Received: 21 October 2020 | Revised: 7 February 2021 | Accepted: 11 February 2021

DOI: 10.1002/ece3.7411

ORIGINAL RESEARCH

Ecology and Evolution Open Access WILEY

How to make use of unlabeled observations in species distribution modeling using point process models

Emy Guilbault[1] | Ian Renner[1] | Michael Mahony[2] | Eric Beh[1]

[1]Faculty of Science, School of Mathematical and Physical Sciences, The University of Newcastle, Callaghan, NSW, Australia

[2]Faculty of Science, School of Environmental and Life Sciences, The University of Newcastle, Callaghan, NSW, Australia

Correspondence
Emy Guilbault, Faculty of Science, The University of Newcastle, Callaghan, NSW, Australia.
Email: Emy.Guilbault@uon.edu.au

Abstract

1. Species distribution modeling, 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 such models. However, the species observations used can have varying levels of quality and can have incomplete information, such as uncertain or unknown species identity.
2. In this paper, we develop two algorithms to classify observations with unknown species identities which simultaneously predict several species distributions using spatial point processes. Through simulations, we compare the performance of these algorithms using 7 different initializations to the performance of models fitted using only the observations with known species identity.
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 did not use the misspecified data. We applied the best-performing methods to a dataset of three frog species (*Mixophyes*).
4. These models represent a helpful and promising tool for opportunistic surveys where misidentification is possible or for the distribution of species newly separated in their taxonomy.

KEYWORDS

classification, ecological statistics, EM algorithm, machine learning, misidentification, mixture modeling, presence-only data

1 | INTRODUCTION AND BACKGROUND

Species distribution modeling 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 (Inoue et al., 2017; Jewell et al., 2007; Nezer et al., 2016; Peterman et al., 2013; 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 datasets can be tricky (Aarts et al., 2012; Burnham & Anderson, 2002; 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 contain information about species presences, in contrast to PA data which records both where species have been found present and where they have not been found (Renner et al., 2015; Warton & Shepherd, 2010). Although PA data are generally of higher quality, it is also less common than PO data because it requires more rigorous planning to visit a set of predetermined sites. On the other hand, PO datasets are very common, arising from surveys or opportunistic sightings, but they usually have lower quality (Ruete & Leynaud, 2015; van Strien et al., 2013). Point process models (PPMs) are a common tool for fitting SDMs to analyze PO data (Mi et al., 2014; Renner et al., 2015; Warton & Shepherd, 2010) and have been used to fit models for real datasets and simulated data (Baddeley et al., 2006, 2015; Illian et al., 2012; Renner & Warton, 2013).

Unreliable or unknown species identification is the main concern in ecology especially for PO data from citizen science. Another issue can arise from confounded records when species taxonomy changes (Mahony et al., 2006). For example, *Mixophyes* frogs are now classified in three genetically distinct species while previously only one species was recognized. The *Mixophyes* frogs are not an isolated case. Padial and De la Riva (2006) noted that taxonomy inflation and new species discovery had contributed to an increase of 48.7% in new species of various organisms by that time. In particular, they refer to a study from Köhler et al. (2005) where amphibian species counts had increased by 25% from 1992 to 2004. This increase in reclassified and new species raises challenges to conservation biology (Catenazzi, 2015; Padial & De la Riva, 2006). Conservation planning efforts depend on clear identification of species and understanding of their distributions and habitat requirements (Franklin, 2013; Guisan et al., 2013). Other than cleaning datasets with missing information, little else is typically done in SDMs to account for misspecification. These practices can lead to missing information and thus incomplete predictions. Consequently, there are new challenges in building appropriate species distribution models for such species, for which the *Mixophyes* example serves as an illustration.

One way we can consider dealing with unknown species identities is to relabel them using mixture modeling or machine learning algorithms. Mixture modeling is a common tool used to represent complex distributions and aims to identify different groups within a dataset while modeling heterogeneity (Fernández Martinez, 2015; Hui, 2016). In communities or groups of species, it is possible to classify or cluster species according to covariate information through their preferences by using finite mixture modeling (Dunstan et al., 2013; Fernández-Michelli et al., 2016; Frame & Jammalamadaka, 2007; McLachlan & Peel, 2000). One particular application of this approach is to deal with over-dispersed data and to model ecological processes in parallel for different species (Matthews et al., 2001; Tracey et al., 2013; Zhang et al., 2004).

Machine learning algorithms are also becoming more common in statistical ecology because they can make use of unknown information and recognize specific structure in the data (Browning et al., 2018; Hastie et al., 2001; Thessen, 2016). Several algorithms exist such as unsupervised learning algorithms that can group observations with similar characteristics. Supervised learning algorithms use separate labeled datasets for classification and semisupervised learning algorithms learn from partially labeled data within the studied dataset to classify the observations (Fernández-Michelli et al., 2016; Vo et al., 2018; Wendel et al., 2015; Zhou, 2018). 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 apply machine learning procedures for classification procedures (Benaglia et al., 2009; Iovleff, 2018), but none accommodate the intersection of point process models with mixture modeling or machine learning algorithms.

In this paper, we develop new tools for fitting models to multispecies PO data with partial species identification by combining the PPM framework with mixture modeling and machine learning approaches to accommodate incomplete labeling. Our proposed methods rely on classification of points with unknown species labels based on the locations with known species labels. Hence, these methods will only assign classifications of known species in the region with verified species labels. The first 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*. The second tool fits mixtures of PPMs to all available data with an expectation–maximization (EM) algorithm and uses them to classify the unlabeled points. This method will be called *mixture 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. In this article, we will first define the new algorithms developed in Section 2. Then, we describe how we apply these methods to simulated data sets showcasing differences in abundance, correlation between species distributions, and percentage of data with unknown species labels in Section 3, as well as to the *Mixophyes* dataset we previously mentioned in Section 3.3. We present the results of these analyses in Section 4 and provide a discussion in Section 5.

2 | NEW MODELING METHODS

2.1 | Background

In ecology, we will consider a spatial point pattern as the distribution of species observation records over a specific window or study area $\mathscr{A}$. The point pattern intensity is defined as the density of points per unit area throughout $\mathscr{A}$. For Poisson point processes, we model the intensity of species i as a log-linear function of covariates $\mathbf{x}$:

$$\ln\left(\mu_i(s)\right) = \beta_{i,0} + \sum_j x_j(s) \times \beta_{i,j} \tag{1}$$

with $\mu_i(s)$ being the intensity of species i at location s. Here, x_j contains the values of covariate j, with which we associate the parameter $\beta_{i,j}$. We fit a point process model by maximizing the log-likelihood, as follows:

$$\ell\left(\boldsymbol{\beta}_i, s\right) = \sum_{1}^{m_i} \ln\left(\mu_i\left(s\right)\right) - \int_{\mathscr{A}} \mu\left(s\right) ds \qquad (2)$$

here $\boldsymbol{\beta}_i$ is the set of parameters associated with the covariates **x** and **s** is the set of $\mathbf{m}_i$ presence locations. The integral is intractable so we rely on numerical quadrature to get an estimate.

2.2 | Notation

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

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

Let $\mathbf{q} = \left\{s_{M+N+1}, \ldots, s_{M+N+Q}\right\}$ 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.

The proposed Loop and Mixture methods presented in Sections 2.3 and 2.4 assign some of the observations with uncertain species identities in **u** to the set of locations with known species identities in **s**, as in the right panel of Figure 1.

2.3 | 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 $\mathbf{a}_1 \in \mathbf{u}, \mathbf{a}_2 \in \mathbf{u}, \ldots, \mathbf{a}_K \in \mathbf{u}$, 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 $\hat{\mu}_i(s)$ for all $s \in \{\mathbf{s} \cup \mathbf{u}\}$ for $i \in \{1, \ldots, K\}$ maximizing the likelihood in ((2)).
3. Derive an $(M + N) \times K$ matrix of membership probabilities $\boldsymbol{\omega}$, where

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

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

$$\omega_i\left(s\right) = \begin{cases} 1\left(s \in \mathbf{s}_i\right) & : s \in \mathbf{s} \\ \dfrac{\hat{\mu}_i\left(s\right)}{\sum_{i=1}^{K} \hat{\mu}_i\left(s\right)} & : s \in \mathbf{u} \end{cases} \qquad (3)$$

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.

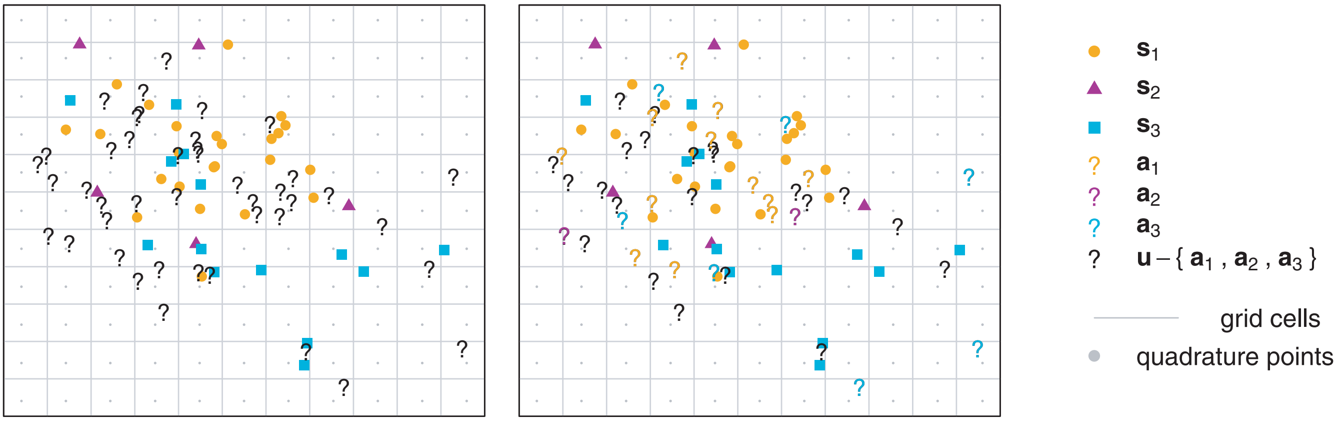


FIGURE 1 Three illustrative point patterns. The orange dots, purple triangles, and turquoise squares represent locations with known species identity, $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$. The gray dots represent quadrature points **q**, which are spaced evenly along a regular grid such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

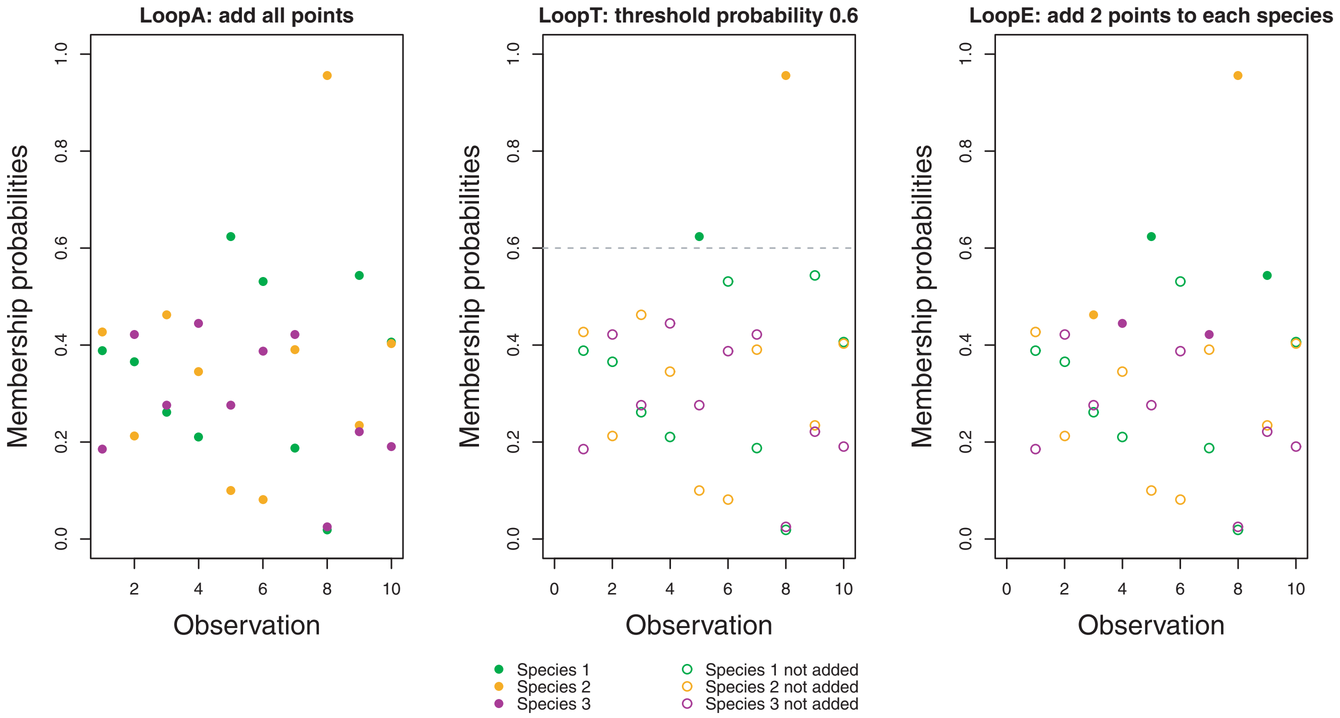


FIGURE 2 Points added to each species are represented by full circles, the ones that we do not add are open circles. (Left) LoopA function. We add all points with unknown species labels to each species. (Middle) LoopT function. We add all points with membership probabilities higher than the current threshold δ_h, set to 0.6 in the middle panel. (Right) LoopE function. We add the points a_h with the highest membership probabilities for each species, illustrated for $a_h = 2$ in the right panel

4. Define an augmented vector for species i as $\mathbf{y}_i = \mathbf{s}_i \cup \mathbf{a}_i$ for all $i \in \{1, \ldots, K\}$, where $\mathbf{a}_i$ consists of a subset of the vector of observations with unknown species labels $\mathbf{u}$. We define $\mathbf{a}_i$ as follows

- For the **LoopA** method, $\mathbf{a}_i = \mathbf{u}$ (left panel of Figure 2). The A in LoopA reflects the fact that we add all points in $\mathbf{u}$.
- For the **LoopT** method, $\mathbf{a}_i = \mathbf{u}_{[\omega_i(s) \geq \delta_h]}$, where δ_h 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 LoopT method augments the locations with known species identity i ($\mathbf{s}_i$) with the subset of locations with unknown species identity ($\mathbf{u}$) that have membership probabilities that are higher than the current threshold δ_h for the species i (middle panel of Figure 2). The T in LoopT reflects the fact that we add points with membership probabilities above a certain threshold.
- For the **LoopE** method, $\mathbf{a}_i = \mathbf{u}_{\left[\omega_i(s) \geq \boldsymbol{\omega}_{i,(M+N-a_h+1)}\right]}$, where $\boldsymbol{\omega}_{i,(j)}$ represents the jth smallest entry of vector $\boldsymbol{\omega}_i$, the ith column of $\boldsymbol{\omega}$, and a_h represents the number of locations to be augmented. That is, the LoopE method augments the locations with known species identity i ($\mathbf{s}_i$) with the subset of locations with unknown species identity ($\mathbf{u}$) with the a_h highest membership probabilities (right panel of Figure 2). The E in LoopE reflects the fact that we add an equal number of point for each species.

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

$$w_i(t) = \frac{c_1 \times c_2 \times \omega_i(t)}{1 + \sum_{s \in \{\mathbf{y}_i \cup \mathbf{q}\}} 1\left(c(s) = c(t)\right)\omega_i(s)} \tag{4}$$

This derivation of the 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 define the quadrature weight of the point at location t to be the product of the point's membership probability for the given species by the area of the grid cell, divided 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:

$$\Delta_{\ell_h} = \frac{\sum_{i=1}^{K} \left| \ell_h^i(\boldsymbol{\beta}) - \ell_{h-1}^i(\boldsymbol{\beta}) \right|}{\sum_{i=1}^{K} \ell_{h-1}^i(\boldsymbol{\beta})} < \varepsilon \tag{5}$$

for some choice of ϵ, where $\ell_h^i(\boldsymbol{\beta})$ is the fitted log-likelihood for the ith species at the hth iteration.

The maximum number of iterations varies for the different methods, as follows:

- For the **LoopA** method, the maximum number of iterations is set by the user. We set the default number of iterations to be 50.
- For the **LoopT** method, the maximum number of iterations is determined by:

$$\frac{\delta_{max} - \delta_{min}}{\delta_{step}} + 1 \tag{6}$$

- For the **LoopE** method, the maximum number of iterations is $\left\lfloor \frac{N}{K} \right\rfloor - a_1$, where $\lfloor c \rfloor$ rounds the number c down to the nearest integer, and a_1 is the first value of a_h chosen by the user. In the case of decimal numbers, only the floor is considered as we cannot add more points than available per species.

2.4 | Mixture of PPM method

Mixture methods can be fitted by maximizing a log-likelihood function and reclassifying the locations with uncertain identity using an EM algorithm framework. We developed the tool such that both soft and hard classification methodology are available. Various initialization schemes can be used, and we have chosen to use four different schemes, described below:

1. 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 of the unknown labeled location u to all the point locations $\mathbf{s}$. For each u, we consider the k closest neighbors in $\mathbf{s}$ regardless of species. Then, we calculate the membership probability of location s for species i using:

$$\omega_i(s) = \begin{cases} 1\left(s \in \mathbf{s}_i\right) & : s \in \mathbf{s} \\ \dfrac{z_i(s)}{\sum_{i=1}^{K} z_i(s)} & : s \in \mathbf{u} \end{cases} \tag{7}$$

where

$$z_i(s) = \sum_k \min_k \frac{1}{d_{i,k}(s)} \tag{8}$$

where the $d_{i,k}(s)$ are the kth distances for the species i at the location s.

- For the **kmeans method**, we define $\omega_i(s)$ as in ((7)) but define $z_i(s)$ as

$$z_i(s) = \frac{\min\left(d_i^C(s)\right)}{d_i^C(s)} \tag{9}$$

where $d_i^C(s)$ is the distance to the ith centroid of the ith cluster. The kmeans initialization is performed by the kmeans function in R, where we repeat the initialization multiple times as defined by the parameter nstart in R.

- For the **random method**, we define $\omega_i(s)$ as in ((7)) and $z_i(s)$ is drawn randomly from a uniform distribution:

$$z_i(s) : U\left[0, 1\right] \tag{10}$$

The random method is used as an uninformative approach for comparison to other methods.

- For the **CoinF**, we set the initial membership probabilities as follows:

$$\omega_i(s) = \begin{cases} 1\text{: } s \in \mathbf{y}_i \\ 0\text{: otherwise} \end{cases} \tag{11}$$

where we define the augmented vector $\mathbf{y}_i$ similarly to Step 4 of the Loop algorithm in Section 2.3 and $\mathbf{a}_i$ is defined as the vector of observations with unknown species labels randomly assigned to one of the species.

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

2. For soft classification: Create a list of point patterns, one for each species, each containing the locations with known identity $\mathbf{s}_i$ as well as the locations of the observations with unknown identity $\mathbf{u}$. For each point pattern, we define the quadrature weights as in ((4)), using the membership probabilities ω_i defined in Step 1.

For hard classification: Assign the locations in $\mathbf{u}$ to belong to one of the K species based on the membership probabilities $\boldsymbol{\omega}$. The classification is based on the highest membership probability.

3. Fit a point process model for each pattern defined in Step 2.
4. E step: Compute the predicted intensities $\hat{\mu}_i(s)$ for each species.
5. Calculate the predicted intensity of the mixture of K densities using:

$$v(s) = \sum_{i=1}^{K} v_i(s) = \sum_{i=1}^{K} \pi_i \times \hat{\mu}_i(s) \tag{12}$$

Here, $\widehat{\mu}_i(s)$ is the intensity at location s for the ith species and π_i is the mixing proportion or weight of the ith species and is given by:

$$\pi_i = \frac{\sum_{s \in y_i} \omega_i(s)}{\sum_{i=1}^{K} \sum_{s \in y_i} \omega_i(s)} \tag{13}$$

where $\omega_i(s)$ represents the membership probability of the ith species at the location s. The resulting $\nu_i(s)$ is thus the mixture intensity of the ith species.

6. Update the membership probabilities for the locations with unknown species identity $\mathbf{u}$ using

$$\boldsymbol{\omega}_i(s) = \begin{cases} 1\left(s \in \mathbf{s}_i\right) & : s \in \mathbf{s} \\ \dfrac{\nu_i(s)}{\sum_{i=1}^{k} \nu_i(s)} & : s \in \mathbf{u} \end{cases} \tag{14}$$

7. For soft classification, M step: Update the quadrature weights for all locations in $\mathbf{s}$ and $\mathbf{u}$ as in Step 2. If any location with an unknown label $u \in \mathbf{u}$ has a membership probability of $\omega_i(u) = 0$ for species i, that location is removed from the point pattern of species i before proceeding to the next step for the current iteration.

For hard classification, M step: Assign 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. For soft classification, fit an updated PPM using the updated quadrature weights and membership probabilities

For hard classification, compute a marked PPM based on the updated classifications.

9. Calculate the model log-likelihood using

$$\ell(\boldsymbol{\beta}) = \sum_{s \in \mathbf{s} \cup \mathbf{u}} f(s, \boldsymbol{\beta}) = \sum_{s \in \mathbf{s} \cup \mathbf{u}} \ln \sum_{i=1}^{K} \pi_i \times \nu_i\left(s, \beta_i\right) \tag{15}$$

where $f(s, \boldsymbol{\beta})$ is the mixture density function defined at locations $s \in \mathbf{s} \cup \mathbf{u}$ and parameterized by $\boldsymbol{\beta}$.

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

$$\frac{\left|\ell_h(\boldsymbol{\beta}) - \ell_{h-1}(\boldsymbol{\beta})\right|}{\left|\ell_{h-1}(\boldsymbol{\beta})\right|} < \varepsilon \tag{16}$$

where $\ell_h(\boldsymbol{\beta})$ is the log-likelihood at the hth iteration and ε is a prespecified tolerance level.

When the model has converged, we use hard classification for the locations $\mathbf{u}$ with unknown species identity when evaluating model performance.

3 | SIMULATION FRAMEWORK AND APPLICATION

3.1 | 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 vary based on sample size, the correlations ρ_{i-j} among the species distributions, and the proportion of observations with unknown labels, we simulated point patterns in which relative abundance patterns and correlation among distributions vary. In summary, we tested the following cases:

- test 1: $m_1 = 80$, $m_2 = 60$, $m_3 = 40$; $\rho_{1-2} = 0.09$, $\rho_{1-3} = -0.42$, $\rho_{2-3} = 0.20$;
- test 2: $m_1 = 60$, $m_2 = 60$, $m_3 = 60$; $\rho_{1-2} = 0.09$, $\rho_{1-3} = -0.42$, $\rho_{2-3} = 0.20$;
- test 3: $m_1 = 80$, $m_2 = 60$, $m_3 = 40$; $\rho_{1-2} = 0.85$, $\rho_{1-3} = -0.09$, $\rho_{2-3} = 0.20$;
- test 4: $m_1 = 60$, $m_2 = 60$, $m_3 = 60$; $\rho_{1-2} = 0.85$, $\rho_{1-3} = -0.09$, $\rho_{2-3} = 0.20$.

We chose low values for abundances as they would be small enough such that potential value of adding points with unknown species identities could be investigated. We chose these cutoffs for correlation to create clearly distinguishable contexts. We note that species are independent, and we do not investigate interactions between the species.

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. The hidden points form $\mathbf{u}$.

Simulations were conducted using the version 4.0.2 of R (R Development Core Team, 2017). We implemented 1,000 simulations of each of the 4 sets of test patterns previously described using a high performance computing cluster from the University of Newcastle, on 512 GB nodes powered by 3.0 GHz Intel Xeon Gold (E5-6154) processors. We display the Hard classification implementation hereafter because it showed the best performances. We tested the effects of different parameters on method performance for the following methods:

- knn: the value of k neighbors,
- kmeans: the number of random initializations *nstart*,
- LoopT: the maximum threshold $\delta_{\max}$, minimum threshold $\delta_{\min}$ and the step size δ_{step},

- LoopE: initial number of points added to the point pattern a_1.

We show how these parameters affect the performance in Appendix A.

3.2 | Suite of evaluation tools

We considered 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 chose the highest membership probability for each observation to determine the labeling of hidden points and compared it with its true label when computing the accuracy:

$$\text{Accuracy} = \frac{\text{Number of correct labels}}{N} \tag{17}$$

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

We also compared the final membership probabilities of the reclassified point labels of each point to 1 (the true weight) with a residual sum of squares (RSS).

$$\text{RSS} = \sum_{i=1}^{K} \sum_{s \in \mathbf{u} \cap \mathbf{t}_i} (\omega_i(s) - 1)^2 \tag{18}$$

where $\omega_i(s)$ is the final membership probability for location s for the reclassified point of species i computed using the methods outlined in Sections 2.3 and 2.4. Considering residual sum of squares (RSS) alone does not provide a reliable comparison because the number of unknown observations can vary, so we considered meanRSS instead to standardize the measure for all fitted models:

$$\text{meanRSS} = \frac{\text{RSS}}{N} \tag{19}$$

We also obtained these performance measures for models fitted using only the locations with known species identity, hereafter referred to as the "individual PPM" method. In this way, we have a baseline with which to judge whether the mixture and Loop methods outperform the standard approach of discarding points with unknown species identity.

We also computed performance measures based on predicted intensities. We compared the true distribution from which we generated the points to the predicted distributions of the various models we fitted. We used 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) (Es, 1997; Swanepoel, 1988). The function is defined as:

$$\text{IMSE} = E\left(\int_{-\infty}^{+\infty} (\hat{f}_n(x) - f(x))^2 dx \right) \tag{20}$$

where $\hat{f}_n(x)$ is an estimator of the density function $f(x)$. Because the scale of the IMSE depends on the magnitude of the true intensity, we rescaled both true and predicted intensities to have a common mean to make for an equitable comparison. We computed the IMSE using the values of the true and predicted intensities at the quadrature points $\mathbf{q}$, and sum across the three species (sumIMSE):

$$\text{sumIMSE} = \sum_{i=1}^{K} \sum_{q=1}^{Q} (\hat{\mu}_i^{\bar{\mathbf{t}}_i}(s_{M+N+q}) - \mu_i(s_{M+N+q}))^2 \tag{21}$$

where $\hat{\mu}_i^{\bar{\mathbf{t}}_i}(s)$ is the predicted intensity of species i at location s rescaled to have mean $\bar{\mathbf{t}}_i$. We also displayed the standard error of the point predictions as a measure of uncertainty. We weight the standard error measure by the number of points for an equitable comparison across different percentages of hidden observations.

3.3 | Application to eastern Australian frogs

Our study case dataset uses presence-only records from the online database of the Atlas of Living Australia (ALA, 2018). On this platform, any person that sees a frog in the wild can report the coordinates and other relevant information. We focused the analysis on the three northern species of *Mixophyes* genus that have been recently separated in Mahony et al. (2006). We cleaned our dataset by including only observations of adult specimens with date information and through verification by a specialist of these species, M. Mahony. The observations with known species labels were those for which we have associated genetic information as well as any observations reported after the taxonomic split in 2006. The rest of the observations were considered as having unknown species labels. We also included data from Oza et al., (2012) as part of the known labeled points. Altogether, we count 181 out of the 444 observations with unknown labels (approximately 40.8%).

We extracted relevant covariates for these species on a 5 km × 5 km grid from different sources as presented here (Table 1).

We fitted models using the methods that performed best in the simulation study and compared them with the individual PPM method for which no points with unknown labels were used.

4 | RESULTS

Here, we present the model performances on the simulated data parameters (abundance, correlation, and percentage of points with hidden species labels). We explore the role of different parameters within the various mixture and loop methods in Appendix A. The individual PPM results will be used as a point of comparison with the other methods as the individual PPM method does not include any of the points with unknown labels. We choose to use Pearson's correlation coefficient when computing sumcor. We conclude the section by comparing maps and membership probabilities of the *Mixophyes* species.

TABLE 1 Description and origin of the different covariates used in the analysis of the *Mixophyes* dataset

Name	Description	Source
Bio05	Max Temperature of Warmest Month	BBCVL
Bio06	Min Temperature of Coldest Month	BBCVL
Bio11	Mean Temperature of Coldest Quarter	BBCVL
Bio13	Precipitation of Wettest Month	BBCVL
Bio18	Precipitation of Warmest Quarter	BBCVL
Altitude	Altitude	BBCVL
Dist road	Distance to the nearest roads	UC Davis Biogeo group
Dist stream	Distance to the nearest hydrological features	Bureau of Meteorology (2014)

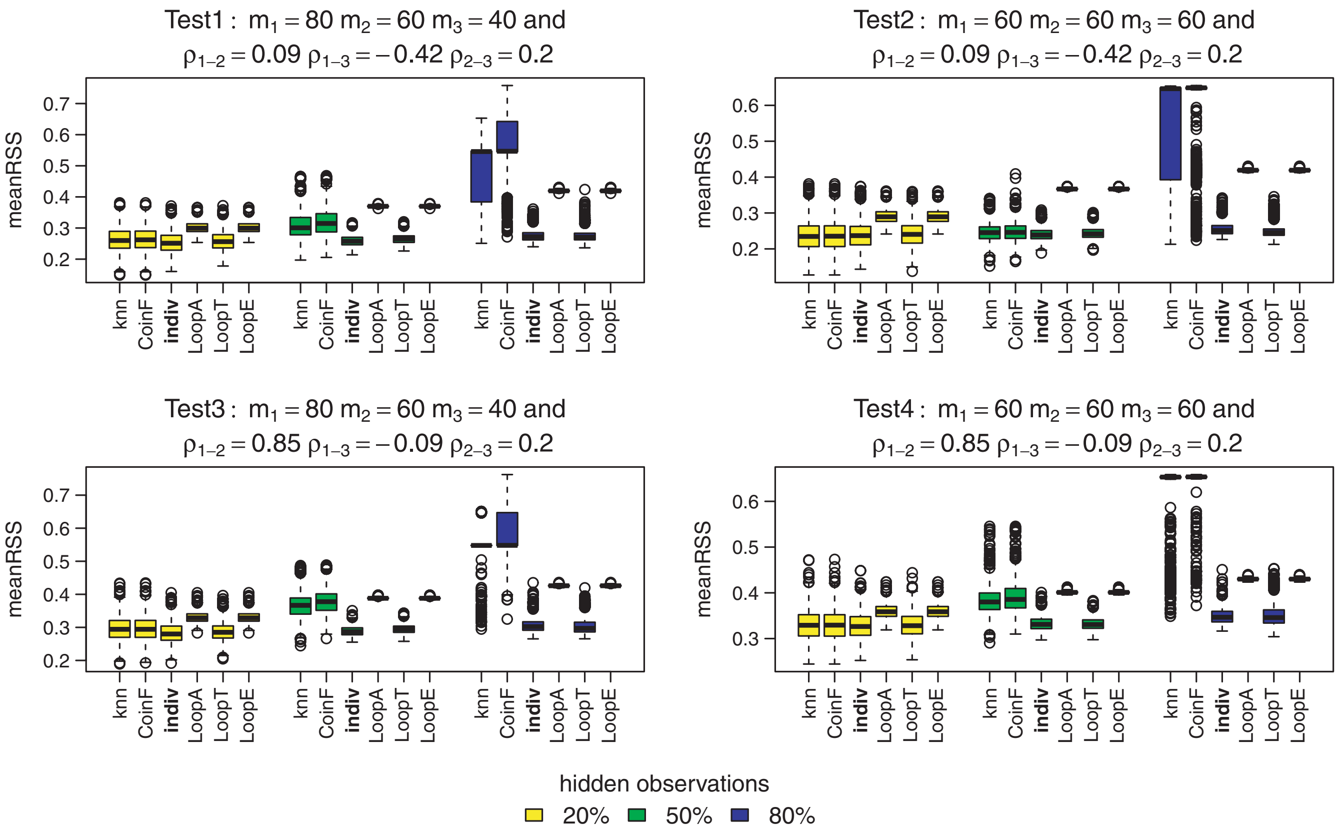
4.1 | Testing species distributions

In this section, we compare the results of varying abundance, the correlation between species distributions, and the percentage of hidden observations on the performance measures and membership weights for classification as presented in Section 3. We only present the best-performing methods in this section: knn mixture, LoopA, LoopT, LoopE, and the individual PPM and coinF method for reference.

4.1.1 | Relabeling performance measures

In terms of relabeling, only LoopT consistently performs as well or better than the individual PPM method across all simulation designs and percentage of hidden observations, as shown in Figure 3. The mixture methods are more competitive than the LoopA and LoopE methods at 20% and 50% of hidden observations but still do not perform as well as the individual PPM or LoopT methods.

Comparing accuracy, all three Loop methods perform comparably to the individual PPM method. The knn and coinF methods are equally competitive at 20% of hidden observations but their performances get worse than the other methods for 50% and 80% percentages in Figure 4.


Test1 : $m_1 = 80$ $m_2 = 60$ $m_3 = 40$ and $\rho_{1-2} = 0.09$ $\rho_{1-3} = -0.42$ $\rho_{2-3} = 0.2$

Test2 : $m_1 = 60$ $m_2 = 60$ $m_3 = 60$ and $\rho_{1-2} = 0.09$ $\rho_{1-3} = -0.42$ $\rho_{2-3} = 0.2$

Test3 : $m_1 = 80$ $m_2 = 60$ $m_3 = 40$ and $\rho_{1-2} = 0.85$ $\rho_{1-3} = -0.09$ $\rho_{2-3} = 0.2$

Test4 : $m_1 = 60$ $m_2 = 60$ $m_3 = 60$ and $\rho_{1-2} = 0.85$ $\rho_{1-3} = -0.09$ $\rho_{2-3} = 0.2$

FIGURE 3 MeanRSS for the best methods: knn, coinF, individual PPM (reference), LoopA, LoopT, and LoopE. 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%. For each method, we fitted models to the three simulated point patterns using four simulated predictors. A low meanRSS value indicates a high performance

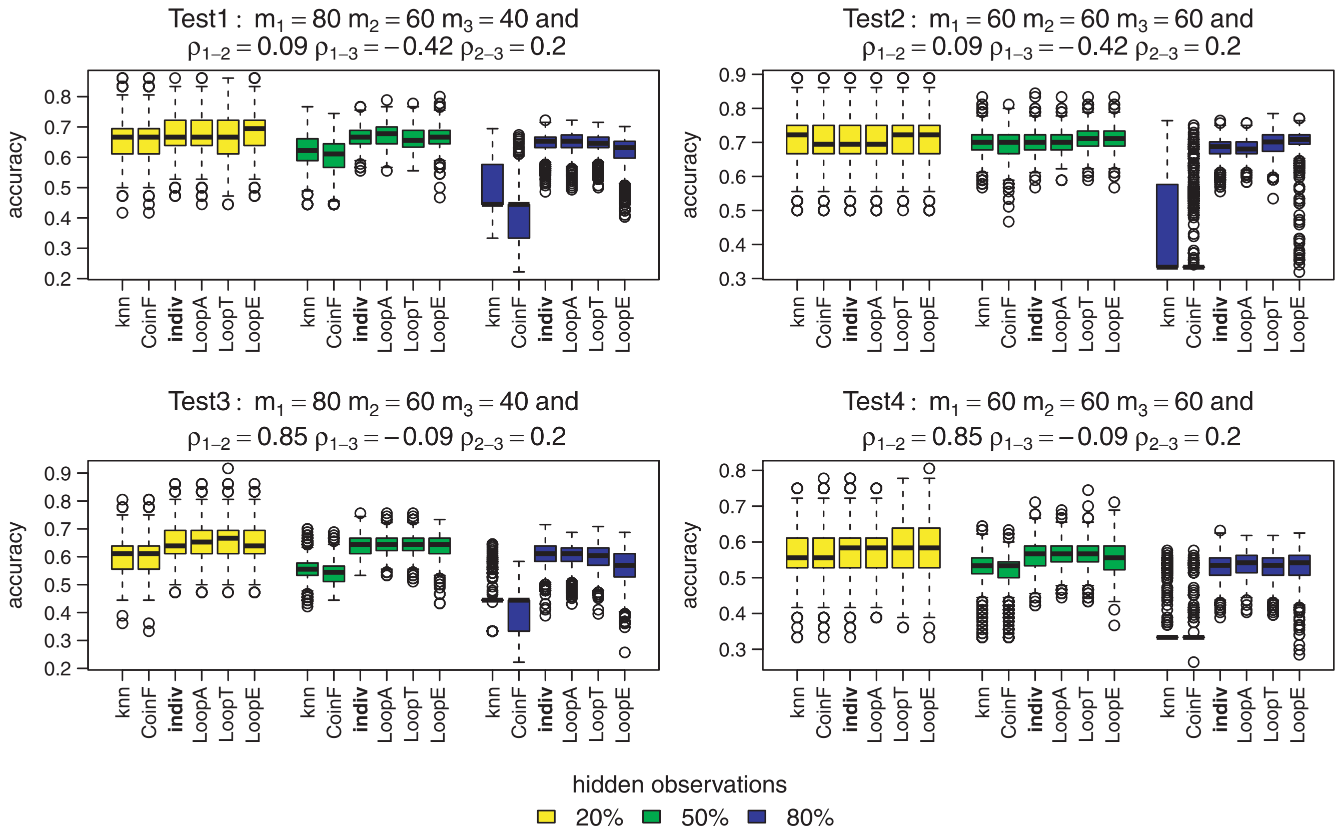


FIGURE 4 Accuracy for the best methods: knn, coinF, individual PPM (reference), LoopA, LoopT, and LoopE. Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50% and in blue for 80%. For each method, we fitted models to the three simulated point patterns using four simulated predictors. A high accuracy value indicates a high performance

4.1.2 | Predicted intensity performance measures

Now, we consider performance based on predicted intensity. The LoopT method performs as well or better than the individual PPM method according to sumIMSE as shown in Figure 5. The LoopA, LoopE, knn, and CoinF methods are mostly never competitive with the other methods at high percentage of hidden observations.

The relative performance is different when using sumcor as the performance measure as shown in Figure 6. It looks like LoopT is consistently best, and the individual PPM method and LoopE methods are broadly comparable for nonhighly correlated distributions. The knn and coinF methods perform almost equally to the individual PPM method when a relatively low percentage of observations have hidden labels and when distributions are highly correlated.

Comparisons of the estimated standard errors appear in Appendix A. Standard errors for the predicted intensities increase, as expected, when the number of observations used in the models decreases, as shown in Figures A5 and A6. This is evident from the higher standard errors for higher percentages of observations with hidden labels as well as for the individual PPM method, which does not add any points.

4.1.3 | Final membership probabilities and classification

Figures 7-10 show the final membership probabilities of the locations with hidden species identity corresponding to each species. The higher the membership probability is to 1, the better the classification performance. It appears that the high correlation among the species distributions as in tests 3 and 4 results in lower classification performance. When there are differences in abundance (test 1 and test 3), the mixture methods seem to show superior performance for the most abundant species and worse performance for the least abundant species.

4.2 | The *Mixophyes* case

4.2.1 | Prediction of *Myxophies*' species distribution

In this section, we fit the best-performing method within each category (knn among the mixture methods and LoopT among the Loop methods) to analyze the distribution of the *Mixophyes* species and compare the predictions to the individual PPM approach in which

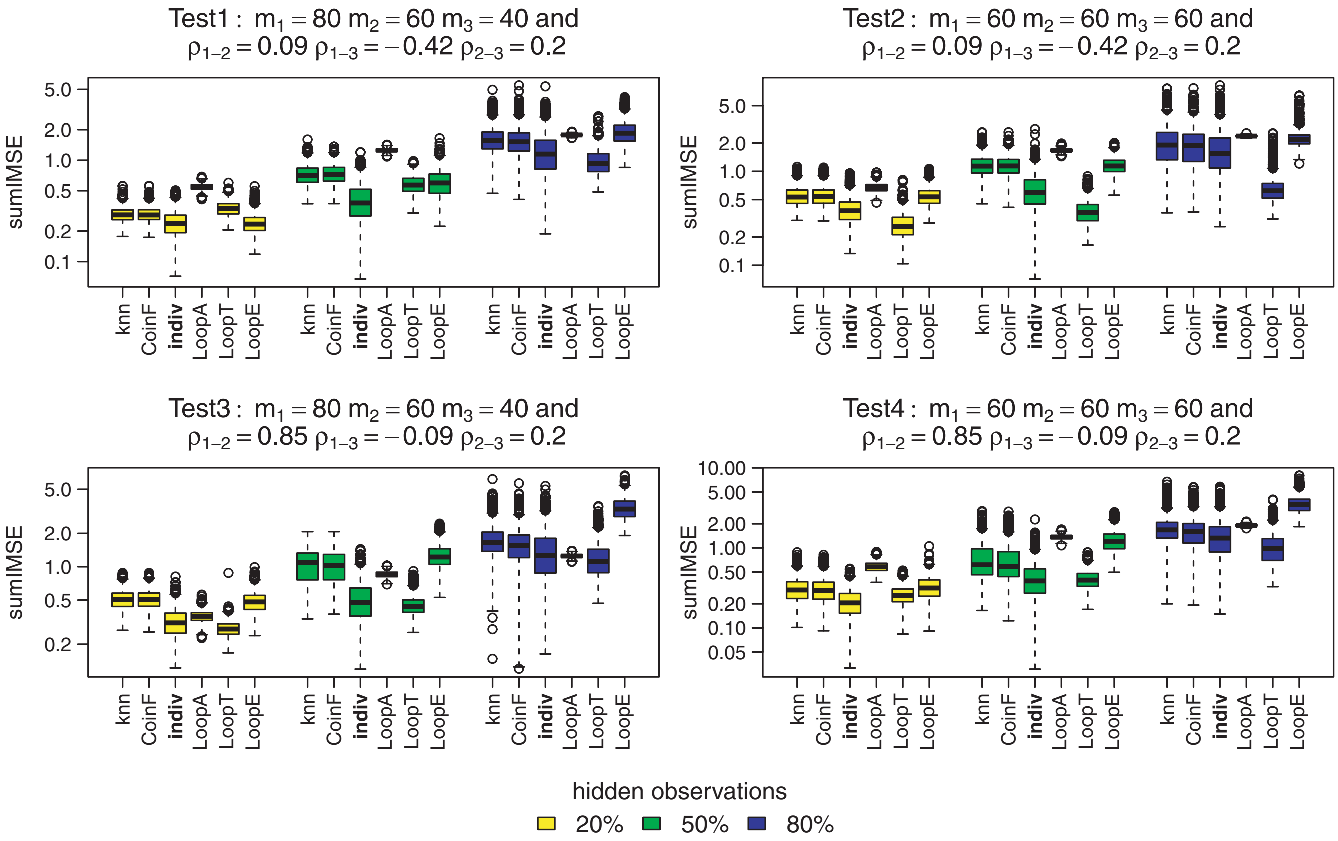


FIGURE 5 SumIMSE (logarithmic scale) for the best methods: knn, coinF, individual PPM (reference), LoopA, LoopT, and LoopE. Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50% and in blue for 80%. For each method, we fitted models to the three simulated point patterns using four simulated predictors. A low sumIMSE value indicates a high performance

no unlabeled observations are included in the model. The resulting fitted intensity maps are shown in Figure 11. Both the knn mixture method and the LoopT method add small areas of distribution for *Mixophyes schevilli*. The maps from the LoopT method show increased areas of relatively high intensity in the south for *Mixophyes carbinensis* and *Mixophyes coggeri*.

4.2.2 | Classification of *Myxophies* observations

Differences in the predicted distributions are also shown by the classification of the locations with uncertain identities in Figure 12. While there is broad agreement in the south for the knn mixture method and the LoopT method, the LoopT method classifies more records as *M. coggeri* in the north and *M. carbinensis* in the central part, while the knn mixture method classifies more records as *M. schevilli* in the north and central parts. This may reflect the fact that the mixture methods tend to have high classification for the most abundant species, and *M. schevilli* had the highest number of verified records among the three species.

The colors of the question marks in Figure 12 are based on the final membership probabilities, with higher membership probabilities leading to bolder colors. This Figure indicates that the mixture knn method tends to result in lower membership probabilities than the LoopT method except for the most abundant species *M. schevilli*, which is also supported by Figure 13, in which the final membership probabilities for the LoopT method tend to be more variable, with the third quartile markedly higher for each species. The final membership probabilities appear more balanced for the LoopT method, whereas the knn mixture method tends to favor the most abundant species, *M. schevilli*.

5 | DISCUSSION

In this article, we present a new modeling framework implemented 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 modeling and a machine learning framework with 7 different implementation methods overall. We tested our algorithms in different contexts where we vary the abundances of our species (equal across the species or different), the correlation between them (two distributions are highly correlated or all have low correlation), and the proportion of unknown species identities (20%, 50%, and 80%). We compared our methods with the individual PPM method which

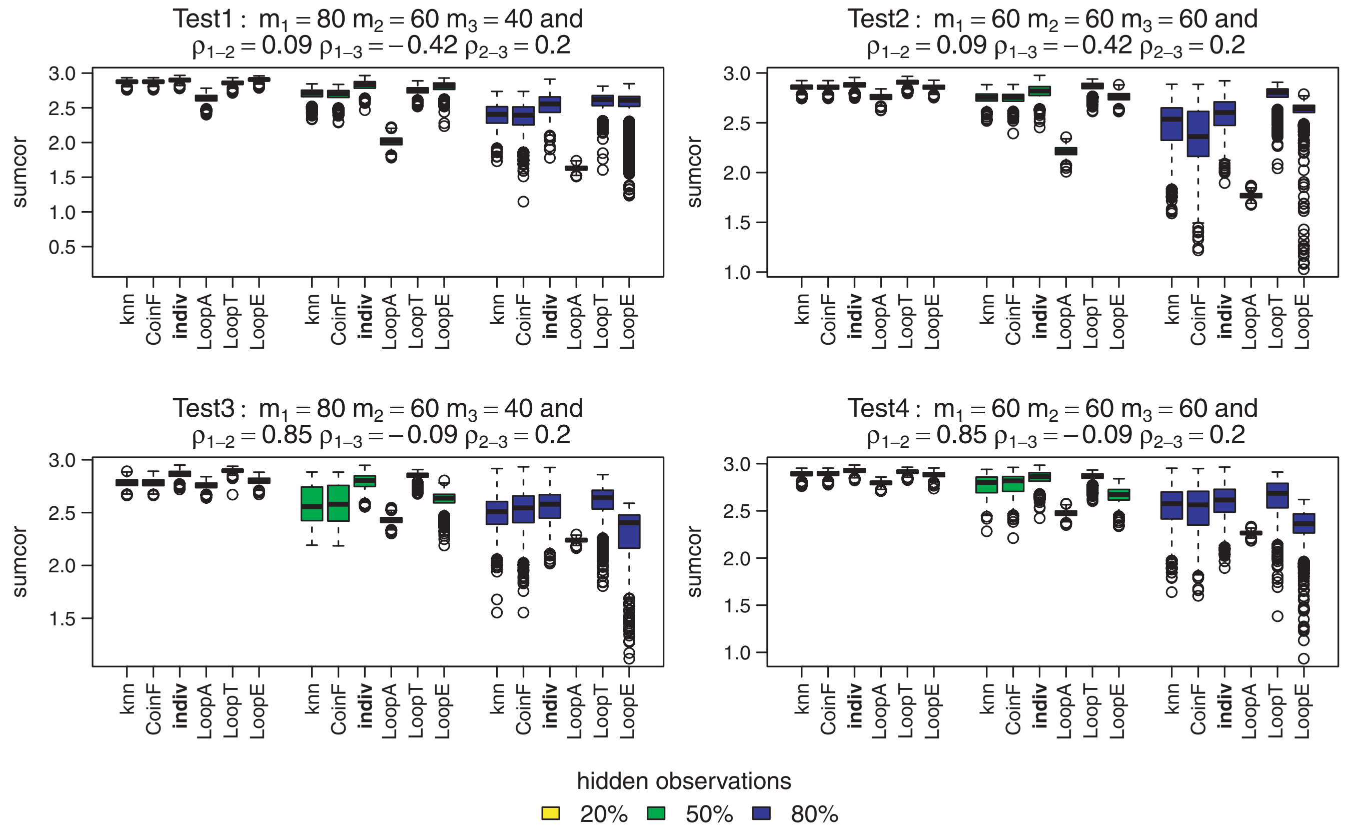


FIGURE 6 Sumcor for the best methods: knn, coinF, individual PPM (reference), LoopA, LoopT, and LoopE. Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50% and in blue for 80%. For each method, we fitted models to the three simulated point patterns using four simulated predictors. A high sumcor value indicates a high performance

ignores locations with unknown species identities to see whether the proposed algorithms allow us to fit distributions that are closer to the initial processes.

The novelty of these methods makes it difficult to compare to other existing tools that do not combine point pattern processes, mixture models, or semisupervised learning methods. Most mixture models also use the EM algorithm but are implemented for Gaussian mixture distributions only (Benaglia et al., 2009; Di Zio et al., 2007; Quost & Denoeux, 2016; Scrucca et al., 2016). A few implementations use both mixture and semisupervised learning but do not use presence-only data or point pattern processes (Figueirido & Jain, 2002; Frame & Jammalamadaka, 2007; Melnykov & Maitra, 2010; Woillez et al., 2012). Flexible R packages such as Flexmix (Leisch, 2004), mixtools (Benaglia et al., 2009) and MixAll (Iovleff, 2018) are not suitable to our design. The work of Taddy and Kottas (2012) is noteworthy in that it models a mixture of marked point processes in a Bayesian framework, but it does not allow for semisupervised learning and therefore cannot accommodate settings such as ours in which some points have unknown species labels. However, as the goal is to investigate whether there is any benefit from adding points with unknown species labels when fitting models, comparison to the individual PPM method which does not add any unlabeled points allows us to compare the proposed methods to a natural baseline.

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. The results show that some methods benefit from adding points with unknown species labels, leading to improved performances. We noticed a discrepancy in performances that is 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 methods performed the best. In particular, the LoopT method showed consistently good performances across all measures studied. For this method, only the points with the highest membership probabilities are added. We explore the roles of the δ_h parameters in Appendix A. We set the maximum and minimum thresholds at $\delta_{\max} = 0.5$ and $\delta_{\min} = 0.1$ and a step size of $\delta_{\text{step}} = 0.1$ as it appears to be the best combination. LoopE showed competitive results to LoopT or the individual PPM method looking at predictive performances in the case of not highly correlated distributions, but may not be able to distinguish highly correlated distributions. For this method, we add the a_h points with highest membership probabilities, with the number of points a_h increasing at each iteration. We explore the role of this parameter in Appendix A. For the LoopA method, we add all unknown points to the known points; thus, the reclassification and

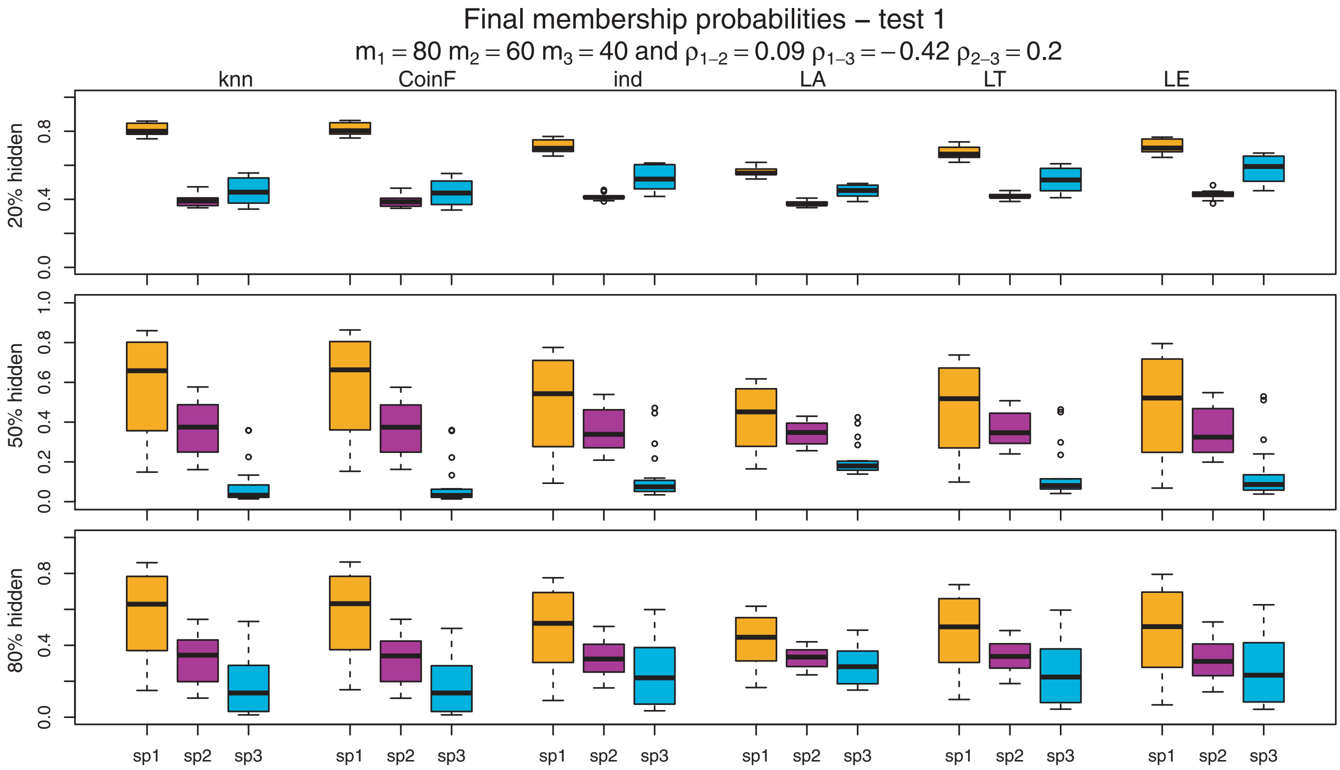


FIGURE 7 The boxplots display the estimated membership probabilities of the correct species for points with hidden labels in test 1. Each color boxplot represents a different species. Each row corresponds to the different percentage of hidden observations tested: 20%, 50%, and 80%. Test 1 is based on simulated point patterns with abundances of $m_1 = 80$, $m_2 = 60$, $m_3 = 40$; and correlations between the species distributions of $\rho_{1-2} = 0.09$, $\rho_{1-3} = -0.42$, $\rho_{2-3} = 0.20$

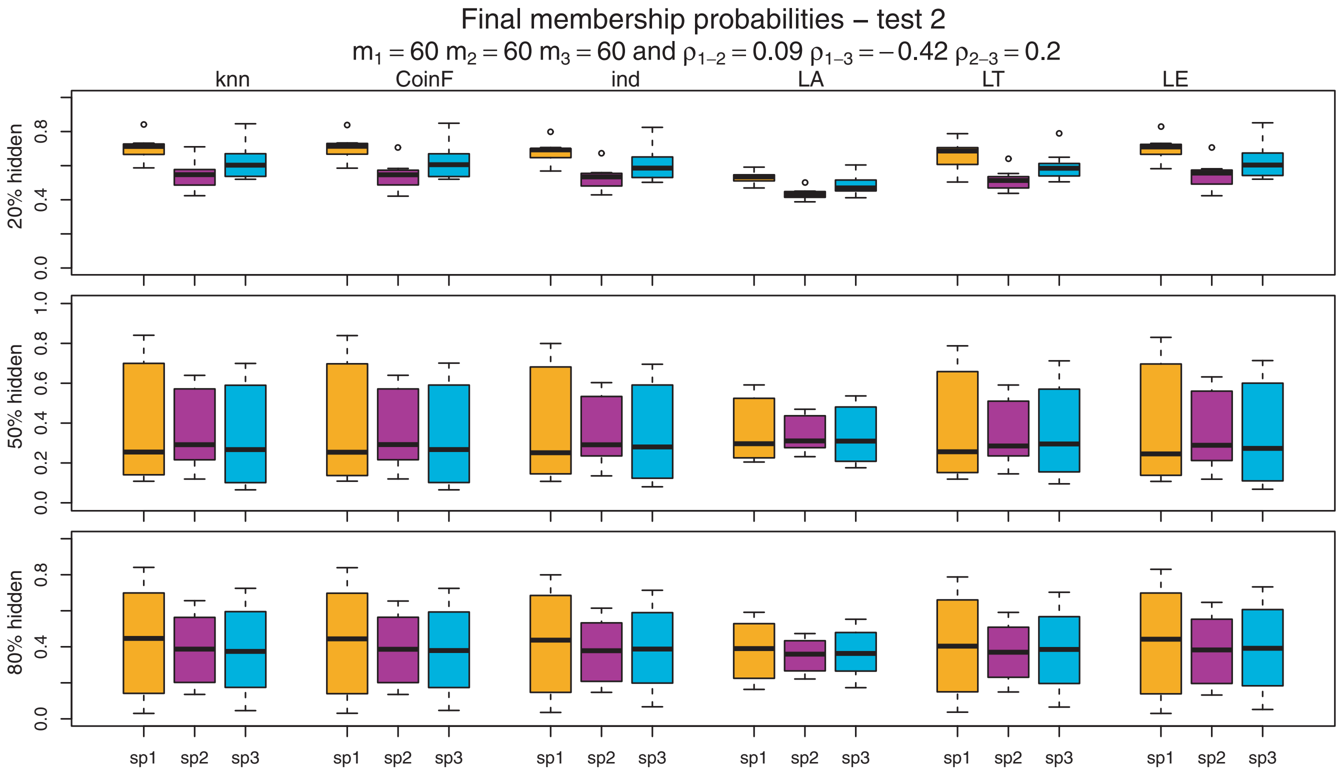


FIGURE 8 The boxplots display the estimated membership probabilities of the correct species for points with hidden labels in test 2. Each color boxplot represents a different species. Each row corresponds to the different percentage of hidden observations tested: 20%, 50%, and 80%. Test 2 is based on simulated point patterns with abundances of $m_1 = 60$, $m_2 = 60$, $m_3 = 60$; and correlations between the species distributions of $\rho_{1-2} = 0.09$, $\rho_{1-3} = -0.42$, $\rho_{2-3} = 0.20$

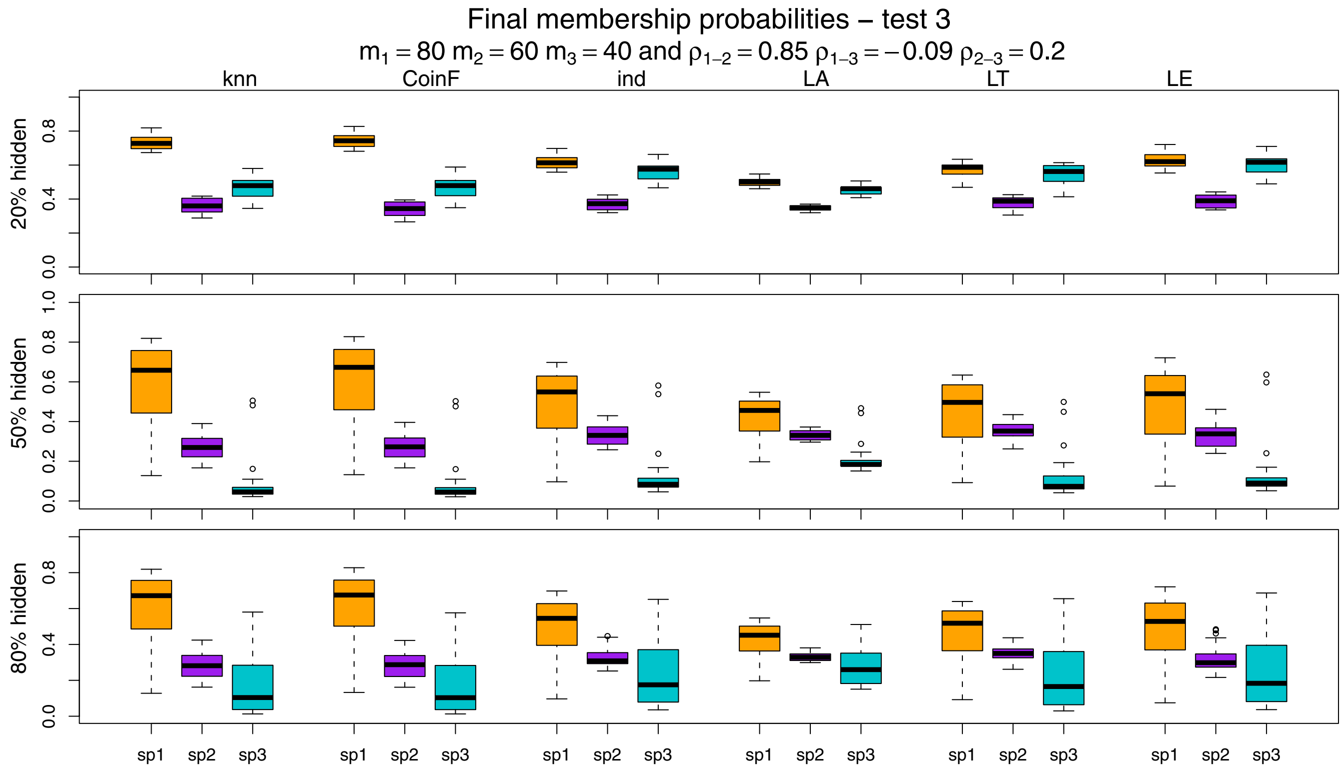


FIGURE 9 The boxplots display the estimated membership probabilities of the correct species for points with hidden labels in test 3. Each color boxplot represents a different species. Each row corresponds to the different percentage of hidden observations tested: 20%, 50%, and 80%. Test 3 is based on simulated point patterns with abundances of $m_1 = 80$, $m_2 = 60$, $m_3 = 40$; and correlations between the species distributions of $\rho_{1-2} = 0.85$, $\rho_{1-3} = -0.09$, $\rho_{2-3} = 0.20$

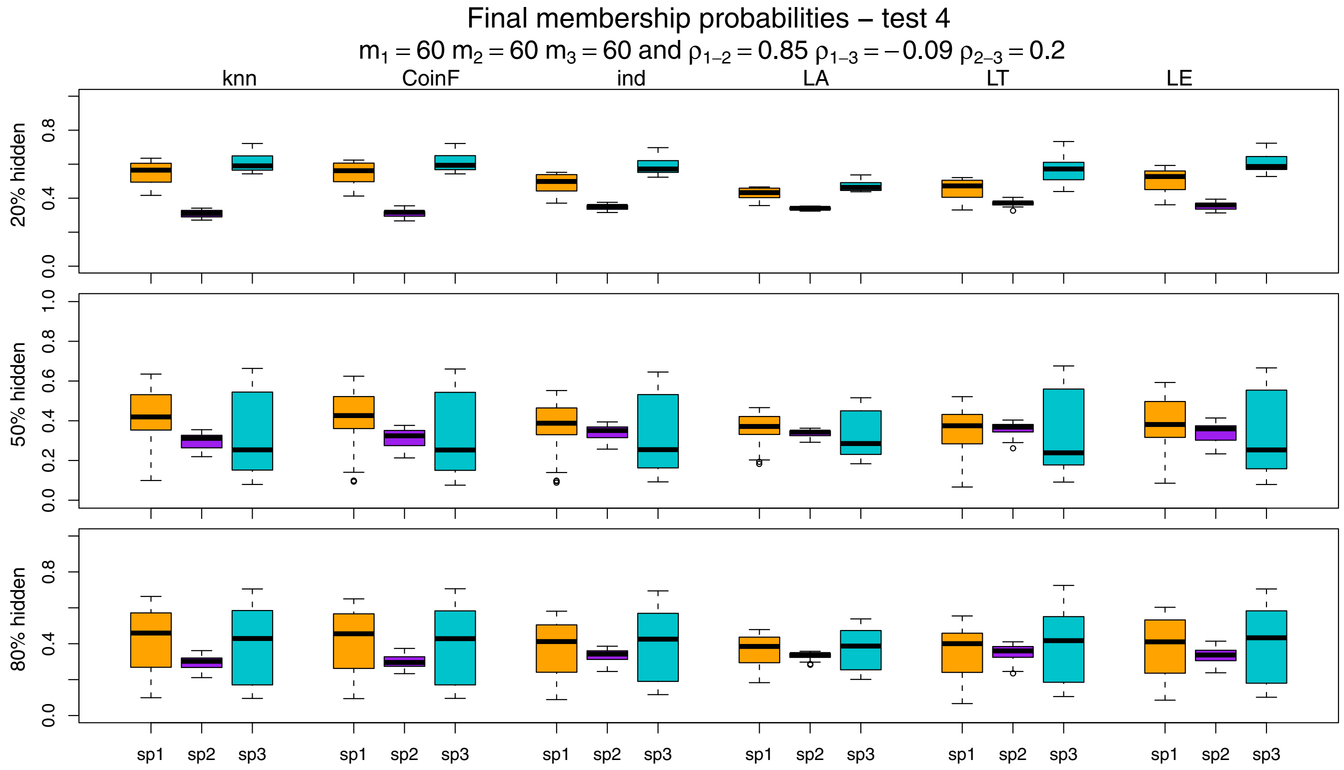


FIGURE 10 The boxplots display the estimated membership probabilities of the correct species for points with hidden labels in test 4. Each color boxplot represents a different species. Each row corresponds to the different percentage of hidden observations tested: 20%, 50%, and 80%. Test 4 is based on simulated point patterns with abundances of $m_1 = 60$, $m_2 = 60$, $m_3 = 60$; and correlations between the species distributions of $\rho_{1-2} = 0.85$, $\rho_{1-3} = -0.09$, $\rho_{2-3} = 0.20$

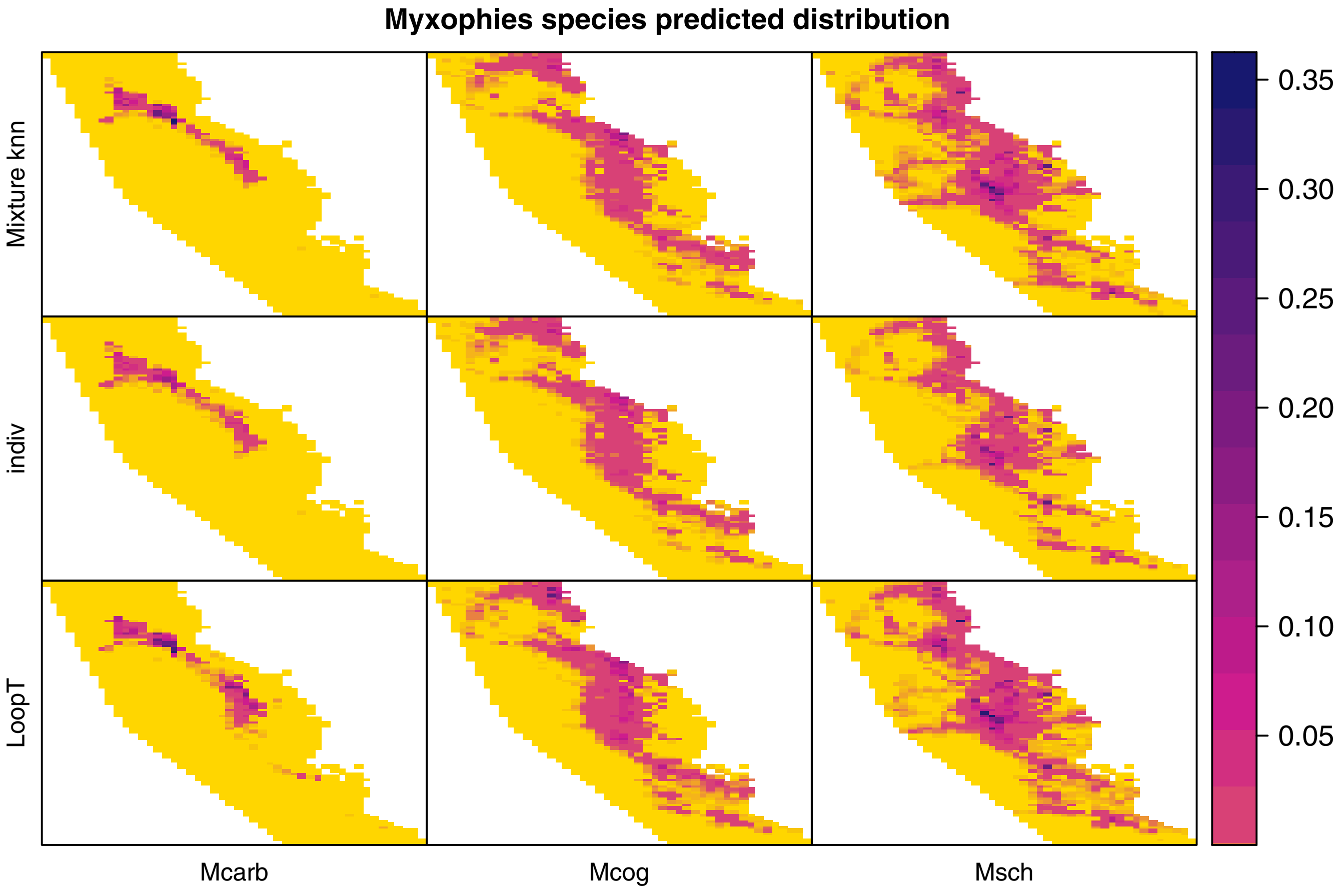


FIGURE 11 Distribution of the *Mixophyes* species predicted intensities for the mixture knn initialization method, the individual PPM method without the reclassified points and the LoopT method. Mcarb stands for *Mixophyes carbinensis*, Mcog stands for *Mixophyes coggeri*, and Msch stands for *Mixophyes schevilli*

density distribution can be biased by the unknown points not belonging to a certain species. The LoopA method displayed similar results as the mixture methods for predictive performances (sumIMSE) but outperformed these methods in relabeling.

The methods using the mixture algorithm tend to perform worse than the Loop methods and the individual PPM method for moderate and high percentages of hidden observations (50% and 80%). However, mixture methods performed relatively better for highly correlated distributions in their predictive performances, which relate to the mixture methods' ability to distinguish multiple distributions inside one distribution. We note that the mixture methods displayed high membership probabilities for the most abundant species. Indeed, the method makes use of mixing proportions, which give further emphasis to the most dominant species. Hence, they tend to favor the most abundant species while not classifying well the other species with lower abundances. The methods (kmeans, random) not presented previously in the results are presented in Appendix A (see Figures A1-A4). All methods showed relatively similar performance to each other across all measures.

Contrary to what we found, 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). Studies link the performance of the knn method to the metric chosen to calculate the nearest neighbor distances and the value of the number *k* of nearest neighbors (Guo et al., 2003; Weinberger & Saul, 2009; Wu et al., 2008). Even the established kmeans method shows drawbacks as its performance depends on overlapping densities and whether the distributions are roughly circular. The choice of the centroid is also not consistent and chosen at random for the first calculation (Wu et al., 2008; Yoo et al., 2012). The coinF method, which randomly assigns species labels, is in line with the other mixture methods and never reaches the performance of the Loop methods. Consequently, we have shown that the loop methods outperform not only the individual PPM method but also a method that randomly assigns species labels. A future research area could look into the different metrics to evaluate nearest neighbors (knn) or the centroid choice (kmeans).

We also tested the best-performing method LoopT and the knn method on the *Mixophyes* dataset. As mentioned in the results, the knn method will favor the most abundant species of the dataset and in our context assigned more points with unknown species labels to *M. schevilli*, while the LoopT method produced more balanced species assignments. The value of the proposed Mixture and Loop methods is to make use of observations with uncertain

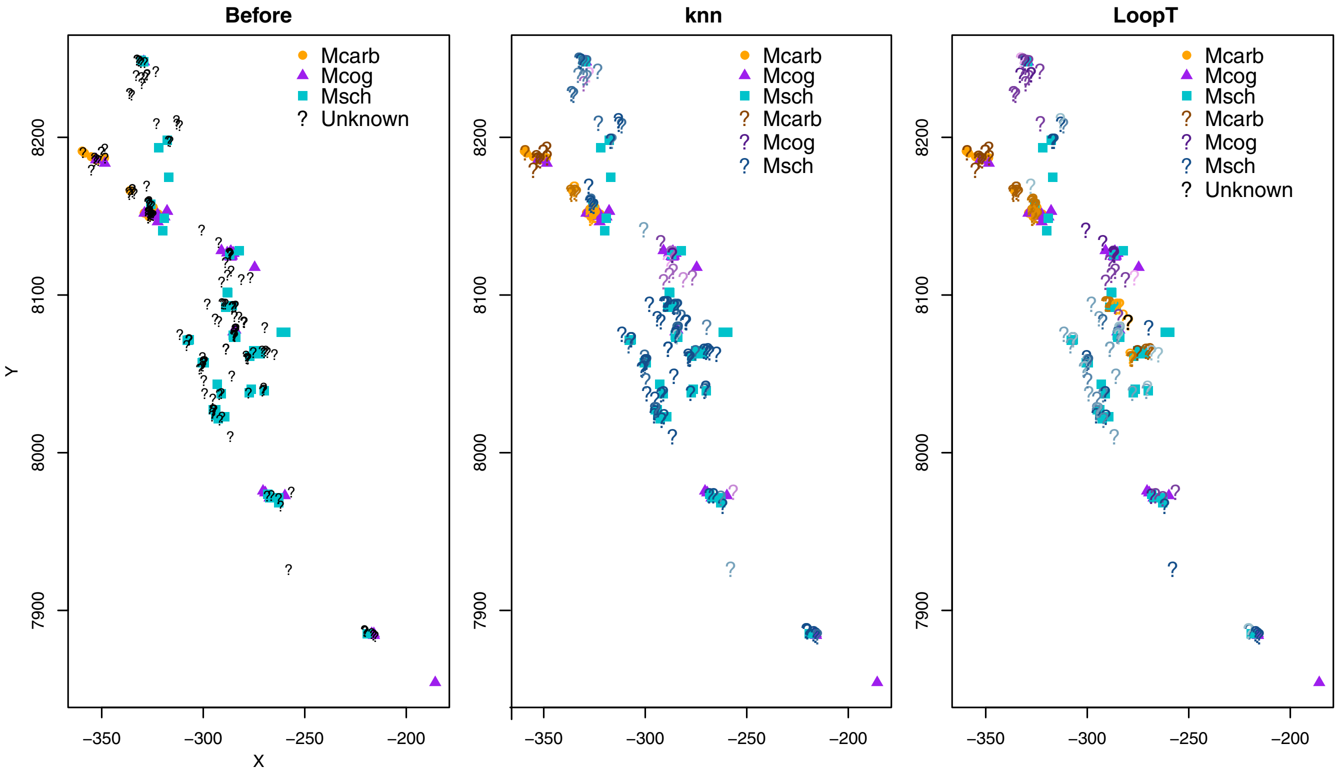


FIGURE 12 Observed locations for the *Mixophyes* data set. On the left, the points with unknown species labels have not been classified. The remaining maps show the final classification for the knn mixture (middle) and LoopT (right) methods. The orange dots, purple triangles, and blue squares represent labeled points, while the question marks represent the points with unknown labels. The color of the question marks indicates their classification, with black representing an unclassified point, and the intensity of colored question marks representing the final membership probabilities, with bolder colors representing higher probabilities

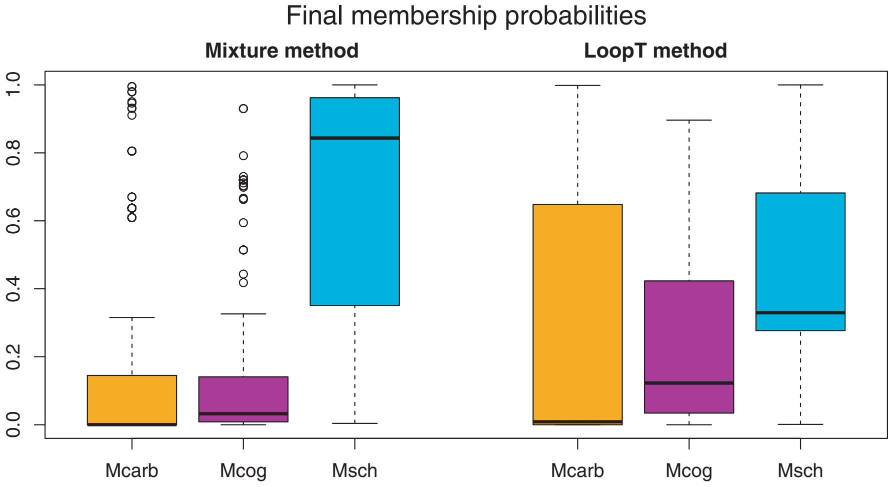


FIGURE 13 Final membership probability per species for the knn mixture method and the LoopT method. Each color represents a different species: Mcarb; *M. carbinensis*, Mcog; *M. coggeri* and Msch; *M. schevilli*

species identities, and our results suggest that the LoopT method provides the best combination of accuracy in prediction and classification.

There are more factors to consider for real ecological datasets that can influence how a model will perform. First, the abundances tested in the simulation are quite low (40 points at the lowest) 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. In our model, we included all covariates used to generate the true point patterns; however, for real datasets we may not have access to the best covariates or know which ones precisely determine the species distributions. Applying a lasso penalty to help in variable selection may therefore provide a natural way forward in this context. Finally, a key reality when dealing with presence-only data is the observer bias, for 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 proposed methods could be extended to accommodate these approaches of accounting for observer bias.

6 | 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 ones 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. They can be adapted to account for other considerations in modeling presence-only data.

ACKNOWLEDGMENTS

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

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

AUTHOR CONTRIBUTIONS

Emy Guilbault: Conceptualization (equal); Data curation (supporting); Formal analysis (lead); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing—original draft (lead); Writing—review and editing (equal). **Ian Renner:** Conceptualization (equal); Formal analysis (supporting); Methodology (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (supporting); Writing—review and editing (equal). **Michael Mahony:** Conceptualization (supporting); Data curation (equal); Resources (supporting); Supervision (supporting); Writing—review and editing (supporting). **Eric Beh:** Project administration (supporting); Supervision (supporting); Writing—review and editing (supporting).

ETHICAL APPROVAL

This does not apply to our research.

OPEN RESEARCH BADGES

This article has earned a Preregistered Badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data are available at https://github.com/EmyGlblt/LoopMixArticle.

DATA AVAILABILITY STATEMENT

Data: The different occurrences come from the Atlas of Living Australia. The data were completed and the coordinates were verified by Michael Mahony. Observations coming from? for the same species were also added to the dataset. Data locations, the environmental covariates, and the R script to analyze the data are available at: https://doi.org/10.5061/dryad.vx0k6djqw. Rscript: An example of the scripts used for the simulation is available here: https://github.com/EmyGlblt/LoopMixArticle. A Rmarkdown document details the steps and the implementation of our functions in the same github location.

ORCID

Emy Guilbault https://orcid.org/0000-0001-6881-9926
Ian Renner https://orcid.org/0000-0003-3116-2486

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

ALA (2018). *Atlas of living Australia occurrence downloaded*. http://www.ala.org.au

Baddeley, A., Gregori, P., Mateu, J., Stoica, R., & Stoyan, D. (2006). Modelling spatial point patterns in R. In A. Baddeley, P. Gregori, J. Mateu, R. Stoica, & D. Stoyan (Eds.), *Case studies in spatial point process modeling* (Vol. *185*), Lecture Notes in Statistics. Springer. 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. 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

Bureau of Meteorology (2014). *Commonwealth of Australia*. http://www.bom.gov.au/water/geofabric/

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer. https://doi.org/10.1007/b97636

Catenazzi, A. (2015). State of the world's amphibians. *Annual Review of Environment and Resources*, *40*, 91–119. https://doi.org/10.1146/annurev-environ-102014-021358

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 Martinez, D. (2015). *Mixture-based clustering for the ordered stereotype model*. Thesis, School of Mathematics Statistics and Operations Research, Victoria University of Wellington, Wellington. https://doi.org/10.13140/RG.2.1.1945.4806

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

Frame, S. J., & Jammalamadaka, S. R. (2007). Generalized mixture models, semi-supervised learning, and unknown class inference. *Advances in Data Analysis and Classification*, *1*(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

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*(3), 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., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*, 1424–1435. https://doi.org/10.1111/ele.12189

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. *Lecture Notes in Computer Science*, *2888*, 986–996.

Hastie, T. J., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning data mining, inference, and prediction* (Vol. *1*). Springer. 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. *Ecology and Evolution*, *3*, 5225–5236. 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, School of Mathematics and Statistics, UNSW Sydney, Sydney. https://doi.org/10.1017/S0004972715000945

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.

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

Köhler, J., Vieites, D. R., Bonett, R. M., García, F. H., Glaw, F., Steinke, D., & Vences, M. (2005). New amphibians and global conservation: A boost in species discoveries in a highly endangered vertebrate group. *BioScience*, *55*, 693–696.10.1641/0006-3568(2005)055[0693:NAAGCA]2.0.CO;2.

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

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. 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

Oza, A. U., Lovett, K. E., Williams, S. E., & Moritz, C. (2012). Recent speciation and limited phylogeographic structure in *Mixophyes* frogs from the Australian wet tropics. *Molecular Phylogenetics and Evolution*, *62*, 407–413. https://doi.org/10.1016/j.ympev.2011.10.010

Padial, J. M., & De la Riva, I. (2006). Taxonomic inflation and the stability of species lists: The perils of ostrich's behavior. *Systematic Biology*, *55*, 859–867. https://doi.org/10.1080/1063515060081588

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., & Denoeux, 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 Development Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. 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://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., … 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*(1), 289–317. https://doi.org/10.32614/RJ-2016-021

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*, e8621. 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, Curtin University, Perth. 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., Tran, Q. N., & Vo, B.-T. (2018). Model-based learning for point pattern data. *Pattern Recognition*, *84*, 136–151. 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

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. *The Journal of the Acoustical Society of America*, *131*, EL184–EL190. https://doi.org/10.1121/1.3678685

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., & Steinberg, D. (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. https://doi.org/10.1007/s10916-011-9710-5

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Guilbault E, Renner I, Mahony M, Beh E. How to make use of unlabeled observations in species distribution modeling using point process models. *Ecol Evol*. 2021;11:5220–5243. https://doi.org/10.1002/ece3.7411

APPENDIX A

Testing algorithm parameters

For the parameters involved in each initialization method, we choose different values to test our model:

- knn method:$k \in \{1, 3, 5, 10\}$
- kmeans method:$nstart \in \{10, 15, 30, 50\}$
- LoopT method: $\delta_{max} \in \{0.5, 0.7, 0.9\}$, $\delta_{min} \in \{0.1, 0.3, 0.5, 0.7\}$, $\delta_{step} \in \{0.05, 0.1, 0.2\}$ where only one of the three parameters is varying at a time.
- LoopE method:$a \in \{1, 5, 10, 15, 20, 30, 40\}$

The results presented in Table A1 correspond to simulations with 80% of hidden observations because no major differences were found for 20% and 50% of hidden observations. The knn, kmeans, and random method did not show any differences when the parameters k and $nstart$, respectively, vary. All the other methods present the best performances for the parameters values displayed in Table A1.

The choices of δ_{max}, δ_{min}, and δ_{step} control the rate and breadth of points added to the set of locations with known species labels. As such, they control the growth of the predicted distribution as we reduce δ_h for each iteration h. Setting a higher value of δ_{max} such as 0.9 suggests that we first augment the distribution with only those points in which we are very confident in belonging to the species and therefore initially grow the predicted distribution slowly, while setting a lower value such as 0.5 suggests that we grow the distribution more rapidly to begin. These two values tended to be the best, with $\delta_{\text{max}} = 0.5$ being optimal for test 4 with equal abundance and high correlation between distributions. The lower the value of δ_{min}, the more we will grow the predicted distribution. The optimal value tends to be around 0.1, suggesting that growing the distributions significantly and therefore making use of most of the locations with unknown points is worthwhile. The value of δ_{step} controls the rate of growth between iterations. There does not seem to be a consistent winner among the three choices tested.

For the LoopE method, we increase at each iteration the number of points to add starting from a defined number a_1. While the a_h points with highest membership probabilities are added, these membership probabilities may be small for large values of a_h, and this could explain that this method is not always doing as well as other methods. With LoopT, each species distribution grows based on equitable confidence bounds, while with LoopE, each species distribution grows based on equitable numbers of points added. The risk of LoopT is in growing the most abundant species too quickly while the risk of LoopE is in growing the least abundant species too quickly. Additionally, because we add the same numbers of points for each species, having not highly correlated species distributions may result in adding points to the wrong species. On another note, it seems that the initial number of points a_1 did not influence the performances.

Testing species distribution for mixture methods

We present analogous results of Figures 3-6 for different initialization methods.

Estimated standard error

Here, we present the boxplots for the standard error of the predictions presented in the manuscript. The LoopT and LoopA methods tend to have lower standard error in particular at 80% of hidden observations. As expected, the individual PPM method exhibits the highest standard errors, particularly for 50% and 80% of hidden observations.

Framework in R

We show below the steps in R to use the function for the framework presented in the manuscript. This document as well as the function used are available on github at https://github.com/EmyGlblt/LoopMixArticle.

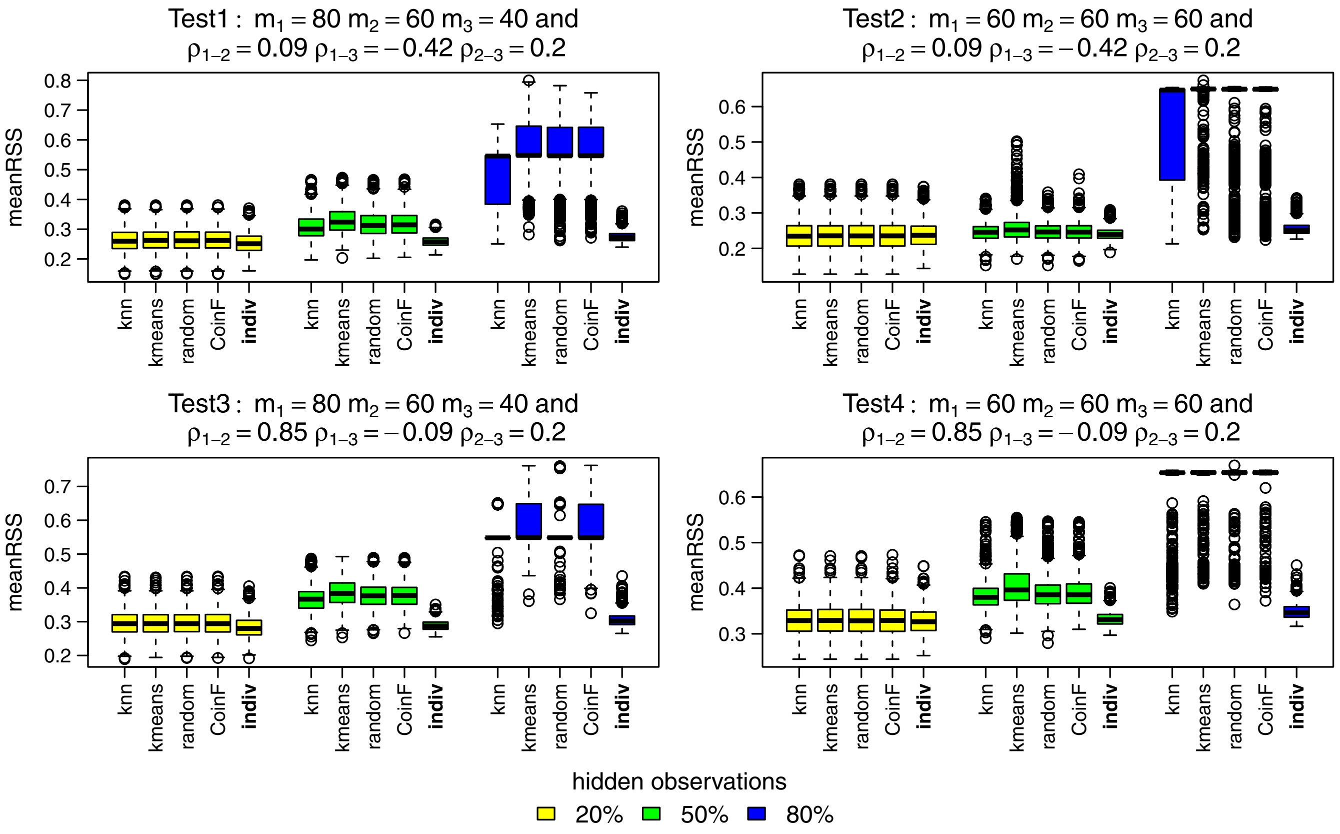


FIGURE A1 MeanRSS for the methods: knn, kmeans, random, coinF, and individual PPM (reference). Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50%, and in blue for 80%

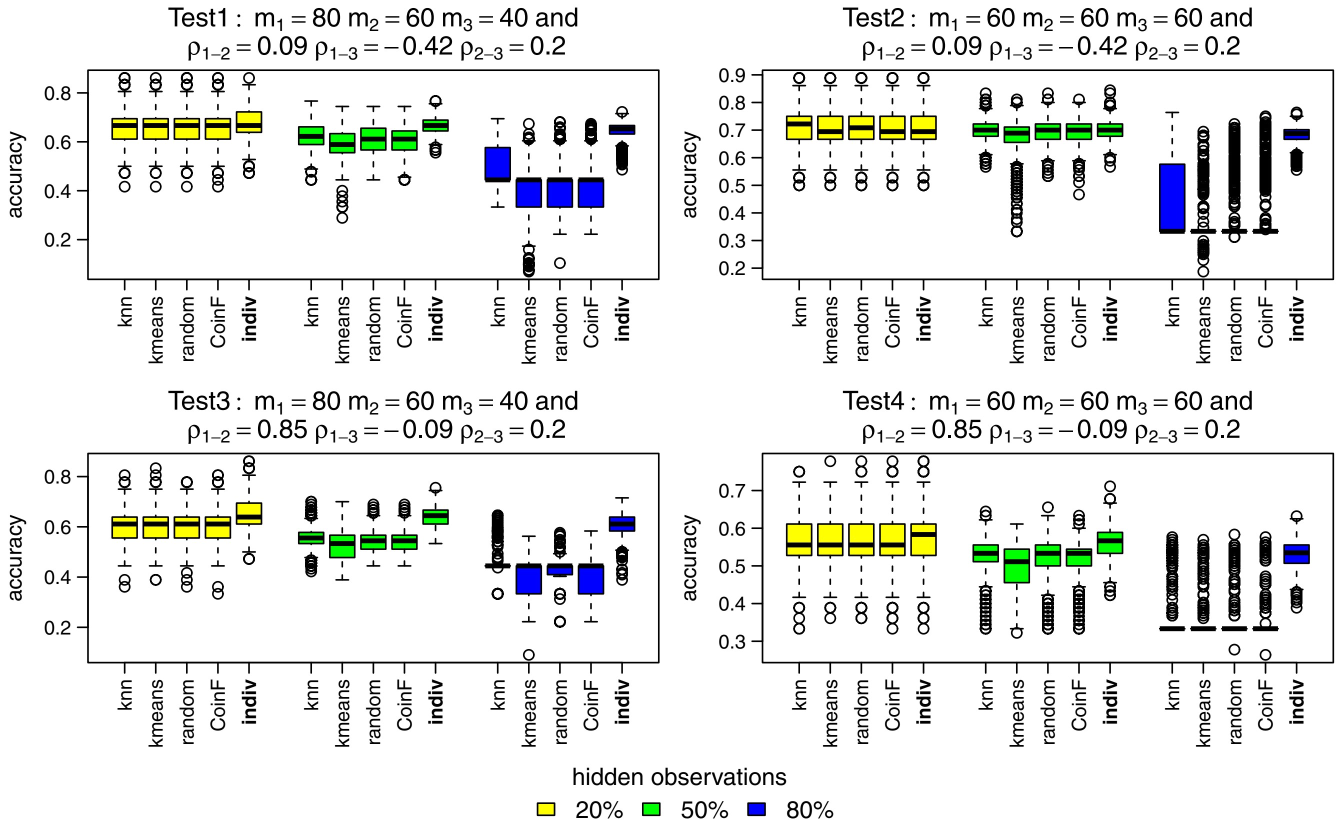


FIGURE A2 Accuracy for the methods: knn, kmeans, random, coinF, and individual PPM (reference). Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50% and in blue for 80%

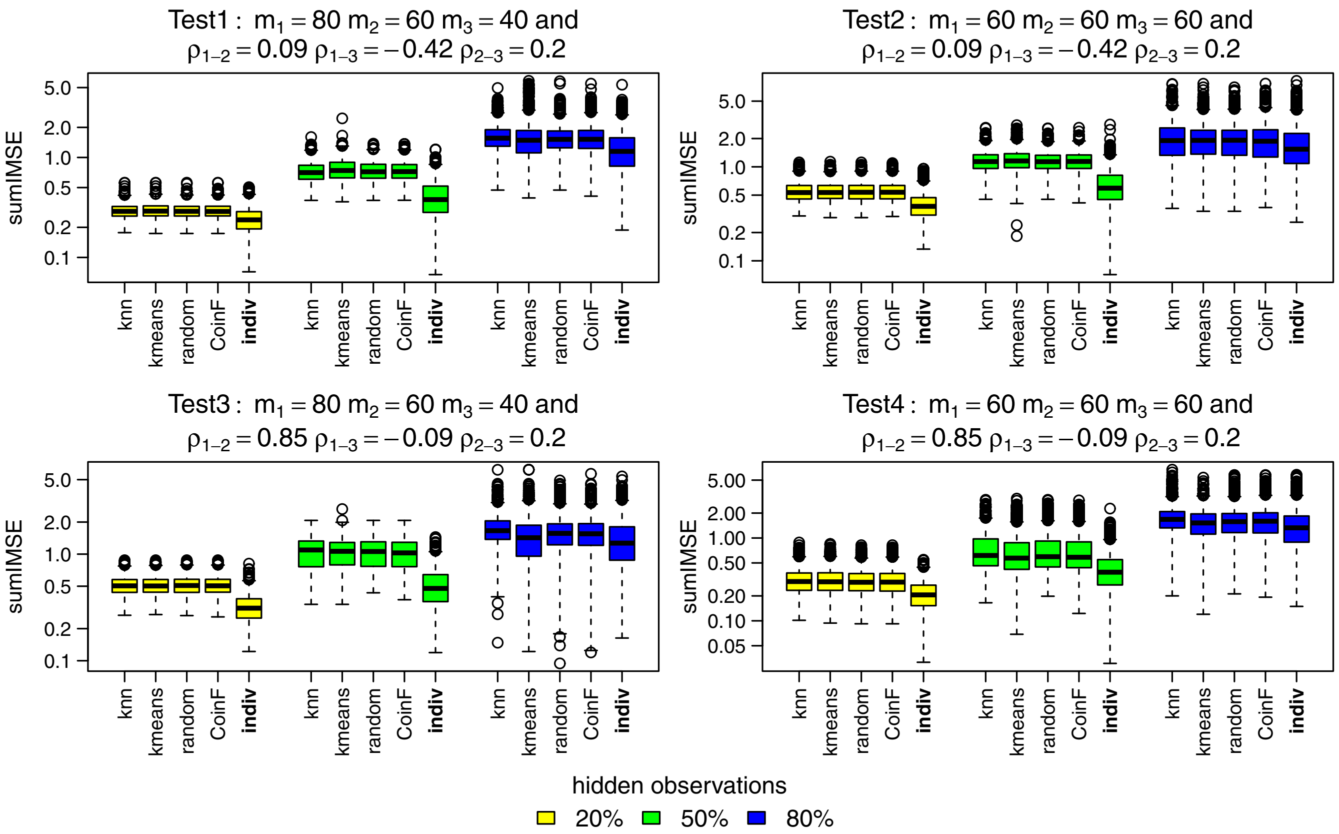


FIGURE A3 SumIMSE (logarithmic scale) for the methods: knn, kmeans, random, coinF, and individual PPM (reference). Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50%, and in blue for 80%

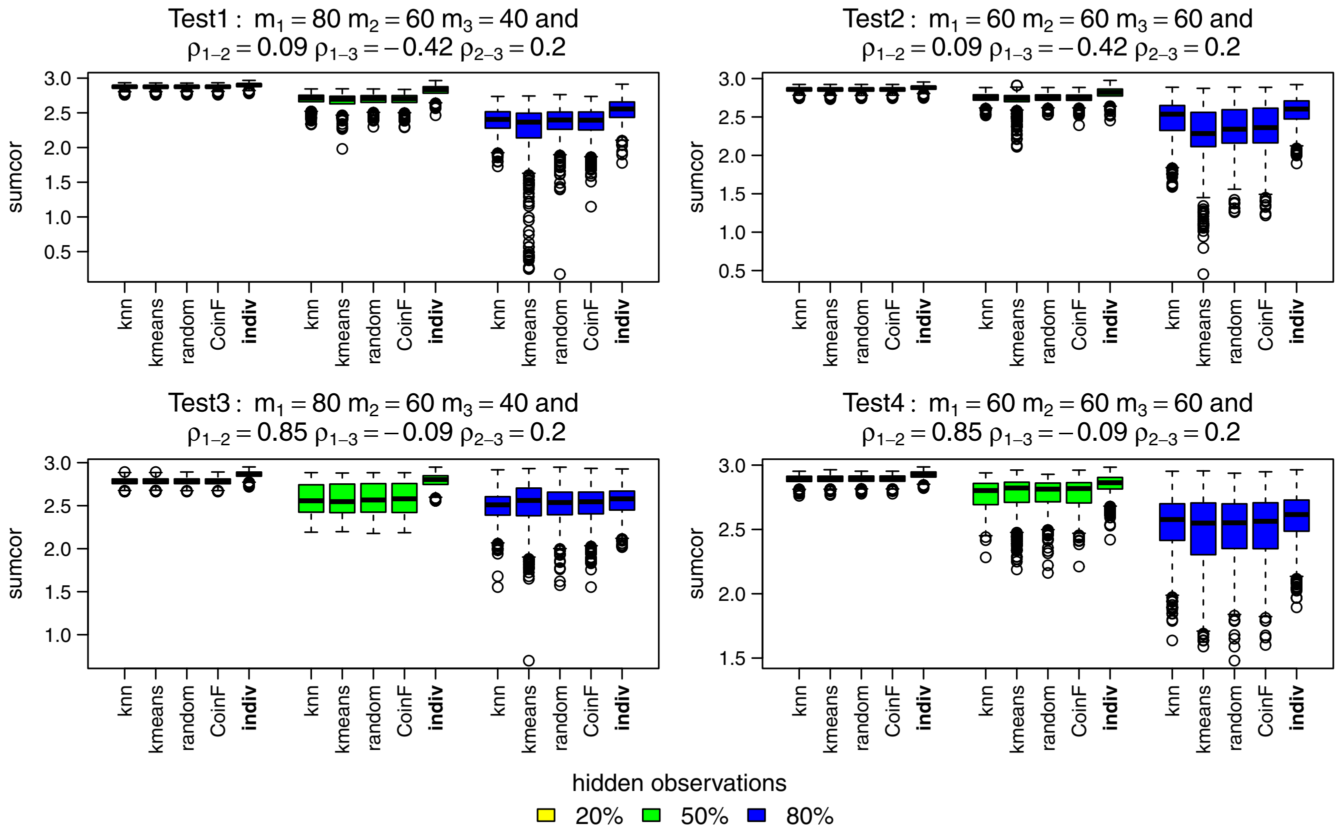


FIGURE A4 Sumcor for the methods: knn, kmeans, random, coinF, and individual PPM (reference). Each color boxplot represents a different percentage of hidden observations: in yellow are the performances for 20% of hidden observations, in green for 50%, and in blue for 80%

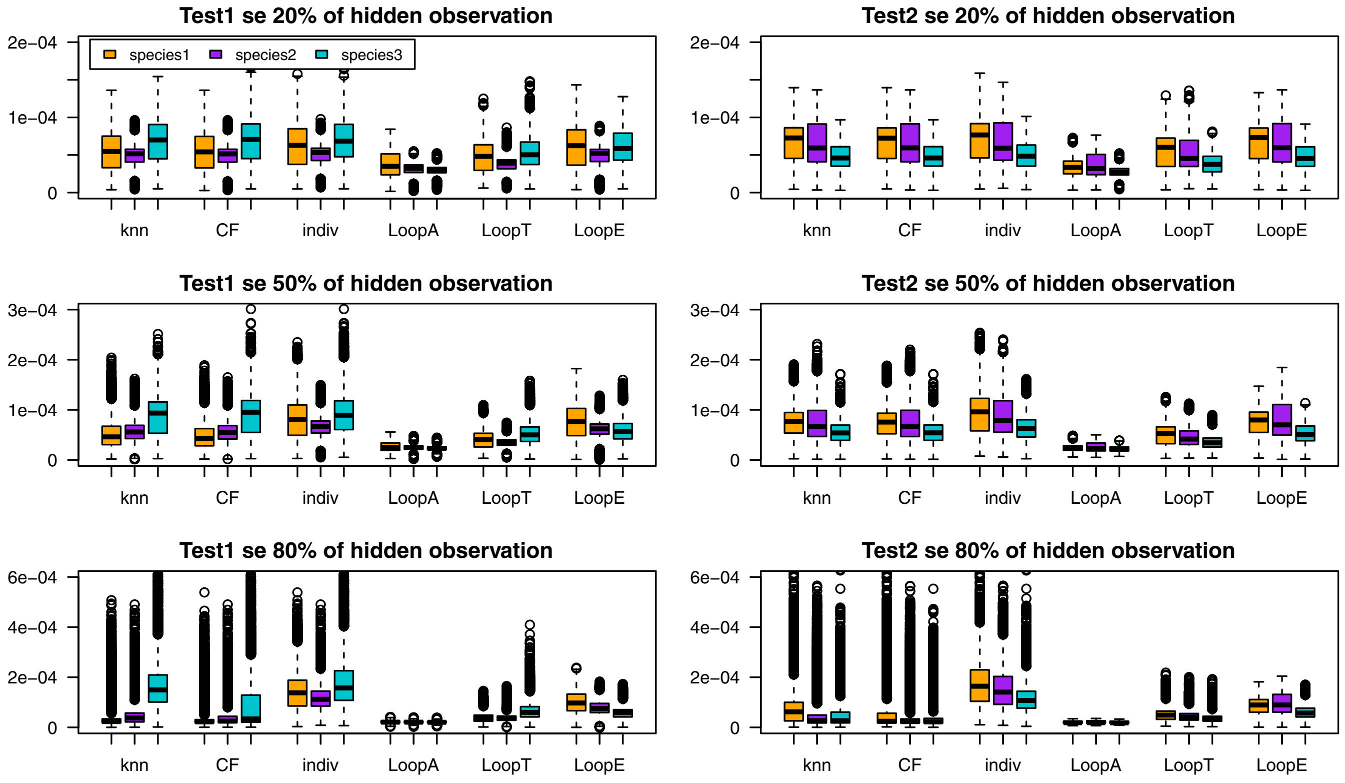


FIGURE A5 Standard error for the best methods: knn, coinF, individual PPM (reference), LoopA, LoopT, and LoopE. Each color boxplot represents a different species: in orange species 1, in purple species 2, and in turquoise species 3. The tests use the following parameters: test 1: $m_1 = 80$, $m_2 = 60$, $m_3 = 40$; $\rho_{1-2} = 0.09$, $\rho_{1-3} = -0.42$, $\rho_{2-3} = 0.20$; test 2: $m_1 = 60$, $m_2 = 60$, $m_3 = 60$; $\rho_{1-2} = 0.09$, $\rho_{1-3} = -0.42$, $\rho_{2-3} = 0.20$

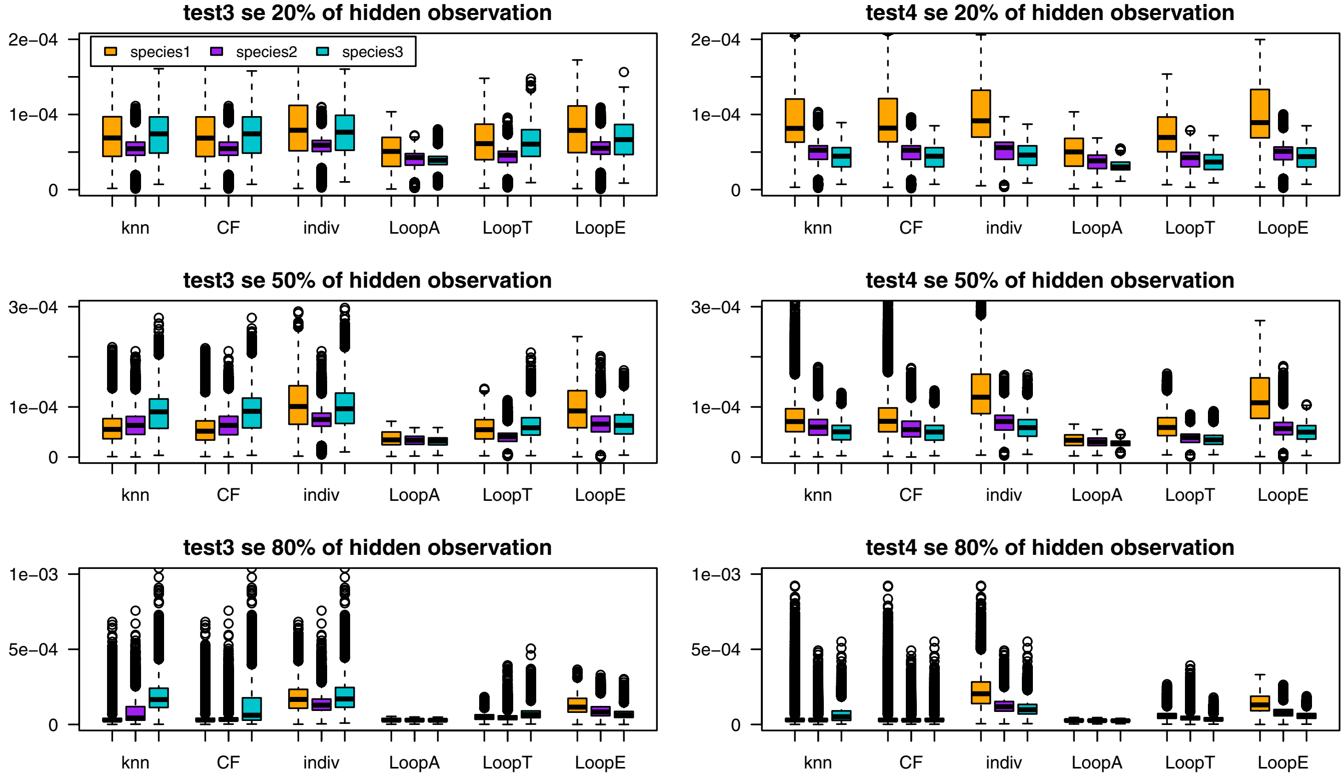


FIGURE A6 Standard error for the best methods: knn, individual PPM (reference), LoopA, LoopT, and LoopE. Each color boxplot represents a different species: in orange species 1, in purple species 2, and in turquoise species 3. The tests use the following parameters: test 3: $m_1 = 80$, $m_2 = 60$, $m_3 = 40$; $\rho_{1-2} = 0.85$, $\rho_{1-3} = -0.09$, $\rho_{2-3} = 0.20$; test 4: $m_1 = 60$, $m_2 = 60$, $m_3 = 60$; $\rho_{1-2} = 0.85$, $\rho_{1-3} = -0.09$, $\rho_{2-3} = 0.20$

TABLE A1 Summary table for parameter testing

		Test 1				Test 2				Test 3				Test 4			
Method	Parameter	MeanRSS	Acc	sumIMSE	sumcor	MeanRSS	Acc	sumIMSE	Sumcor	MeanRSS	Acc	sumIMSE	sumcor	MeanRSS	Acc	sumIMSE	sumcor
knn	k	X															
kmeans	nstart	x but only reach convergence for high starting values															
LoopT	delta max	0.5; 0.9		x		0.5; 0.9		x		0.5; 0.9		x	0.5; 0.9	0.5	0.5; 0.9	x	0.5
	delta min	0.1	x		0.1	0.1	x		0.1	0.1	x		0.1	0.1; 0.7	x		0.1
	delta step	0.1; 0.2		x		0.2		x	0.05	0.1; 0.2		x	0.05	0.1; 0.2		x	0.05, 0.1
LoopE	a	x															

Note: Tests 1, 2, 3, and 4 correspond to the test presented in the simulation part. The x sign means that there are no differences between the different values tested. The numbers displayed to represent the parameters for which the best performance is reached for each performance measure presented in the table.