INTRODUCTION

Citizen science is becoming extremely valuable in modern science (Bonney et al., 2014; Howe, 2008; Raykar et al., 2010). Large world-wide networks of participants are contributing to conservation efforts, generating substantial information and promoting a better understanding and awareness of biodiversity (McKinley et al., 2017). The numerous examples found in the literature range from the world-wide estimation of the abundance of birds (Sullivan et al., 2009), the distribution of jaguars in the Peruvian Amazon (Mengersen et al., 2017) and hard coral cover estimation in the Great Barrier Reef, Australia (Peterson et al., 2020). Many such projects
can now be found in online platforms such as iNaturalist, eButterfly and Zooniverse.

Notwithstanding its popularity and potential, the quality of the data produced by volunteers is often questioned in the scientific community (Bonney et al., 2014; Kosmala et al., 2018). Contributors’ commitment, abilities, training and effort, along with the difficulty of the task, affect their performance (Dennis et al., 2017; Kelling et al., 2015). In crowdsourcing projects, such as those involving manual image classification, there is a growing interest in assessing how well users can perform estimation and classification tasks, specifically in measuring their latent abilities (e.g. Falk et al., 2019).

In this paper, an assessment approach is proposed based on Item Response Theory (IRT) models for projects producing classification or observational data. These methods allow for the estimation of the relative performance of the participant, while accounting for the difficulty of the task. For each participant, the performance is associated with abilities, skills and knowledge. The participants’ abilities and the task’s difficulties are both latent since they are not observable, but inferred from the model. Measures of participants’ abilities, skills, knowledge and contribution is vital when working with citizen science data. Differentiating and clustering them into groups (e.g. beginners, advanced, experts, etc.) helps to produce more reliable data and guides scientists on how to improve proficiency, engagement and training. Statistical inference from crowdsourced data assigns greater weight to information from more competent users (Bird et al., 2014; Johnston et al., 2018; Kosmala et al., 2016; Santos-Fernández et al., 2021). In the absence of a gold standard, ability-derived weights can be used in consensus voting algorithms or for building reputation and leaderboard systems (Callaghan, Rowley, et al., 2019; Silvertow et al., 2015). Extremely low abilities are generally associated with careless, non-genuine respondents and bots, which are generally excluded for being non-informative and representing a burden for inference and computation.

There is an extensive literature on the use of IRT in a wide range of fields, ranging from psychology, political, social sciences and education (Gadermann et al., 2012; Laurens et al., 2012; McGann, 2014) to statistics and computer sciences (De Jong et al., 2010; van der Linden et al., 2010; Meyer & Zhu, 2013; Wauters et al., 2010). For binary classification tasks, of interest in this paper, a common approach found in the literature is the logistic model. Latent variable models that have been used in ecology to describe the occurrence of species as a function of latent predictors (Pollock et al., 2014; Warton et al., 2015) are related to the IRT models. The main distinction is that IRT models use different formulations as we show later.

Multiple frequentist and Bayesian approaches to IRT have been employed, along with a wide range of computational methods. Frequentist implementations of these models are available in a wide range of statistical software packages; see, for example, Paek and Cole (2019) for an implementation in R. In the case of the Bayesian approach, several software implementations have been published using different statistical languages. For instance, Curtis et al. (2010) used BUGS/JAGS; Grant et al. (2016), Luo and Jiao (2018) and Bürkner (2019) used Stan, while Stone and Zhu (2015) employed SAS. Multiple examples based on Bayesian inference can be found in Fox (2010) and in the Stan user manual (Carpenter et al., 2017).

Overall, little attention has yet been paid to estimating the abilities and performance measures for citizen science in ecological settings. In this research, subjects are referred to as participants, citizens, annotators, users, while items refer to information sources (images, videos, audio files, etc.) (Gura, 2013; Ratnieks et al., 2016; Zhang et al., 2013). However, we concentrate on the particular case of species identification using manual image classification.

Data collected in citizen science projects are challenging compared to that found in the traditional item response literature. For example, the number of items for classifications can be large, often reaching hundreds of thousands, and in most cases, it could be larger than the number of users. Additionally, participants classify a different number of items. Some users engage heavily while others contribute lightly (Hsing et al., 2018). However, a user rarely classifies all the items if the number is large. Complete panel data, which is the result of all the participants classifying all the items, are extremely uncommon in citizen science. Instead, a large proportion of the data is missing. Overall, the skills of the respondents are generally heterogeneous, ranging from experts to beginners.

These complications are exacerbated when the focus of citizen science is on the classification of features or on the identification of species in images. In this set-up, camera, site- and species-specific factors can affect the underlying item difficulty level (Willi et al., 2019). Several image specific components may be also critical, for example the camera resolution, brand, flash effect, etc. Other crucial factors are the visibility, time of the day (day or night), landscape and the vegetation. It has also been well-documented that some species are far more difficult to identify than others. Look-alike or closely related species are more challenging, leading frequently to misclassifications (Chambert et al., 2018; Hsing et al., 2018).

In ecological research, items such as images are usually georeferenced (Callaghan, Rowley, et al., 2019; McKinley et al., 2017). An added complication in such a setting is that items that are geographically close to each other tend to have more similar characteristics, such as the probability of containing the target species or the latent item difficulty, compared to those that are further away. For this reason, many citizen science applications account for spatial autocorrelation. Some examples can be found in Pagel et al. (2014); Purse et al. (2015); Arab and Courter (2015); Humphreys et al. (2019); and Altwegg and Nichols (2019). The image difficulties tend to exhibit spatial autocorrelation, which should be accounted for in the IRT models. Ignoring such autocorrelation can lead to biased parameter estimates and underestimate the associated standard errors, and hence, potentially erroneous inference. See, for example, Lichstein et al. (2002); Ver Hoef et al. (2018).

The identification of areas with a strong spatial association is also relevant for survey administration purposes. Distinguishing areas in which the items have relatively high difficulties is vital to provide recommendations on how to improve the quality of the images, generate more useful training and users’ documentation and to better manage the order in which tasks are presented for classification, etc.
High levels of misclassification in images produced by camera traps might also indicate technical issues with the cameras. Additionally, the identification of images with no visible species of interest and covered by vegetation is relevant for data cleaning purposes.

Traditional item response models found in the literature do not consider spatial dependence in the items, which can cause biased parameter estimates when applied to spatial data. The purpose of the article was to enhance existing models by accounting for spatial autocorrelation in the parameters associated with the item difficulties. Specifically, we introduce extensions of the 3PI item response model (U)sing (S)patially dependent item difficulties (henceforth 3PLUS). These approaches are tailored to citizen science data produced from the classification of georeferenced images. We discuss the advantages and limitations of the suggested methods and assess their performance. We also illustrate how to use these new approaches with big datasets, frequently produced in citizen science projects. We take a Bayesian approach for modelling using prior information for the parameters of interest. This has multiple advantages which include better parameter estimates for cases in which the number of samples for some users are relatively small (Choy et al., 2009; Mislevy, 1986). We start with a simulation study, to compare the proposed model to the state of the art. This is followed by a case study of citizen science image classification in the Serengeti, Africa.

2 | MATERIALS AND METHODS

2.1 | Spatial item response models

In this section we lay out the theoretical aspects of the 3Pl model and propose several model extensions. In the case of citizen-elicited image classification, a common approach is to ask the user whether it contains the target class or species. Let the binary response variable Y represent whether a question is correctly answered by the ith citizen (i = 1, ..., I) in the jth image (j = 1, ..., J), and assume that Yj has a Bernoulli distribution, Yj ∼ Bern(pj). In some studies, a citizen i may answer only a subset of the total number of items J, so there may be missingness in Y for several combinations of j. The probability pj of a correct answer can be modelled using the three-parameter logistic model (Baker & Kim, 2004; Lord, 1980).

\[
p_j = \eta_j + (1 - \eta_j) \frac{1}{1 + \exp(-a_j(\theta_i - b_j))}
\]  

Here θi is the latent ability of the ith user, such that a larger value of θ indicates a larger probability of identifying correctly the class of the item. It is also possible to express the ability of the participant based on two factors: (1) the ability to correctly identify the species when it is truly present in the image, which is known as sensitivity, and (2) the ability to classify the image as absent when the species is not present (specificity). This model is described in the spatial context in Santos-Fernández et al. (2021).

The parameter bj represents the difficulty of the jth image and \( a_j > 0 \) represents the slope or discrimination parameter of image j indicating how quickly the function will go from 0 to 1. The parameter \( \eta_j \) is the pseudoguessing parameter of the jth image that is the lower asymptote, which accounts for users’ chance of answering correctly by guessing. The reader is referred to the Supporting Information for a brief illustration of a simple item response model.

The two-parameter logistic (2Pl) model is a reduction arising from setting \( \eta_j = 0 \), while the Rasch model is obtained using \( \eta_j = 0 \) and \( a_j = 1 \) (Cai et al., 2016). The effect of the difficulty, slope and pseudoguessing parameter on \( p_{ij} \) under the 3Pl model is shown in Figure 1.

In the context of a study involving tasks such as classification of geographically located images, it is important to take into account the inherent spatial correlation between the items. To address

![Figure 1](image-url)  

**FIGURE 1** Item characteristic curves of the 3Pl model giving the probability of correctly answering a standard item (b = 0) as a function of the latent user ability θ. In (a), we compare curves from three different values of the slope α with no guessing, that is η = 0. In (b), we compare curves using three pseudoguessing values η = 0, 0.1, 0.2 and a fixed value of α = 1.
this, we introduce a spatial extension of the 3PI model in which we model the difficulties of the task as a spatial random effect. Two implementations are proposed. In the first case, we use a Gaussian process with covariance structure \( \Sigma \) (Banerjee et al., 2014). In the second variation, we adopt a conditional autoregressive (CAR) prior (Besag et al., 1991). We note that the general approach can accommodate other representations of the spatial autocorrelation. CAR priors and other spatial methods have been found to adequately capture variability and spatial autocorrelation in citizen science applications (Altwegg & Nichols, 2019; Arab & Courter, 2015; Arab et al., 2016; Pagel et al., 2014; Purse et al., 2015; Santos-Fernández et al., 2021). Multiple alternatives have been used in the literature, such as Gaussian random fields (Humphreys et al., 2019) and stochastic partial differential equations (SPDE) (Peterson et al., 2020). Under the CAR prior alternative, the distribution of the difficulties \( b_j \) conditional on the first-order neighbours is the average of the \( n_q \) first-order neighbours plus Gaussian noise. The variance is inversely proportional to the number of neighbours.

\[
b_q | \nu_m, \tau_b \sim \mathcal{N} \left( \frac{1}{n_q} \sum_{q=m}^{m} \nu_m \frac{1}{\tau_b} \right),
\]

where \( n_q \) is the number of neighbours for region \( q \), \( q \sim m \) means \( q \) and \( m \) are neighbours and \( q \neq m \). The parameter \( \tau_b \) is the precision of the spatial effect that determines the variance along with \( n_q \). In the CAR model, the covariance matrix is expressed as a function of the adjacency matrix \( W \) which contains the neighbouring structure \( \Sigma_b = (I - \rho W)^{-1} \), where \( 0 < \rho < 1 \) determines the amount of spatial dependence and \( I \) is an \( J \times J \) identity matrix. Usually a sum to zero constraint is used to improve parameter identifiability and model interpretability (e.g. \( \Sigma b_q = 0 \)).

In the Gaussian process approach, we model the difficulty using a multivariate normal distribution \( b_j \sim \text{MVN}(\mu_b, \Sigma_b) \). Here we define \( \mu_b = X \beta \), where \( X \) represents a design matrix of covariates affecting the difficulty and \( \beta \) is the vector of regression coefficients. The covariance matrix \( \Sigma_b \) of dimension \( J \times J \) is written as \( \Sigma_b = C + \sigma_0^2 I \), where \( C \) is a covariance matrix that describes the spatial dependencies between locations, \( \sigma_0^2 \) is the nugget effect and \( I \) is the identity matrix. Several spatial structures based on the Euclidean distance have been proposed for \( C \). Common modeling options have exponential, Gaussian or spherical form (Banerjee et al., 2014; Cressie, 2015):

- **Exponential model**: \( C = \sigma^2 \exp(-3d/a) \),
- **Gaussian model**: \( C = \sigma^2 \exp(-3d/a)^2 \),
- **Spherical model**: \( C = \sigma^2 \left( 1 - \frac{3d}{2a} + \frac{d^3}{2a^3} \right) I(\text{d}/a \leq 1) \),

where \( \sigma^2 \) is the partial sill and \( a \) is the spatial range parameter. The spatial range indicates the extent of the spatial dependence. Therefore, there will exist a negligible spatial autocorrelation between two points, when the distance between them is larger than the spatial range.

When a large number of images are taken from the same spatial location, as in the case of camera traps, the number of parameters associated with the spatial process can be prohibitive. A reasonable assumption to make is that images taken from the same site are realizations from the same site-specific difficulty \( b \). For this case, the index \( j \) will represent a unique spatial location rather than an image id.

Possibly, the greatest issue when using item response models is caused by the model being non-identifiable. This leads to computational challenges, such as the failure of the algorithm to converge. These issues occur when the model has more parameters than can be estimated from the data. Similarly, identifiability problems result when several combinations of parameters produce the same response variable. For a detailed discussion on identifiability, see, for example, Fox (2010). Strategies like anchoring the sum or the mean of the difficulties or the abilities to zero have been suggested to cope with this limitation.

### 2.2 Item explanatory LLTM model

We propose a further extension of the 3PLUS model called item explanatory, which is a special case of the linear logistic test model (LLTM) (Fischer, 1973; Wilson et al., 2008). In this approach, the item parameter \( b_j \) is a linear combination of factors governing the difficulty \( \sum_{m=1}^{M} \beta_j \mu_{jm} \), where \( M \) is the number of factors. This model does not give directly an indication of the difficulties associated with the images, but explains the effect of the item-related factors. For instance, often we are interested in the camera difficulty that accounts for the image quality as well as the complications associated with the site (unique latitude and longitude) in which the camera is placed rather than on the individual image difficulties. Similarly, the intrinsic difficulty associated with the species can be considered. This explains the misclassification error due to species that look like the target one and also mimic animal-specific behaviour and features.

Let \( Y_{ij} \sim \text{Bern}(p_{ij}) \) indicate whether the citizen \( i \) correctly classified an image from the site \( j \) containing species \( l \), where the probability \( p_{ij} \) is given by:

\[
p_{ij} = \eta + (1 - \eta) \frac{1}{1 + \exp[-\alpha (\theta_i - \beta_j - \beta_l)]},
\]

where the parameters \( \beta_j \) and \( \beta_l \) are difficulties associated with the site and species respectively. As above, \( \beta_l \) is designed to capture spatial dependence and will be modelled using the spatial structures previously discussed. The pseudoguessing parameter \( \eta \) can be associated with the species or with the sites, hence the dot index. The first case is more reasonable since it explains which species are more likely to be correctly classified due to guessing. Large of species-related pseudoguessing values indicate species that are easily classified for those users with less proficiency. Similarly, species or site-specific discrimination parameters can be estimated in the model \( \{u_i, v_j\} \), showing which items are more suitable for differentiating between the users.
To avoid confusion, we will retain the notation previously used to refer to the models but we add a prime symbol (′) to denote test model extensions. That is, the test model (3PLUS′) of Equation 6, for example, is an extension of the model introduced in Equation 1 with an extra parameter $\beta_i$.

3 | SIMULATION STUDY

We first present a simulation study to compare the performance of the 3Pl and 3PLUS models using the model from Equation 1. Consider $J = 225$ unique geographical locations in a $15 \times 15$ grid, which is generally large enough for the illustration and has been employed in several practical designs (e.g. Swanson et al., 2015).

For the image difficulties, we generated the spatial autocorrelation using a multivariate normal distribution, $b_j \sim \text{MVN}(\mu_b = 0, \Sigma_b)$. The spatial covariance matrix $\Sigma_b$ was obtained using the simple exponential form, for example $\Sigma_b = \sigma^2 e^{-d/\alpha}$, where $\sigma$ is the partial sill and we set $\alpha = 1$. The parameter $d$ is the Euclidean distance between locations and $\alpha$ is the spatial range. We partitioned the spatial process of the difficulties into regions using Voronoi polygons (Gold, 2016; Okabe et al., 2009). The image locations were the centroids for the polygons and the boundaries were obtained using the Euclidean distance. Figure A4a shows the spatial distribution of the latent difficulty.

In (a) we show the spatial association between the difficulties and regions characterized by clusters of images that are easy (light blue) and hard (dark blue) to classify. Consider five groups of users with fixed abilities: $\theta_i = (-1, -0.5, 0, 0.5, 1)$. We set several values for the discrimination ($a$) ranging from small to large slopes, $a_j = \{0.25, 0.50, ..., 1.50, 1.75\}$. We assume having six target classes, each with the same probability of being correctly classified by guessing and consider that users will have on average one sixth chance of guessing the true target class with extremely low knowledge. Therefore, a pseudoguessing parameter $\eta_j \sim \text{Beta}(1, 5)$ is employed, which is a weakly informative prior commonly used in the literature for this parameter.

A random user and a slope were assigned with replacement to each image id. On each image a random species was generated and a pseudoguessing value was then assigned. Each citizen classifies multiple images and each image is classified several times. Using $p$ from Equation 1 we simulated binary realizations using the Bernoulli distribution.

3.1 | Fitting the Bayesian models

We want to learn about the latent parameters $\theta_i$, $b_j$, $a_j$ and $\eta_j$ using the observed binary variable $Y_{ij}$ that denotes whether the citizen $i$ classified correctly the $j$th image. We simulated spatial dependence using an exponential model. However, keeping in mind that in practice, the true spatial data generating process is unknown, we account for spatial autocorrelation in the model by adopting a more general CAR prior.

We employed Markov chain Monte Carlo simulations, in particular Hamiltonian Monte Carlo (HMC), in the software package Stan (Carpenter et al., 2017) which is based on the no-U-turn sampler (NUTS) (Hoffman & Gelman, 2014). We used three chains each of 50,000 samples after discarding a burn-in of 20,000 samples. We used the prior distributions shown in Table 1. The CAR prior was implemented according to Morris et al. (2019). We also performed a sensitivity analysis using several priors for the precision parameter but did not find substantial changes in the parameter estimates (results not shown).

The parameter $b_{3PI}$ is the difficulty in the traditional model (3PI), while $b_{3PLUS}$ represents the difficulty in the spatial approach. We anchored the ability of the ‘reference’ user by setting it equal to zero. This was done by fitting the model in the MIRT package (Chalmers, 2012) and finding the user with the score closest to zero.

The comparison of the models was based on the following criteria: (a) confusion matrix and the accuracy when estimating the parameters (difficulties, abilities, slopes, pseudoguessing), (b) Root Mean Square Error (RMSE) and (c) Watanabe–Akaike information criterion (WAIC) (Watanabe, 2010).

| TABLE 1 | Prior distributions used in the models |
|---|---|
| $\theta_i \sim \mathcal{N}(0, \sigma_{\theta})$ | # hierarchical informative prior on the abilities |
| $\sigma_{\theta} \sim \text{Cauchy}(0, 5)T(0,)$ | # half Cauchy prior on the user’s abilities sd. $T(0,)$ means lower truncation at 0 |
| $b_{3PLUS} \sim \mathcal{N}(\mu_b, \sigma_b)$ | # informative prior on the difficulties in 3PI model |
| $\mu_b \sim \mathcal{N}(0, 5)$ | # normal prior for the mean of the item difficulty |
| $\sigma_b \sim \text{Cauchy}(0, 5)T(0,)$ | # flat prior for the sd of the item difficulty |
| $b_{3PLUS} \sim \text{CAR}(\tau, W, D)$ | # CAR prior for the spatial model |
| $\tau_j \sim \text{Gamma}(1, 1)$ | # precision of the spatial effect |
| $a_j \sim \mathcal{N}(1, \sigma_a)$ | # normal prior with mean 1 on the slope |
| $\sigma_a \sim \text{Cauchy}(0, 5)T(0,)$ | # half Cauchy prior for the slope sd |
| $\eta_j \sim \text{beta}(1, 5)$ | # weakly informative prior on the pseudoguessing |
| $\tau$ | # precision parameter in the CAR prior |
| $D_{\text{diag}}$ | # diagonal matrix |
| $W_{\text{diag}}$ | # adjacency matrix |

3.2 | Simulation results

The two models produced similar user ability estimates (See Figure A1, Supporting Information). The abilities’ RMSE for both model was similar (0.2451 in the 3PLUS vs. 0.2452 in the 3Pl). The 3PLUS model produces better difficulty estimates (Figure 2a; Figure A4) with respect to accuracy and precision. Figure A3 (a) shows the latent difficulties and figures (b) and (c) depict the estimated values from the 3PI and
3PLUS models respectively. The 3PLUS model recovers the latent difficulty much better. Figure A4b,c contains the Voronoi plots for both models. The yellow labels are the locations where the true difficulty categories were not correctly estimated and were therefore misclassified. The 3PLUS model achieved a prediction accuracy of 76.89% for the difficulty classes compared to 62.22% under the 3Pl model. This model also reduced the RMSE substantially: RMSE_{3PLUS} = 0.300 versus RMSE_{3PL} = 0.499. The correlations between the estimated difficulties and the true values were $r^2_{3PLUS} = 0.957$ versus $r^2_{3PL} = 0.868$. The two models produced similar estimates of the slope ($\alpha$), but the pseudoguessing parameter ($\eta$) was better estimated in the 3PLUS model in terms of RMSE. The estimates showed shrinkage towards the mean, where the amount of shrinkage was determined by the strength priors and hyperparameters used (see Figure A2a,b). The 3PLUS model also produced a smaller WAIC (23,136.2 vs. 23,187.6) and LOO (leave-one-out cross-validation) values (23,139.5 vs. 23,191.3). We also performed a similar simulation study using the linear logistic test models 3Pl$'$ and the 3PLUS$'$ considering species-specific difficulties. We found the 3PLUS$'$ model to be superior in terms of the precision of the difficulties and WAIC (results not shown).

4 | HAKUNA MY DATA: A CASE STUDY OF SPECIES IDENTIFICATION IN THE SERENGETI

4.1 | The Serengeti data

Swanson et al. (2015) describe a citizen science project that identified species from the Serengeti, Tanzania. This project is hosted on Zooniverse (https://www.zooniverse.org/projects/zooniverse/snapshot-serengeti) and captured more than a million images, with a total of 10.8 million classifications produced by 28,000 users. Images were obtained by camera traps in 225 locations. The dataset contains image details of the classifications per user and includes the spatial component given by the location of the cameras. Swanson et al. (2015) published several datasets including the original raw classification data, the consensus voting results and a gold standard dataset. The gold standard dataset included classifications by experts of 4,140 images.

A total of 50 species categories were identified including wildebeest, zebra, buffalo, hartebeest, lion (female and male), etc. The categories impossible to identify and human were also included. The authors noted that subjects tend to use more the ‘nothing here’ classification rather than guessing when they were not sure of the species. Two-dimensional kernel density plots of the species abundance obtained from the images are presented in Figure A6. It shows the three most prevalent species (prey) plus female lion (predator). We assessed the spatial autocorrelation of the proportion of correct classification per site using the Moran’s $I$ test and found evidence of spatial dependence ($p < 0.001$).

Images were captured at different times of the day, including night time. Several other factors affecting the difficulty of the images are included: animal moving or feeding, the presence of babies, the relative position regarding the vegetation and camera placement, etc.

We selected registered users with more than five classifications so that we could obtain suitable estimates of the abilities. For inclusion, images had to have a known location or site id. This produced...
a dataset composed of almost 5 million observations, from 21,347 citizens from 225 locations. The data and codes associated with the case study can be found in the repository https://github.com/EdgarSantos-Fernandez/hakuna.

4.2 | Analysis using the gold standard data and model comparison

Assessing the fit of several model variations and comparisons becomes difficult with large datasets, since they require substantial computational resources and more sophisticated statistical algorithms. This is discussed further in Section 4.3. In the light of this, we adopt the following approach to assessing model fit. We first find the best performing model using the gold standard dataset and then fit it using the whole data. For simplicity, images with more than one species were excluded from the gold standard dataset resulting in 88,517 observations. Exclusion could potentially bias the estimates when the number of images with more than one species present is large, which was not the case here.

We studied 10 model variations of Equation (6) and captured spatial autocorrelation using an exponential covariance matrix (Equation 3). The models are compared in Table 2. We set the mean ability equal to zero to ensure identifiability and improve model interpretability. We also considered spatial priors for the guessing and slopes. However, this modelling option is computationally more expensive, requires a greater number of parameters to be estimated from the data and can be prohibitive for big data problems. We started with the 3Pl and 3PLUS previously compared in the simulations. Eight linear logistic test models were studied, from a very simple (1Pl′) to more complex (3PLUS′). Three variations of the 3PLUS′ were examined, resulting from permutations of the site and species slopes and pseudoguessing parameters. For instance, 3PLUS-2′ and 3PLUS-3′ involve a species-based guessing (αj), while the slopes are site specific for 3PLUS-2′ (αj) and species dependent for 3PLUS-3′ (αl). Spatial autocorrelation was modelled using an exponential covariance structure. The 10 competing models were fit in Stan using three chains, 4,000 iterations and a burn-in of 2,000 reaching convergence. Leave-one-out cross-validation (LOO) was used to evaluate model fit (Vehtari et al., 2017). The LOO measure is also appropriate since we are interested in the applicability to other users/items (see Table 2).

Based on the LOO information criterion values, the 3PLUS-2′ model was found to be superior. In this model, the pseudoguessing parameter is species related, while the slope is indexed to the site that accounts for the camera and location-specific factors.

4.3 | The effect of the covariates on the difficulties

Using the best competing model, we assessed how the covariates affect the difficulties of the task. Six binary covariates were included, indicating whether the animals were standing, resting, moving, eating or interacting and the presence of babies. The summary statistics of the regression coefficients are presented in Table 3 and the corresponding posterior distributions are displayed in Figure 3. These results indicate that juvenile animals were substantially harder to identify correctly. Moreover, images of juveniles tended to be approximately 11% more difficult after controlling for the other covariates and model parameters. Images with moving animals were

### Table 2: Results from the gold standard dataset according to several model variations from the model in Equation (6)

| Model     | Name                | Parameters                          | WAIC      | LOO       |
|-----------|---------------------|-------------------------------------|-----------|-----------|
| 3Pl       | traditional 3Pl model | α; β; βj = 0                         | 94,459.3  | 94,500.1  |
| 3PLUS     | spatial 3Pl model   | α; β; βj = 0; βj: MVN (μj, Σj)       | 86,153.3  | 86,205.8  |
| 1Pl'      | traditional 1Pl test model | α = 0; β; βj = 0; βj                 | 86,175.0  | 86,197.5  |
| 2Pl'      | traditional 2Pl test model | α = 0; β; βj = 0; βj                 | 85,400.0  | 85,444.0  |
| 3Pl'      | traditional 3Pl test model | α; β; βj = 0                         | 85,443.8  | 85,491.8  |
| 1PLUS'    | spatial 1PLUS test model | α = 0; β; βj: MVN (μj, Σj)          | 86,166.1  | 86,187.1  |
| 2PLUS'    | spatial 2PLUS test model | α = 0; β; βj: MVN (μj, Σj)          | 85,391.7  | 85,434.8  |
| 3PLUS-1'  | spatial 3PLUS test model | α; β; βj: MVN (μj, Σj)              | 85,460.3  | 85,508.1  |
| 3PLUS-2'  | spatial 3PLUS test model | α; β; βj: MVN (μj, Σj)              | 85,160.5  | 85,210.8  |
| 3PLUS-3'  | spatial 3PLUS test model | α; β; βj: MVN (μj, Σj)              | 85,678.4  | 85,723.3  |

Bold values indicate to highlight the best model.
substantially easier to classify for the participants. There was little evidence that the other covariates were associated with the difficulty of the images.

4.4 | Item response modelling for big data

Fitting item response models to big datasets can be challenging and potentially prohibitive within the Bayesian paradigm. Operations involving a large number of parameters often exceed the memory, processing and disc capacities, which is exacerbated if simulation-based computation such as MCMC is used. This is even more problematic when accounting for spatial autocorrelation (Datta et al., 2016; Finley et al., 2017; Katzfuss & Cressie, 2012).

In this section, we focus on modelling datasets too large to be fit directly in one machine or even on a modern HPC node. We use the dataset previously described, which is composed of almost 5 million classifications. A true answer (species) was obtained in the non-gold standard images using the majority vote (Swanson et al., 2015), which is a common practice (e.g. Hines et al., 2015; Wauthier & Jordan, 2011). In the gold standard images, the accuracy of the majority vote method was high (0.974), which indicates that it is suitable to estimate the true categories in the non-gold standard images.

We took a divide-and-conquer approach, also known as divide-and-recombine (Neiswanger et al., 2013; Scott et al., 2016; Wang et al., 2016). In this approach, the dataset is split into more manageable subsets or shards using stratified random sampling. The number of observations of each location and species is large. However, the classifications per user varied from a few to several hundred. Thus, we employed stratified sampling with users as grouping criteria, thereby splitting the dataset into 10 relatively similar parts. This yielded 10 shards with approximately half a million observations per shard. Some users will not be represented on all the subsets, while most of the species and sites were estimated on every shard.

Models were fit to the independent subsets on independent machines with no communication, and subposterior estimates were then combined to obtain global estimates from the data using consensus Monte Carlo (Scott et al., 2016). Specifically, we used weighted averages of the posterior MCMC chains, with weights inversely proportional to the variance of the posterior samples. At each iteration, the consensus estimate of a parameter is given by

$$\bar{\theta}_s = \frac{\sum_g W_s \theta_{sg}}{\sum_g W_s}$$

where $s$ is the subset or machine, $g$ is the iteration number and $\theta_{sg}$ generically denotes a parameter in shard $s$ and iteration $g$. Here, $W_s = \text{Var}(\theta_s | y_s)^{-1}$. The consensus Monte Carlo is known to produce an asymptotic approximation to the posterior based on the whole dataset (Scott et al., 2016).

All the results from this section can be reproduced using the R Markdown file in the repository https://github.com/EdgarSantos-Fernandez/hakuna, which illustrates (a) how to request resources on an HPC, (b) how to fit the models in Stan and (c) how to combine the posterior estimates.

Shards were fitted in parallel on a high-performance computing (HPC) system using 5 CPUs, three processors and requested 280 Gb of memory. The estimates of the latent parameters were obtained using the best performing model (3PLUS-2') found in the previous section. We used three chains, a warm-up period of 3,000 samples.
and 8,000 iterations. It took approximately 24 hr to fit each model and to compute the information criteria. To combine the subposteriors into the whole dataset posterior, we modified the function `consensusMCindep()` from the R package `parallelMCMCCombine` (Miroshnikov & Conlon, 2014). This approach is suitable for normally distributed parameters (difficulties, abilities, slopes) and has been found to work satisfactorily also for non-normal variables (Scott et al., 2016) such as the pseudoguessing which is modelled using beta distributions. Summary statistics of some of the parameters of interest are shown in Table 2.

On each machine, we obtained subposterior estimates for each of the parameters. For instance, in Figure A7, we show the estimates of the first nine users with good, moderate and poor classification abilities (three on each group). For example, users 1, 3 and 5 have a high ability scores, while 2, 7 and 8 have low abilities. In some cases, there could be a moderate variability on the estimates obtained from different shards, for example user = 2. This is because, some users will have a better performance in some subsets than in the others due to randomness. See Figure A12 for scatter plots comparing the posterior ability estimates in the 10 shards ($\theta_{s1}$, $\theta_{s2}$, etc.) and in the recombined consensus posterior ($\theta$).

Figure 4 shows the user proficiency levels as a function of the proportion of correct classifications. The estimates of abilities can be used in citizen science programs in several ways. For example, in the Serengeti dataset, the true answers were obtained using a majority vote (MV) method. This method is known to fail for very difficult tasks (Raykar et al., 2010). However, this can be overcome using a weighted approach [e.g. weighted majority vote, WMV (Hines et al., 2015; Lintott et al., 2008; Littlestone et al., 1989)] with votes' weights proportional to the ability score as shown in Figure 4. For example, let us consider images from the gold standard dataset and two species, namely giraffe (easy) and reedbuck (hard to identify). As expected, in easy images (giraffe), both methods produce similar accuracy (0.988 using MV vs. 1 using WMV) because most of the participants will have a high probability of correct classification. However, the WMV outperforms the MV substantially in images with reedbucks (0.960 vs. 0.880). This is because the vote in the MV approach can be dominated by inexperienced participants in images that are hard to classify. This is solved in the WMV, by giving more weight to more capable participants.

After a number of contributions, ability scores can be used to identify those participants that require training and qualification. This can easily be done, by selecting those participants from the beginner’s group with a large number of classifications. Moreover, many projects are hosted using Amazon mechanical Turk, where participants are paid for their contributions. Bonuses can be assigned to those with high-quality work, based on the abilities and the number of classifications. Other platforms relying on gamification can also use the ability estimates to rank participants.

Similarly, we obtained posterior distributions for the difficulties in identifying the species (Figure 5; Figure A8). Species with mean positive values are more difficult to identify, for example striped hyena. Other categories with negative estimates such as giraffe and zebra had a high probability of correct classification. Figure 5 shows the posterior estimates of the species difficulties as a function of the proportion of correct classification. On average, citizens misclassified images containing humans with probability 0.10 and this could be considered as an indicator of careless responses.

We also obtained the minimum probability of a correct answer for species due to guessing, for users with extremely low abilities (Figure 6; Figure A9). Interestingly, large pseudoguessing estimates
FIGURE 5  Species difficulties and 95% highest posterior density interval as a function of the proportion of species correct classification. The size of the dots represents the number of times a given species is presented for classification. The category human is used as a baseline.

FIGURE 6  Posterior mean and highest density interval of the species pseudoguessing versus the proportion of correct classification. The size of the dots represents the number of times the species appeared in the images for classification.
are obtained for lions, the only species for which the gender was required (female and male). These species had a higher probability of being correctly identified by less experienced users. Figure A14 displays scatter plots of the correlation between shards, showing a high correlation between the combined estimates ($\eta$) and the ones obtained from the $s$th shards/machine ($\eta_{s,1}$, $\eta_{s,2}$, etc.). Site-related difficulties were also produced. In Figure 7, we show the posterior estimates of the site difficulties. Images from some sites were harder to classify (for instance site 43). A comparison between the 225 sites' subposterior estimates among the 10 shards is presented in Figure A15. These results indicate a high degree of agreement between site difficulties obtained from the different subsets and compared to the consensus values.

5 | DISCUSSION AND CONCLUSION

World-wide networks with thousands of contributors are engaging in citizen science activities. The information provided by this formidable workforce is indispensable in helping to reach aspirational targets such as the United Nations Sustainable Development Goals (Fritz et al., 2019; Hsu et al., 2014).

In the ecological context, the contribution of citizens varies from project to project, but a considerable amount falls into one of these categories: (a) providing verifiable records such as images and audio files from animals, plants, etc. or survey responses, (b) classifying these records to help estimate the distribution of species and also to train machine learning algorithms for automated classification, (c) reporting observations/counts without producing verifiable records (unsubstantiated observations).

However, the data produced by citizens, and hence by these programs, are generally more inaccurate and/or imprecise compared to data from professionals/experts or from traditional experimental or observational studies. Therefore, some sort of data quality assessment is required at some point of the scientific process. This issue has received a lot of attention in recent years. Prior research has also noted the importance of measuring the subjects’ contribution to citizen science projects (Silvertown et al., 2015). Although item response models provide a natural framework in which to investigate and perhaps adjust for these challenges, the literature on their usefulness for crowdsourced programs, particularly in the ecological context, is limited.

Here, we introduced a new family of item response models specially tailored for ecological applications, and an approach to implementing these models for large datasets. These methods can be used in a wide range of applications involving datasets elicited via crowdsourcing. We focused on ecological spatial data frequently collected in citizen science applications and specifically on citizen-elicited annotations of images. We demonstrated the application of these models for real-world data, specifically in the presence of inconveniently large datasets. We found that our proposed framework of modelling outperforms the benchmark (3Pl) method in terms of precision and accuracy, in a simulation experiment, considering factors such as the number of geotagged locations and the number of participants. Similarly, our approach performed better for real-life data. The application of the divide-and-recombine approach for
massive item response problems is novel and can be translated and extended to other settings.

Our methodology provides a more suitable way of assessing the proficiency of the participants, which is useful for comparing them on equal grounds, for measuring their contribution and for building reputation schemes. Clustering the participants into groups based on their abilities, is relevant for citizen science programs and crowdsourcing. For instance, the latent classifications are often obtained by majority or consensus vote. Weighted variations of the voting systems can be easily implemented using as weights the posterior abilities with some penalization based on the variability of these estimates. More effective compensation, for example, payment for their contribution, rewards and bonuses can be readily implemented, for example, in projects using platforms such as Amazon Mechanical Turk. Finally, ability assessment is a powerful resource to detect software bots, careless participants, etc, which should have the lowest ability scores (beginner group in Figure 4).

Our models also identify challenging and commonly misclassified classes of responses (e.g. species) in order to produce better guidance, qualifications and users' training. Obtaining difficulties associated with images gives indications of cameras' technical malfunctioning, misplacement, light conditions, etc. We are currently developing the R package staircase (STAistical Item Response models for Citizen Science applications in Ecology) (https://github.com/EdgarSantos-Fernandez/staircase, staircase) that will allow practitioners to fit the new family of item response models, producing estimates of the participants' abilities along with other relevant ecological measures. The package adopts the Bayesian modelling framework and puts particular emphasis on big data.

Item response models generally require rectangular data, for instance with items in the rows and participants in the columns as in the example shown in Supporting Information. In such datasets, items are classified by multiple participants and missing data are possible. In more traditional settings, such as educational applications, each item needs to be classified in the order of hundreds of times, to obtain suitable estimates of the parameters. However, in our case, we can relax this requirement because the spatial dependency between items allows borrowing strength in space. Secondly, the Bayesian paradigm of modelling allows the use of smaller sample sizes by using hierarchical priors or other methods such as empirical Bayes.

These methods empower citizen science project managers in multiple ways. By design, projects such a Snapshot of the Serengeti allow images to be classified by multiple participants until a consensus is obtained and the image is retired. In other set-ups, the study is designed to achieve the estimates of the parameters with a specified precision or until a given sample size is reached. In this spirit, images that are hard to classify, due, for example, to a high disagreement between participants, can be assigned to more competent or skilled participants (those with a higher ability score) such as those we identified in Figure 4. Similarly, estimates of the images' difficulties allow assignment of images in a more optimal way, with beginners assigned to easy images first. The Bayesian architecture that we implemented allow imputing or predicting, borrowing strength from data to predict where there are gaps in information or regions in the parameter space.

Our approach could be extended in a number of ways. For example, many projects rely on unsubstantiated observations (case III mentioned above) such as the United Kingdom's breeding bird survey (Harris et al., 2015), where participants report observations/counts and no records such as images or audio files are provided. This makes it more difficult to verify the quality of the submitted data. To ensure data quality, these programs generally require qualification of participants before engaging in the project. For example, many surveys have several courses and training programs where participants learn from images and audio to identify species. In these training activities, the performance of the participants is assessed based on a set of audio/images where the ground truth is known (training set). Generally, the whole or a section of the training set is completed by all the participants. This allows fitting item response models for estimating the participants' abilities and proficiency, considering the difficulty of the task. This is helpful in many ways. For example, we can weigh the evidence the participants submit afterwards, so that data from participants with higher ability scores should have more weight in the models. Moreover, beginners can be assigned to less challenging areas.

In this work, participants' abilities were considered to be constant across the classification period. This is reasonable because the period is generally short for many citizen science projects. However, in some circumstances, this condition does not hold, for example, an increased level of skill is achieved with time as participants learn and train. Consequently, variations such as those suggested by Martin and Quinn (2002); Dunson (2003), Wang et al. (2013), and Weng et al. (2018) can be adopted. Additionally, the assumption of normally distributed posterior abilities' parameters could be violated when using crowdsourcing data, for example, when the latent distribution is a mixture distribution. We could tackle this problem in our Bayesian framework by implementing an approach similar to that suggested by Reise et al. (2018). The estimation of multiple ability components known as multidimensionality or MIRT could also provide a better understanding of what skills are required to correctly classify certain types of images. Several MIRT modelling alternatives can be found in Reckase (2009); and van der Linden (2016). In the case study, we considered participants with five classifications or more because obtaining suitable ability estimates requires several data points, keeping in mind that very precise estimates are generally obtained after hundreds of classifications. This makes the inferences challenging for participants who engage with only a few classifications. Our case study involves a relatively small geographical scale. Further work should consider items collected over broader scale surveys, at country and continent levels, or even globally.

In many projects, observations are spatially aggregated in hotspots, for example, near roads or tracks. Other regions remain unobserved or have only a few observations. Our model assumes spatial dependence in the item difficulties and it allows interpolation of the latent difficulty by borrowing strength from neighbouring
sites or kriging. It only requires generation of the coordinates of the unobserved areas, which will create entries in the covariance matrix to establish spatial autocovariance between observed and prediction sites. We could also obtain estimates of the probability that a given participant (with a known ability estimate) correctly classifies an image taken from those hypothetical locations. A possible extension is to account for spatial dependence in the guessing and slope parameters. We assessed this option in our case study, but it increased the computational burden and was prohibitive for our big data problem. It also made the model overparameterized, causing MCMC convergence issues. Other factors affecting the quality of the photograph and the difficulty such as camera-specific parameters and time of the day could be included in the model if available.

In terms of scalability, variational approximations could be used to speed up the computation on big datasets (Hui et al., 2017; Natesan et al., 2016). As most citizen science projects are ongoing, we are currently exploring an online updating approach for processing data as it becomes available (Callaghan, Poore, et al., 2019; Särkkä, 2013; Schifano et al., 2016). In this proposal, we use the posterior estimates from past datasets to construct the prior distributions used by the new models.

Summing up, this paper has introduced a new item response statistical framework for citizen science data, which accounts for spatial dependence. The models allow several popular spatial structures including conditional autoregressive priors, Gaussian processes, etc. It enhances our understanding of citizen science and provides new ecological insights, such as the difficulty associated with species and how much guessing is involved. The models also produce estimates of site-related difficulties. One of the most useful outputs is the categorization of participants into clusters based on their skill levels, which can help program managers to design more targeted and more cost-effective training and qualification.

These findings have several important implications for future citizen science implementations, including better use of the data and the implementation of more efficient majority voting algorithms to estimate the true answers on images. From the computational point of view, it guides handling and making inferences using big datasets. More importantly, this work helps to add more rigour to the scientific process involving citizen science data.

ACKNOWLEDGEMENTS

This research was supported by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS) and the Australian Research Council (ARC) Laureate Fellowship Program under the project ‘Bayesian Learning for Decision Making in the Big Data Era’ (ID: FL150100150). The manuscript was substantially improved based on the detailed comments and suggestions of two reviewers, the associate editor and the editor. The authors thank Erin Peterson for the helpful discussions during the preparation of the manuscript. The authors declare that they have no conflict of interest. Data analysis and visualizations were undertaken in R software (R Core Team, 2018) using the packages RStan (Stan Development Team, 2018), tidyrverse (Wickham, 2017), ggcorrnoi (Garrett et al., 2018) and ggrepel (Slowikowski, 2019), bayesplot (Gabry & Mahr, 2018) and ggforce (Pedersen, 2019).

AUTHORS’ CONTRIBUTIONS

E.S.-F. and K.M. conceptualized and designed the study and drafted the manuscript; E.S.-F. wrote the R codes. Both the authors made a substantial contribution and approved the final submission.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13623.

DATA AVAILABILITY STATEMENT

The dataset and the R/Stan codes used in the case study (Section 4.3) can be found in the repository https://github.com/EdgarSantos-Fernandez/hakuna. These files have been archived on Zenodo (https://doi.org/10.5281/zenodo.4702772, Santos-Fernandez) and the latest released version can be found in https://zenodo.org/badge/latestdoi/353582115. See other examples in https://github.com/EdgarSantos-Fernandez/staircase.

The following files are part of the repository:

serengeti.RDS: contains the Serengeti dataset.
Item_response_modeling_for_big_data.Rmd: this R markdown file includes the full R/Stan codes for fitting the models, combining the posterior estimates and for visualizing the results.
Item_response_modeling_for_big_data.pdf: this is the knitted R markdown output file.

ORCID

Edgar Santos-Fernandez https://orcid.org/0000-0001-5962-5417
Kerrie Mengersen https://orcid.org/0000-0001-8625-9168

REFERENCES

Altwegg, R., & Nichols, J. D. (2019). Occupancy models for citizen science data. Methods in Ecology and Evolution, 10, 8–21. https://doi.org/10.1111/2041-210X.13090
Arab, A., & Courter, J. R. (2015). Spatio-temporal trend analysis of spring arrival data for migratory birds. Communications in Statistics - Simulation and Computation, 44, 2535–2547. https://doi.org/10.1080/03610918.2013.809100
Arab, A., Courter, J. R., & Zelt, J. (2016). A spatio-temporal comparison of avian migration phenology using citizen science data. Spatial Statistics, 18, 234–245. https://doi.org/10.1016/j.spa.2016.06.006
Baker, F. B., & Kim, S.-H. (2004). Item response theory: Parameter estimation techniques. CRC Press.
Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). Hierarchical modeling and analysis for spatial data. CRC Press.
Besag, J. York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics, 43, 1–20. https://doi.org/10.1007/BF00116466
Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krockeck, M., Stuart-Smith, J. F., Peci, G. T., Barrett, N., & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. Biological
Santos-Fernandez E, Mengersen K. Understanding the reliability of citizen science observational data using item response models. Methods Ecol Evol. 2021;00:1–16. https://doi.org/10.1111/2041-210X.13623