delta_{\min} = 0.7$ (Figure 13). $\delta_{\min}$=0.1 displays the better performances.

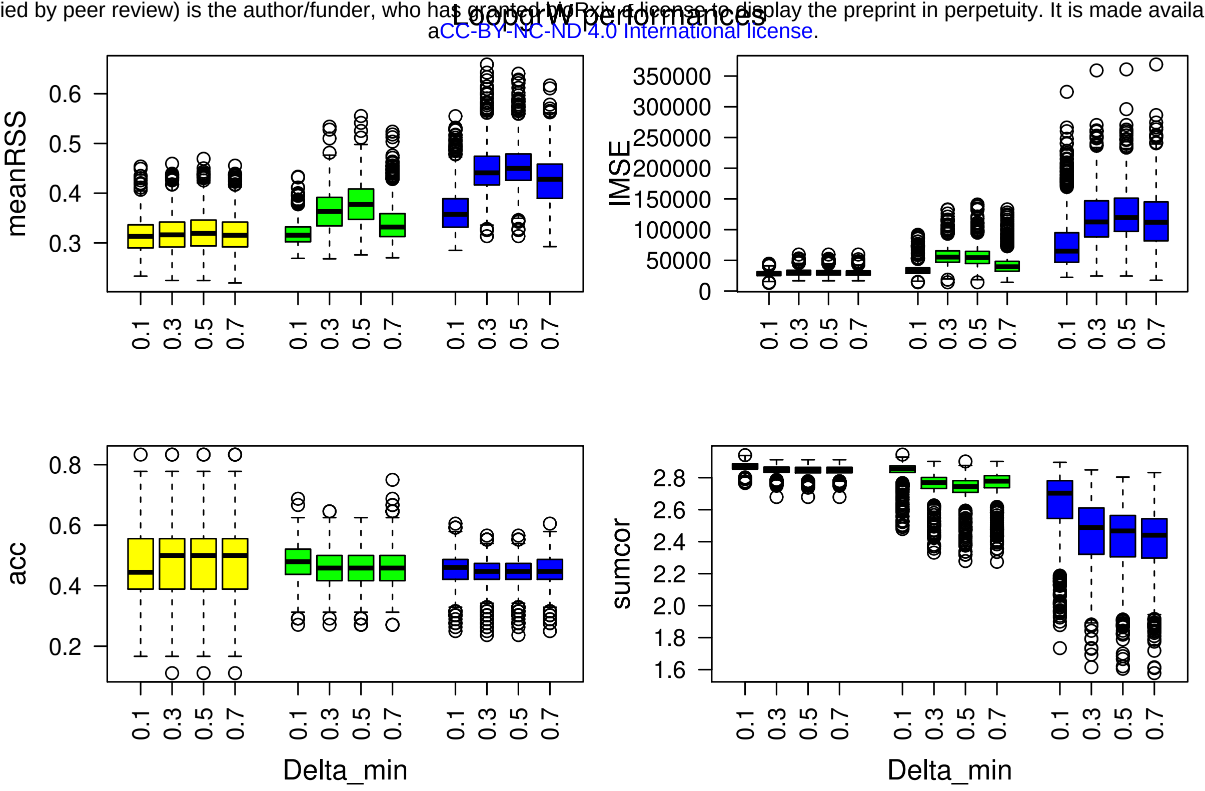

Figure 13: Model performances for the Loop grW method and for different values of $\delta_{\min}$. Each color represents a different proportion of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$

.

Figure 14 shows different performance measures as we vary δ_{step}. There do not appear to be major differences in classification performance, although 0.1 appear slightly better for meanRSS. With 50% and 80% of hidden observations, predictive performance display a curve performances where performances get better (IMSE decreases and sumcor increase) from 0.01 till 0.1 and then get worse (IMSE increases and sumcor descreases) from 0.1 to 0.2. δ_{step}=0.1 displays the best performances accross all measures.

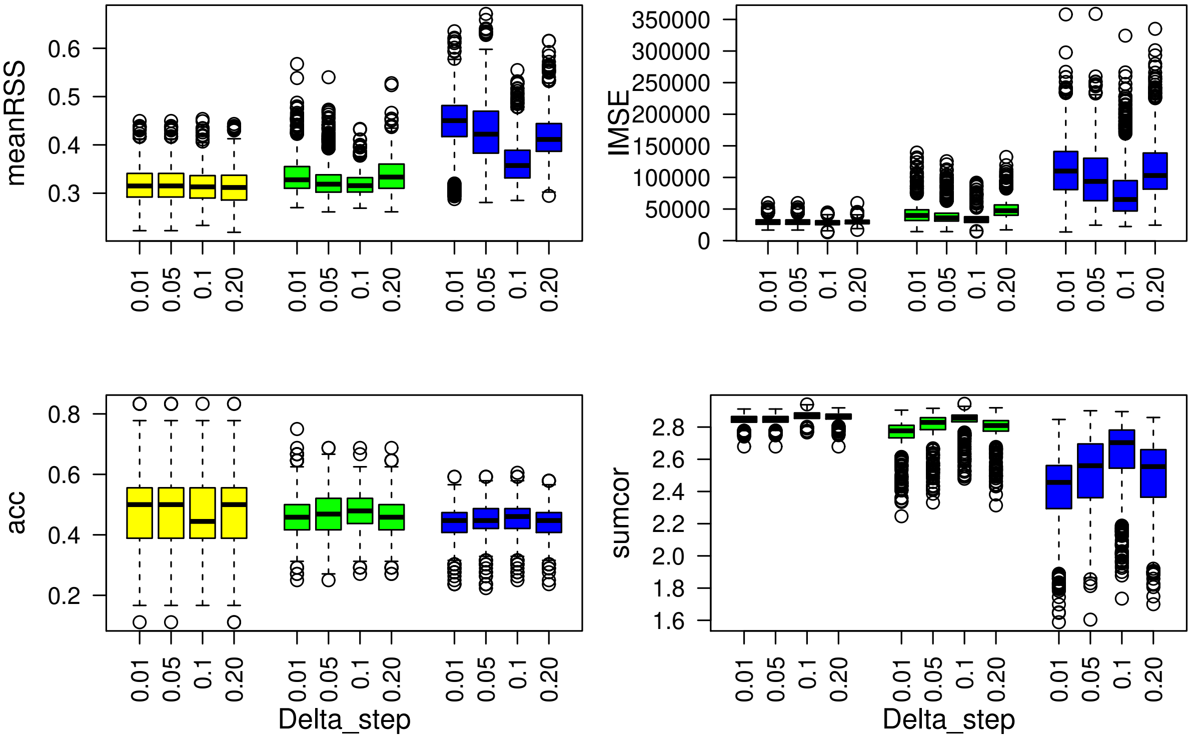

Figure 14: Model performances for the Loop grW method and for different values of weight step. Each color represents a different proportion of hidden observations: in yellow are the performances with 20% of hidden observations, in green with 50% and in blue with 80%. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$

.

311 5.2.3 Loop hgW method

312 In the Loop hgW method, we vary the number of points a added at each iteration. In Figure 15, we can
313 see that there is no variation in performances when the number of added points a increases.

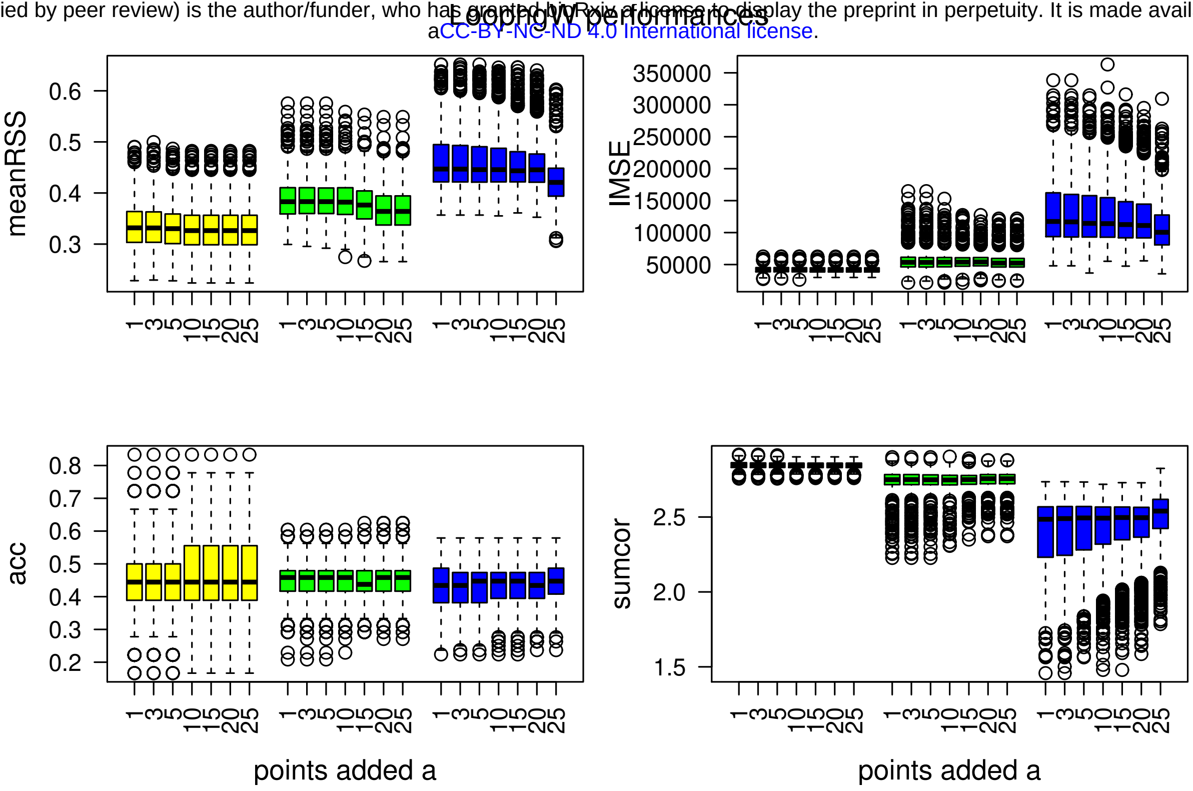

Figure 15: Model performances for the Loop grW method. Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50% and in blue for 80%. The parameters of abundances and correlation are: $m_1 = 32$, $m_2 = 42$, $m_3 = 23$; $\rho_{1,2} = 0.85$, $\rho_{1,3} = -0.09$, $\rho_{2,3} = 0.20$

The results for the other combination of abundances and correlation are showed in the Appendix.

6 Discussion

In this article, we present a new modelling tool in R that aims to incorporate the observed locations with unknown species identities to improve species distributions. These tools accommodate two ways of reclassifying information using mixture modelling and the machine learning framework with 7 different initialization methods. We tested our algorithms in different contexts where we vary the abundances of our species (similar or different), the correlation between them (two distribution are correlated or none are correlated) and the proportion of unknown species identities (20%, 50% and 80%). The different methods were compared to the individual method which ignores locations with unknown species identities to see whether the proposed algorithms allow us to fit distributions that are closer to the initial processes.

In the results we presented the three best methods. They showed varying performance depending on the aspects of the model and the performance measure considered. The novelty of these tools, makes it difficult to compare to other existing tools that either do not consider point pattern process (Frame & Jammalamadaka, 2007; Frühwirth-Schnatter, 2006; Hui, 2016; Martinez, 2015; Melnykov & Maitra, 2010; Quost & Denœux, 2016), Poisson distributions (Figueirido & Jain, 2002; Hui *et al.*, 2015; Scrucca *et al.*, 2016; Woillez *et al.*, 2012), count data (Benaglia *et al.*, 2009; Iovleff, 2018; Leisch, 2004) or implementation of mixture (Witten, 2011; Wendel *et al.*, 2015) or semi-supervised learning frameworks (Di Zio *et al.*, 2007; Fraley & Raftery, 1998; Jeffries & Pfeiffer, 2001; Taddy & Kottas, 2012).

The other methods (kmeans, random, equal and normal) not presented previously in the results are presented in the Appendix. They show relatively worse performance across all measures, although at times, the normal loop method is competitive with the individual PPM and the Loop hgW methods. We note that this method performs slightly better when the distributions are correlated.

We have noticed differences in performance, that are more significant when we increase the proportion of observations with hidden labels. While at 20% of hidden observations, all methods performed fairly similarly, at 50% and 80% of hidden observations, the loop grW method in particular showed the best predictive performances regardless of differences in abundance and correlation among species distributions. For this method, only the points with the highest membership probabilities are added. We set the maximum and minimum thresholds at $\delta_{\text{max}} = 0.5$ and $\delta_{\text{min}} = 0.1$ and a step size of $\delta_{\text{step}} = 0.1$, but we could expect that performances may be better or worse with different choices of these parameters as shown in the results. These choices appeared to produce superior performances for most measures than other values of these parameters considered. Higher values of δ_{min} led to worse performances. This result can be seen as counterintuitive as we can expect that having a smaller interval of weight for example could improve this particular performances. It will in other words reduce the interval of weights and better discriminate the points of uncertain identity. As for δ_{step}, choosing a value that is too small may lead to iterations where no points are added, while choosing a value that is too large may be too discriminating and does not allow to reclassify the points.

The Loop hgW method did not perform as good as the Loop grW method even if it has been shown to be as good as the individual method in some contexts. For this method, we add initially a certain number of points a that is increased at each iteration. While the a points with highest membership probabilities are added, these membership probabilities may be small for large values of a, and this could explain that this method is not always doing as good as the best method.

Interestingly, the knn method was the best of the four mixture methods tested, outperforming the kmeans, random and equal initialization options. Previous studies using the EM algorithm for classification and clustering data show that such algorithms are highly dependent on the initialization method (Figueirido & Jain, 2002; Melnykov & Maitra, 2010; O'Hagan *et al.*, 2012). Additionally, even very popular methods like kmeans have some drawbacks. Its performance is dependent on overlapping densities and whether the distributions are roughly circular or not. The choice of the centroid is also not consistent and chosen at random for the first calculation (Yoo *et al.*, 2012, 2007; Wu *et al.*, 2008). In our simulations, kmeans, random and equal methods showed very different results and always performed worse than the other methods as well as mainly overestimating (kmeans and random) and underestimating (equal) the predicted intensities compared to the true process.

Despite outperforming the other mixture modelling methods, the knn method was still not competitive with the machine learning methods or the individual PPM method when the proportion of hidden observations are 50% or 80%. However, the knn method was quite consistent in the predicted intensities and showed similar results to the individual method for the sumcor measure at 50% or 80% of hidden observations. Other studies have found that the performance of the knn method is linked to the metric chosen to calculate the nearest neighbor distances and the value of the number k of nearest neighbors (Weinberger & Saul, 2009; Guo *et al.*, 2003; Wu *et al.*, 2008).

We tested how the number of neighbors k can influence the model and found that for any combination of abundance and correlation, all the measures of performances decrease when the values of k increase. It is expected as the neighboring points are further away from one another and could conflate species habitat preferences with differing species abundances, but requiring more neighbor points can also stabilize the distances. The way of choosing the value of k by utilizing different distance metrics could also impact the performances as previously noted, but we shall leave this aspect of the analysis for future consideration.

In our simulations, we have considered a relatively general case of point patterns and we only varied species abundance and correlation among distributions in addition to the proportion of observations with hidden information. For real ecological data sets, there are more factors to consider that can influence how a model will perform. First, the abundances tested in the simulation are quite low (20-40 points) and some methods can show convergence issues in this context. While we use the spatstat package (Baddeley *et al.*, 2015) to fit PPMs, we could make use of similar functions in the ppmlasso package (Renner & Warton, 2013) which integrate regularization methods like the lasso penalty that can boost performances with low sample sizes. A related point is that we included all covariates that were used to generate the true point patterns in our models. In real situations, however, we may not have access to the best covariates or know which ones truly determine the species distributions. Applying a lasso penalty to help in variable selection may therefore be provide a natural way forward in this context. Finally, a key reality when dealing with presence-only data is the presence of observer bias, in which sampling effort varies throughout the study region. Some models apply a correction for observer bias in the prediction (Hefley *et al.*, 2013; Lahoz-Monfort *et al.*, 2014; Warton *et al.*, 2013) and our tools would be able to accommodate such improvements.

7 Conclusion

The new algorithms presented in this article aim to reclassify observations that have uncertain or unknown labels in order to better predict point pattern distributions. We showed that machine learning based models performed better in a general context than mixture based models no matter the initialization method and also better than the individual PPM method that does not include the points with unknown labels. Our simulations showed encouraging results in this context with good performances in some cases, although there are some improvements to implement in order to make the tools more appropriate for real life data.

Acknowledgments

Computational resources used in this work were provided by Intersect Australia Ltd.

Authors' contributions

EG and IR conceived the ideas and designed methodology; EG and IR built the algorithms; EG analyzed the data; EG and IR led the writing of the manuscript. MM had an overview on the project. MM and EB reviewed the paper. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

Rscript: An example of the scripts used for this paper is available here: functions to use for the test simulation 📎 and the script example 📎.

References

Aarts, G., Fieberg, J. & Matthiopoulos, J. (2012) Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, **3**, 177–187. https://doi.org/10.1111/j.2041-210X.2011.00141.x.

Baddeley, A., Gregori, P., Mateu, J., Stoica, R. & Stoyan, D. (2006) *Modelling Spatial Point Patterns in R. In: Baddeley A., Gregori P., Mateu J., Stoica R., Stoyan D. (eds) Case Studies in Spatial Point Process Modeling.*, volume 185 of *Lecture Notes in Statistics*. Springer, New York, NY. https://doi.org/10.1007/0-387-31144-0_2.

Baddeley, A., Rubak, E. & Turner, R. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London. https://doi.org/10.1201/b19708.

Benaglia, T., Chauveau, D., Hunter, D.R. & Young, D. (2009) mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, **32**, 1–29. https://doi.org/10.18637/jss.v032.i06.

Berman, M. & Turner, T.R. (1992) Approximating point process likelihoods with glim. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **41**, 31–38. https://doi.org/10.2307/2347614.

Browning, E., Bolton, M., Owen, E., Shoji, A., Guilford, T., Freeman, R. & McPherson, J. (2018) Predicting animal behaviour using deep learning: Gps data alone accurately predict diving in seabirds. *Methods in Ecology and Evolution*, **9**, 681–692. https://doi.org/10.1111/2041-210x.12926.

Burnham, K.P. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York, NY, 2 edition. https://doi.org/10.1007/b97636.

Di Zio, M., Guarnera, U. & Rocci, R. (2007) A mixture of mixture models for a classification problem: The unity measure error. *Computational Statistics and Data Analysis*, **51**, 2573–2585. https://doi.org/10.1016/j.csda.2006.01.001.

Dunstan, P.K., Foster, S.D., Hui, F.K.C. & Warton, D.I. (2013) Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, **18**, 357–375. https://doi.org/10.1007/s13253-013-0146-x.

Es, B. (1997) A note on the integrated squared error of a kernel density estimator in non-smooth cases. *Statistics and Probability Letters*, **35**, 241–250. https://doi.org/10.1016/S0167-7152(97)00019-9.

Fernández-Michelli, J.I., Hurtado, M., Areta, J.A. & Muravchik, C.H. (2016) Unsupervised classification algorithm based on em method for polarimetric sar images. *ISPRS Journal of Photogrammetry and Remote Sensing*, **117**, 56–65. https://doi.org/10.1016/j.isprsjprs.2016.03.001.

Figueirido, M.A. & Jain, A.K. (2002) Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, **24**, 381–396. https://doi.org/10.1109/34.990138.

Fraley, C. & Raftery, A.E. (1998) How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, **41**, 578–588. https://doi.org/10.1093/comjnl/41.8.578.

Frame, S.J. & Jammalamadaka, S.R. (2007) Generalized mixture models, semi-supervised learning, and unknown class inference. *Advances in Data Analysis and Classification*, **1**, 23–38. https://doi.org/10.1007/s11634-006-0001-9.

Franklin, J. (2013) Species distribution models in conservation biogeography: developments and challenges. *Diversity and Distributions*, **19**, 1217–1223. https://doi.org/10.1111/ddi.12125.

Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Springer series in statistics. Springer. https://doi.org/10.1007/978-0-387-35768-3.

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

Guisan, A., Tingley, R., Baumgartner, J.B., Naujokaitis-Lewis, I., Sutcliffe, P.R., Tulloch, A.I., Regan, T.J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T.G., Rhodes, J.R., Maggini, R., Setterfield, S.A., Elith, J., Schwartz, M.W., Wintle, B.A., Broennimann, O., Austin, M., Ferrier, S., Kearney, M.R., Possingham, H.P. & Buckley, Y.M. (2013) Predicting species distributions for conservation decisions. *Ecol Lett*, **16**, 1424–35. https://doi.org/10.1111/ele.12189.

Guo, G., Wang, H., Bell, D., Bi, Y. & Greer, K. (2003) Knn model-based approach in classification. R. Meersman, Z. Tari & D.C. Schmidt, eds., *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pp. 986–996. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_62.

Hastie, T., J., Tibshirani, R. & Friedman, J. (2001) *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, volume 1. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21606-5.

Hefley, T.J., Tyre, A.J., Baasch, D.M. & Blankenship, E.E. (2013) Nondetection sampling bias in marked presence-only data. *Ecol Evol*, **3**, 5225–36. https://doi.org/10.1002/ece3.887.

Hui, F.K.C. (2016) *Mixing it Up: New Methods for Finite Mixture Modelling of Multi-Species Data in Ecology*. Ph.D. thesis. https://doi.org/10.1017/S0004972715000945.

Hui, F.K.C., Warton, D.I. & Foster, S.D. (2015) Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics*, **9**, 866–882. https://doi.org/10.1214/15-aoas813.

Illian, J.B., Sørbye, S.H. & Rue, H. (2012) A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (inla). *The Annals of Applied Statistics*, **6**, 1499–1530. https://doi.org/10.1214/11-aoas530.

Inoue, K., Stoeckl, K., Geist, J. & Ricciardi, A. (2017) Joint species models reveal the effects of environment on community assemblage of freshwater mussels and fishes in european rivers. *Diversity and Distributions*, **23**, 284–296. https://doi.org/10.1111/ddi.12520.

Iovleff, S. (2018) *MixAll: Clustering and Classification using Model-Based Mixture Models*. R package version 1.4.2.

Jeffries, N. & Pfeiffer, R. (2001) A mixture model for the probability distribution of rain rate. *Environmetrics*, **12**, 1–10. https://doi.org/10.1002/1099-095X(200102)12:1<1::AID-ENV425>3.0.CO;2-N.

Jewell, K.J., Arcese, P. & Gergel, S.E. (2007) Robust predictions of species distribution: Spatial habitat models for a brood parasite. *Biological Conservation*, **140**, 259–272. https://doi.org/10.1016/j.biocon.2007.08.017.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Wintle, B.A. (2014) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515. https://doi.org/10.1111/geb.12138.

Leisch, F. (2004) FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, **11**, 1–18. https://doi.org/10.18637/jss.v011.i08.

Mahony, M., Donnellan, S.C., Richards, S.J. & Donald, K. (2006) Species boundaries among barred river frogs, mixophyes (anura: Myobatrachidae) in north-eastern australia, with descriptions of two new species. *Zootaxa*, **1228**, 35–60. https://doi.org/10.5281/zenodo.172713.

Martinez, D.F. (2015) *Mixture-based Clustering for the Ordered Stereotype Model*. Thesis, School of Mathematics Statistics and Operations Research. https://doi.org/10.13140/RG.2.1.1945.4806.

Matthews, J., Steiner, L. & Gordon, J. (2001) Mark-recapture analysis of sperm whale (physeter macrocephalus) photo-id data from the azores (1987-1995). *Journal of cetacean research and management*, **3**, 219–226.

McLachlan, G.J. & Peel, D. (2000) *Finite Mixture Models*. Wiley, New York. https://doi.org/10.1002/0471721182.

Melnykov, V. & Maitra, R. (2010) Finite mixture models and model-based clustering. *Statistics Surveys*, **4**, 80–116. https://doi.org/10.1214/09-ss053.

Mi, X., Bao, L., Jianhua, C. & Ma, K. (2014) Point process models, the dimensions of biodiversity and the importance of small-scale biotic interactions. *Journal of Plant Ecology*, **7**, 126–133. https://doi.org/10.1093/jpe/rtt075.

Nezer, O., Bar-David, S., Gueta, T. & Carmel, Y. (2016) High-resolution species-distribution model based on systematic sampling and indirect observations. *Biodiversity and Conservation*, **26**, 421–437. https://doi.org/10.1007/s10531-016-1251-2.

O'Hagan, A., Murphy, T.B. & Gormley, I.C. (2012) Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics and Data Analysis*, **56**, 3843–3864. https://doi.org/10.1016/j.csda.2012.05.011.

Peterman, W.E., Crawford, J.A. & Kuhns, A.R. (2013) Using species distribution and occupancy modeling to guide survey efforts and assess species status. *Journal for Nature Conservation*, **21**, 114–121. https://doi.org/10.1016/j.jnc.2012.11.005.

Quost, B. & Denœux, T. (2016) Clustering and classification of fuzzy data using the fuzzy em algorithm. *Fuzzy Sets and Systems*, **286**, 134–156. https://doi.org/10.1016/j.fss.2015.04.012.

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

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I. & O'Hara, R.B. (2015) Point process models for presence-only analysis. *Methods in Ecology and Evolution*,

6, 366–379. https://doi.org/10.1111/2041-210x.12352.

Renner, I.W. & Warton, D.I. (2013) Equivalence of maxent and poisson point process models for species distribution modeling in ecology. *Biometrics*, **69**, 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x.

Ruete, A. & Leynaud, G.C. (2015) Goal-oriented evaluation of species distribution models' accuracy and precision: True skill statistic profile and uncertainty maps. Technical report, PeerJ PrePrints. https://dx.doi.org/10.7287/peerj.prepints.1208v1.

Schank, C.J., Cove, M.V., Kelly, M.J., Mendoza, E., O'Farrill, G., Reyna-Hurtado, R., Meyer, N., Jordan, C.A., González-Maya, J.F., Lizcano, D.J., Moreno, R., Dobbins, M.T., Montalvo, V., Sáenz-Bolaños, C., Jimenez, E.C., Estrada, N., Cruz Díaz, J.C., Saenz, J., Spínola, M., Carver, A., Fort, J., Nielsen, C.K., Botello, F., Pozo Montuy, G., Rivero, M., de la Torre, J.A., Brenes-Mora, E., Godínez-Gómez, O., Wood, M.A., Gilbert, J., Miller, J.A. & Thuille, W. (2017) Using a novel model approach to assess the distribution and conservation status of the endangered baird's tapir. *Diversity and Distributions*, **23**, 1459–1471. https://doi.org/10.1111/ddi.12631.

Scrucca, L., Fop, M., Murphy, T.B. & Raftery, A.E. (2016) mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *the R journal*, **8**, 289–317. https://doi.org/10.21236/ada459792.

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

Taddy, M.A. & Kottas, A. (2012) Mixture modeling for marked poisson processes. *Bayesian Analysis*, **7**, 335–362. https://doi.org/10.1214/12-ba711.

Thessen, A. (2016) Adoption of machine learning techniques in ecology and earth science. *One Ecosystem*, **1**. https://doi.org/10.3897/oneeco.1.e8621.

Tracey, J.A., Zhu, J., Boydston, E., Lyren, L., Fisher, R.N. & Crooks, K.R. (2013) Mapping behavioral landscapes for animal movement: a finite mixture modeling approach. *Ecological Applications*, **23**, 654–669. https://doi.org/10.1890/12-0687.1.

Tran, N.Q. (2017) *Classification, Novelty Detection and Clustering for Point Pattern Data*. Thesis, Faculty of Science and Engineering, Department of Electrical and Computer Engineering. http://hdl.handle.net/20.500.11937/59025.

van Strien, A.J., van Swaay, C.A. & Termaat, T. (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, **50**, 1450–1458. https://doi.org/10.1111/1365-2664.12158.

Vo, B.N., Dam, N., Phung, D., N. Tran, Q. & Vo, B.T. (2018) Model-based learning for point pattern data. *Pattern Recognition*, **84**. https://doi.org/10.1016/j.patcog.2018.07.008.

Warton, D.I., Renner, I.W. & Ramp, D. (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS One*, **8**, e79168. https://doi.org/10.1371/journal.pone.0079168.

Warton, D.I. & Shepherd, L.C. (2010) Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. *The Annals of Applied Statistics*, **4**, 1383–1402. https://doi.org/10.1214/10-aoas331.

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

Wendel, J., Buttenfield, B.P. & Stanislawski, L.V. (2015) An evaluation of unsupervised and supervised learning algorithms for clustering landscape types in the united states. *Cartography and Geographic Information Science*, **43**, 233–249. https://doi.org/10.1080/15230406.2015.1067829.

Witten, D.M. (2011) Classification and clustering of sequencing data using a poisson model. *The Annals of Applied Statistics*, **5**, 2493–2518. https://doi.org/10.1214/11-aoas493.

Woillez, M., Ressler, P.H., Wilson, C.D. & Horne, J.K. (2012) Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery. *J Acoust Soc Am*, **131**, EL184–90. https://doi.org/10.1121/1.3678685.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y. *et al.* (2008) Top 10 algorithms in data mining. *Knowledge and information systems*, **14**, 1–37. https://doi.org/10.1007/s10115-007-0114-2.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F. & Hua, L. (2012) Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, **36**, 2431–2448. 10.1007/s10916-011-9710-5.

Yoo, I., Hu, X. & Song, I.Y. (2007) Biomedical ontology improves biomedical literature clustering performance: a comparison study. *International Journal of Bioinformatics Research and Applications*, **3**, 414–428. https://doi.org/10.1504/IJBRA.2007.015010.

Zhang, L., Liu, C. & Davis, C.J. (2004) A mixture model-based approach to the classification of ecological habitats using forest inventory and analysis data. *Canadian journal of forest research*, **34**, 1150–1156. https://doi.org/10.1139/x04-005.

Zhou, Z.H. (2018) A brief introduction to weakly supervised learning. *National Science Review*, **5**, 44–53. https://doi.org/10.1093/nsr/nwx106.