Integrating citizen-science and planned-survey data improves species distribution estimates

Author(s): Viviane Zulian, David A. W. Miller and Gonçalo Ferraz

Source: *Diversity and Distributions*, December 2021, Vol. 27, No. 12 (December 2021), pp. 2498-2509

Published by: Wiley

Stable URL: https://www.jstor.org/stable/10.2307/48632843

Received: 7 April 2021 | Revised: 24 August 2021 | Accepted: 5 September 2021

DOI: 10.1111/ddi.13416

Diversity and Distributions WILEY

RESEARCH ARTICLE

Integrating citizen-science and planned-survey data improves species distribution estimates

Viviane Zulian[1] | David A. W. Miller[2] | Gonçalo Ferraz[1]

[1]Programa de Pós-Graduação em Ecologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[2]Department of Ecosystem Science and Management, Pennsylvania State University, University Park, Pennsylvania, USA

Correspondence
Viviane Zulian, Programa de Pós-Graduação em Ecologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, CP 15007, Porto Alegre, RS 91501-970, Brazil.
Email: zulian.vi@gmail.com

Funding information
Instituto Humanize; Programa de Internacionalização (PrInt/CAPES); Fundo Brasileiro para a Biodiversidade (FUNBIO); Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Number: PP 312606/2013-3; The Rufford Foundation, Grant/Award Number: 19835-1 and 23709-2

Editor: Martin Jung

Abstract

Aim: Mapping species distributions is a crucial but challenging requirement of wildlife management. The frequent need to sample vast expanses of potential habitat increases the cost of planned surveys and rewards accumulation of opportunistic observations. In this paper, we integrate planned-survey data from roost counts with opportunistic samples from eBird, WikiAves and Xeno-canto citizen-science platforms to map the geographic range of the endangered Vinaceous-breasted Parrot. We demonstrate the estimation and mapping of species occurrence based on data integration while accounting for specifics of each dataset, including observation technique and uncertainty about the observations.

Location: Argentina, Brazil and Paraguay.

Methods: Our analysis illustrates (a) the incorporation of sampling effort, spatial autocorrelation and site covariates in a joint-likelihood, hierarchical, data integration model; (b) the evaluation of the contribution of each dataset, as well as the contribution of effort covariates, spatial autocorrelation and site covariates to the predictive ability of fitted models using a cross-validation approach; and (c) how spatial representation of the latent occupancy state (i.e. realized occupancy) helps identify areas with high uncertainty that should be prioritized in future fieldwork.

Results: We estimate a Vinaceous-breasted Parrot geographic range of 434,670 km^2, which is three times larger than the "Extant" area previously reported in the IUCN Red List. The exclusion of one dataset at a time from the analyses always resulted in worse predictions by the models of truncated data than by the Full Model, which included all datasets. Likewise, exclusion of spatial autocorrelation, site covariates or sampling effort resulted in worse predictions.

Main conclusions: The integration of different datasets into one joint-likelihood model produced a more reliable representation of the species range than any individual dataset taken on its own, improving the use of citizen-science data in combination with planned-survey results.

KEYWORDS

citizen-science, data integration models, endangered species, geographic range, occupancy models, species distribution models, Vinaceous-breasted Parrot

1 | INTRODUCTION

Wildlife management depends on knowledge about species' geographic ranges, which is also a key element of threat assessment criteria used by the International Union for Conservation of Nature (IUCN, Mace et al., 2008). Despite their unequivocal relevance, accurate range maps are scarce (Jetz et al., 2012). Efforts to improve knowledge about species ranges are hindered by the extent of necessary field sampling and by the scarcity of funding for monitoring. The sampling challenge is heightened by the inevitable trade-off between data quantity and quality. Planned surveys with replicated samples of a predetermined set of locations using standardized protocols that note the presence or absence of target species provide high-quality information, but they are few and far between. Large and long running planned surveys such as the North American Breeding Bird Survey (BBS; Hudson et al., 2017) or the Pan-European Common Bird Monitoring scheme (PECBM; Gregory et al., 2005) are exceptions to a global pattern of "opportunistic" collection of mostly presence-only data, which records where a species is detected but not where it is searched for and not found, in contrast with presence–absence data, which records where a species is and where it is not detected.

Technological advances have produced many collaborative initiatives where volunteers share wildlife sightings from opportunistic records in easily accessible online platforms. These initiatives fall under the broad umbrella of citizen science (Heigl et al., 2019; Tulloch, 2013; Wiggins & Crowston, 2011). Due to the popularity of birdwatching, citizen-science platforms now hold an extraordinary amount of spatially indexed bird detections. Outstanding examples include the global eBird (Sullivan et al., 2009) and Xeno-canto (Xeno-canto, 2019) platforms, as well as the Brazilian WikiAves (WikiAves, 2019). These platforms hold data for thousands of bird species, with increasing spatial coverage. These huge datasets have the potential to fill gaps in our knowledge of species' distributions (Altwegg & Nichols, 2018; La Sorte & Somveille, 2020; Sullivan et al., 2017). There are, however, wide variations in sampling technique, expertise, and effort among observers, as well as differences in data structures and spatial coverage among citizen-science platforms. The ability to integrate data from different sources is therefore important. This has spurred progress in the construction of statistical species distribution models that integrate multiple data streams for mapping the probability of species presence over a region of interest (Fletcher et al., 2019; Isaac et al., 2020; Miller et al., 2019).

Initial work on data integration methods used presence–absence datasets as an accessory to the analyses of larger presence-only datasets. Seminal papers by Dorazio (2014), Fithian et al. (2015), and Giraud et al. (2016) integrated presence-only data from opportunistic samples with presence–absence data from planned surveys in a spatial point-process, joint-likelihood framework. The resulting data integration models use the sampling effort information in presence–absence data to improve inference from the usually larger, presence-only datasets that lack information about effort. This approach has been extended to account for local habitat heterogeneity (Coron et al., 2018) and data patchiness (Peel et al., 2019). In one wide-ranging study, Pacifici et al. (2017) showed how data integration can include site covariates, account for spatial autocorrelation, address false positive detections, combine counts with presence–absence data and weigh datasets differently based on their quality. Simmonds et al. (2020) recently explored the limits of data integration, asking when more data are not necessarily better. These efforts demonstrated how data integration can not only account for limitations of presence-only data, but also flexibly and robustly harmonize a wide-range of data types (Isaac et al., 2020; Miller et al., 2019).

The early emphasis on integrating widely available, opportunistic data from citizen-science sources with explicit sampling information from planned-survey, presence–absence data may have concealed the extraordinary amount of sampling information contained in citizen-science datasets themselves (but see previous analyses of sampling information from citizen-science sources, e.g. Kéry et al., 2010). The set of data points indicating detection of one focal species in a citizen-science platform may not explicitly convey the effort that went into searching for that species; nonetheless, because platforms gather observations from multiple species, one can find abundant information about sampling effort by looking at where and when non-focal species were detected (Hill, 2012; Phillips et al., 2009). Indeed, citizen-science data frequently include information that can be used to estimate sampling effort, such as number of observers, time and distance travelled during sampling, number of detections of all species or number of species detected. Here, we build on previous work by Fithian et al. (2015), Pacifici et al. (2017), Stauffer et al. (2018) and Miller et al. (2019), to develop a static, integrated occupancy model of species distribution. Our approach assembles detection non-detection information for each sampling unit and accounts for imperfect detection within each data source in the integrated model via the estimation of sampling effort per source. To assess the extent to which our accounting of sampling effort improves distribution models, we employ a cross-validation approach that measures the ability of different models to predict randomly excluded data points. Such assessment of model fit also reveals the extent to which data integration, spatial autocorrelation and site covariates contribute to the modelling task.

Accurate range maps are especially needed for threatened or endangered species in regions that lack planned wildlife surveys, as is often the case in the tropics. The Vinaceous-breasted Parrot (VBP, *Amazona vinacea*) is an endangered species, endemic to the tropical South American Atlantic Forest (BirdLife International, 2017). Showing substantial uncertainty about the species' geographic range, the IUCN reports a "possibly extant" VBP area that is almost three times as large as the "extant" area (Figure 1a, BirdLife International & Handbook of the Birds of the World, 2016). In a recent study of VBP abundance, Zulian et al. (2020) show how ~75% of known communal roost sites are outside the IUCN "extant" area, suggesting current range estimates are inadequate for planning purposes. This motivated us to ask how VBP data sources could be combined to generate a better estimate of the species' range and identify

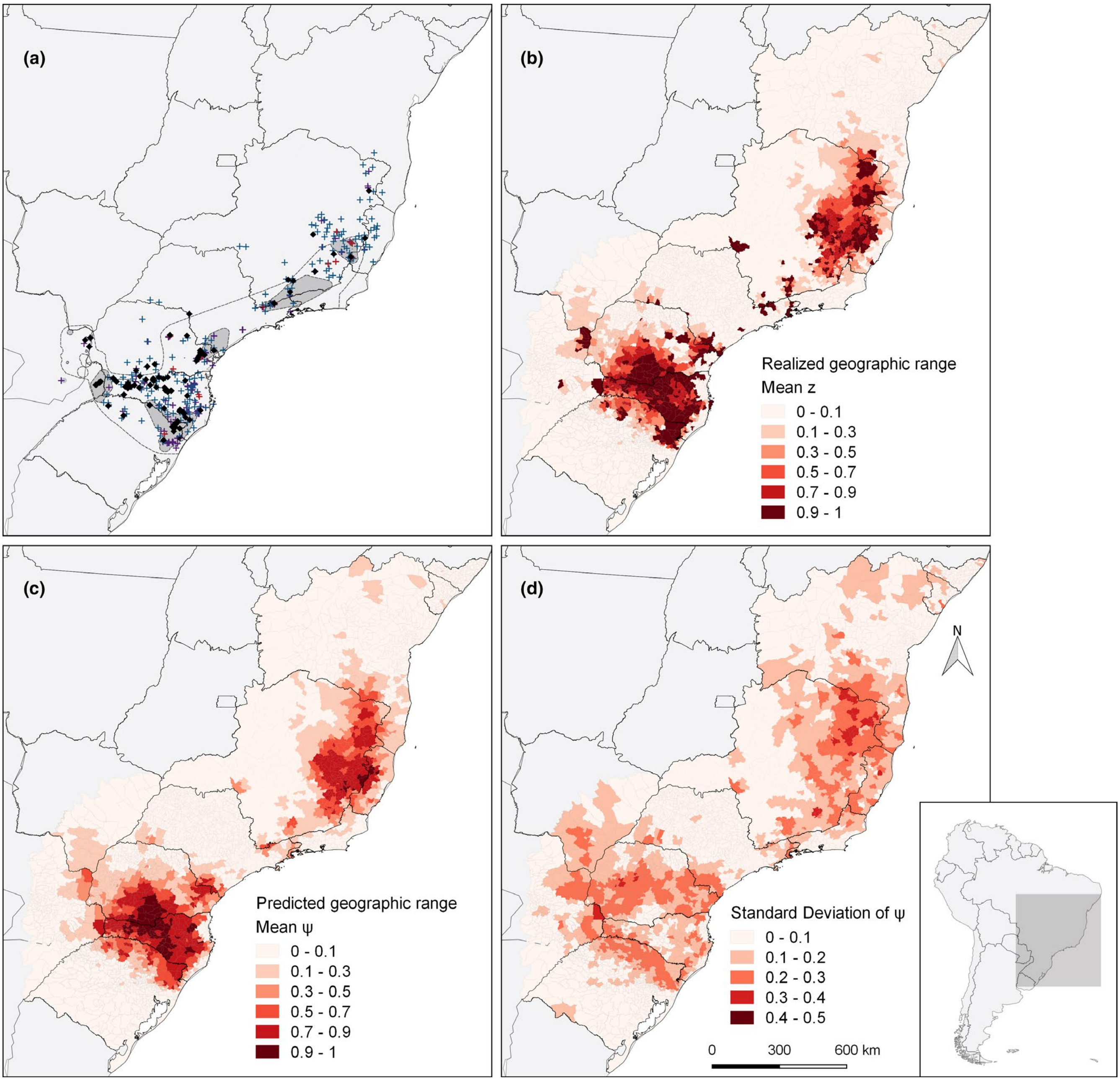

FIGURE 1 Vinaceous-breasted Parrot observations, geographic distribution and uncertainty about the distribution. Panel a maps Vinaceous-breasted Parrot detections analysed in this study with black diamonds indicating the location of roost counts and crosses the location of citizen-science (eBird-in red, WikiAves-in blue and Xeno-canto-in purple) records. Grey polygons represent the IUCN "Extant" range and dashed lines delimit the IUCN "Possibly Extant" range of the Vinaceous-breasted Parrot. Panel b represents realized occupancy (mean *z*). Panels c and d show, respectively, the predicted occupancy (ψ) and the standard deviation of its posterior distribution. Estimates in panels b, c and d are based on the Full Model fit to all (roost counts, eBird, WikiAves and Xeno-canto) datasets. Spatial units correspond to municipalities, with darker tones of red representing higher occupancy (b, c) and higher standard deviation (d)

where the greatest uncertainty in the current distribution exists. We set out to characterize the spatial extent of the current distribution, estimating the local probability of the species' presence (Kéry, 2011) and quantifying the uncertainty about these probability estimates (Rocchini et al., 2011).

We aim here to (a) demonstrate how data integration models can be harnessed to address differences in data collection across multiple datasets by accounting for variation in sampling effort and detection probability between and within datasets; (b) develop an approach to assess the predictive value of including or excluding different data streams in a single integrated model; and (c) assess how modelling decisions affect the predictive power of our models, with particular attention to the choice of occupancy and detection covariates, whether and how to account for residual spatial autocorrelation, and how effort and detection are related. We integrate planned-survey data collected by research teams (Zulian et al., 2020) with citizen-science data from the eBird (eBird, 2019), WikiAves (WikiAves, 2019) and Xeno-canto (Xeno-canto, 2019) platforms to model the VBP geographic range in an eleven-year period.

2 | METHODS

2.1 | Study area

Our study area comprises 2,449,757 km^2 divided into 3,701 municipalities from Argentina, Brazil and Paraguay (Figure 1a). This area includes the entire IUCN-delimited VBP "Possibly Resident" range (BirdLife International & Handbook of the Birds of the World, 2016) and is bounded by the limits of the Atlantic Forest biome (Olson et al., 2001). Considering the absence of VBP records north of the Brazilian state of Bahia (BirdLife International, 2017), we set the northern limit of our study area along the northern borders of that state and the adjacent state of Alagoas.

2.2 | Data collection

We obtained VBP detection–non-detection data for all 3,701 municipalities collected between 1 January 2008 and 31 December 2018. We chose the municipality as our spatial unit because WikiAves data register the location of observations by municipality name, without spatial coordinates and because municipality limits are easily recognized by decision-makers and residents. "Occupancy" is given by the presence of VBPs in a municipality during the eleven-year study period. Our data come from four sources: roost counts, WikiAves, eBird and Xeno-canto. Roost counts were performed by researchers (Zulian et al., 2020), while WikiAves, eBird and Xeno-canto data were uploaded to citizen-science platforms by volunteer observers.

Roost counts were performed between 2014 and 2018 by 26 teams in 74 municipalities of Brazil, Argentina and Paraguay, following methodological guidelines described by Zulian et al. (2020). Between one and 25 counts per site were taken each year, between April and June, on sites known by researchers to have VBP roosts. Roost count data were converted into detection/non-detection histories with counts from the same municipality considered as replicate samples. Counts with at least one parrot received a "1" (detection) and counts with no parrots received a "0" (non-detection) in the binary history. Parrots are observed in relatively narrow time windows near dawn and dusk, but early arrivals or a late departure from the roost influence the observations, so we measured the count's duration in minutes (Time Observing = TObs) as an effort covariate.

We obtained eBird data from birding checklists with observations in our study area and uploaded to the platform throughout the study period. Our analysis included only complete checklists—where the observers recorded all the species they were able to identify—and excluded all checklists, which did not identify a municipality or that potentially spanned more than one municipality due to long distance (>12 km) or long time (>360 min) travelled. Checklists from the same municipality were treated as replicate samples. The checklist structure made it easy to convert eBird data into detection/non-detection format, and we accordingly built eBird detection/non-detection histories that register the detection (1) or non-detection (0) of the VBP for each list of each municipality. eBird effort covariates were the number of species recorded in a list (*SSee*), minutes spent observing (*TObs*) and kilometres travelled (*RLen*).

WikiAves receives observer input in the form of individual photographs or audio recordings of an identified species and has expert moderators checking uploaded content to avoid misidentification. Record location is registered as a municipality name along with information about authorship and comments. We obtained the total number of WikiAves records uploaded to each municipality of our study area and period, and recorded detection/non-detection as only one data point per municipality, without replication at the municipality level. Thus, there is only one vector of WikiAves detection/non-detection data, with length equal to the number of municipalities and values of "1" or "0," respectively, for those municipalities that did or did not have at least one VBP photograph or audio recording. Effort covariates were the number of photos (*NPho*) and audio recordings (*NAud*) submitted to WikiAves per municipality.

Xeno-canto hosts only audio recordings of bird sounds (Xeno-canto, 2019). We used the R package *warbleR* (Araya-Salas & Smith-Vidaurre, 2017) to download the list of all Xeno-canto records from our study area and period. Our Xeno-canto unit data are the set of all audio recordings from one municipality, without replication. We organized these detection/non-detection data in the same vector format as WikiAves' and used the number of recordings (*NAud*) uploaded in each municipality as a covariate of sampling effort. Unlike WikiAves, Xeno-canto does not have its content checked by moderators, but we did confirm identification of all Xeno-canto VBP records. Unlike eBird, neither Xeno-canto nor WikiAves records can be organized as complete lists of every species that an observer identified in a given space and time.

2.3 | Data analysis

We summarized each of our four data sources in a matrix or a vector of detection–non-detection information per municipality, depending, respectively, on whether they had multiple (roost counts, eBird) or a single (WikiAves, Xeno-canto) observation per municipality. Effort covariates matrices (or vectors) took the corresponding data source shape. In our models, the true occupancy state of each municipality (or site) *i* is denoted as z_i, which takes the value 1 when site *i* was occupied and 0 when not. The state of this latent (partially observed) variable follows a Bernoulli distribution with mean ψ_i:

$$z_i \sim \text{Bernoulli}(\psi_i). \tag{1}$$

We allowed the probability ψ_i that site i is occupied by VBPs to vary with respect to three site environment covariates, with a logit link function. As VBPs are endemic to the Atlantic Forest and appear to be associated with both altitude (BirdLife International, 2017) and Araucaria forest cover (BirdLife International, 2017; Cockle et al., 2019; Collar et al., 2017; Tella et al., 2016), we included Atlantic forest cover (AtF_i), Araucaria forest cover (ArF_i) and average altitude (Alt_i) as covariates of municipality i occupancy. Forest cover values are from Ribeiro et al. (*in preparation*) as proportions of the municipality area. Average municipality altitude x, in metres, is from DIVA-GIS (2018), log-transformed as $\log(x+1)$. Our linear model of occupancy also included a spatial random effect to account for unexplained spatial autocorrelated variation (δ_i):

$$\mathrm{logit}(\psi_i) = \beta_0 + \beta_1 * \mathrm{AtF}_i + \beta_2 * \mathrm{ArF}_i + \beta_3 * \mathrm{Alt}_i + \delta_i. \quad (2)$$

This effect follows a conditional auto-regressive (CAR) distribution as applied by Pacifici et al. (2017) in the context of integrated species distribution models. To avoid confounding effects of municipality size variability and to gain sampling replication within spatial units in the CAR analysis, we represented space by a hexagonal lattice overlaid on the study area, with municipalities assigned to the lattice cell that matches their centroid. Cells measured 0.5° latitude across; all the first-order neighbours of each cell were given a weight of 1 when fitting the CAR model.

We fit a joint-likelihood data integration model with a single shared occupancy process: for all four data types, VBP detection in sample j and site i is conditional on the species being present at the site ($z_i = 1$). Departing slightly from the standard accounting of effort based on the number of replicate samples (MacKenzie et al., 2002), we express the conditional probability (p_j^*) of detecting the species as a function of an estimated amount of sampling effort (E_j) for sample j (Miller et al., 2019; Stauffer et al., 2018):

$$p_j^* = 1 - (1-p)^{Ej}, \quad (3)$$

where p is the probability of detection per unit effort. Because we are using indirect, and sometimes several metrics of effort for each data source (our effort covariates), we estimate parameter E_j for each sample j as a linear function of the covariates. Thus, for each data source (RC = roost counts, EB = eBird, WA = WikiAves and XC = Xeno-canto) we have:

$$E_j^{\mathrm{RC}} = \alpha_1 * \mathrm{TObs}_j \quad (4a)$$

$$E_j^{\mathrm{EB}} = \alpha_2 * \mathrm{SSee}_j + \alpha_3 * \mathrm{TObs}_j + \alpha_4 * \mathrm{RLen}_j \quad (4b)$$

$$E_j^{\mathrm{WA}} = \alpha_5 * \mathrm{NPho}_j + \alpha_6 * \mathrm{NAud}_j \quad (4c)$$

$$E_j^{\mathrm{XC}} = \alpha_7 * \mathrm{NAud}_j. \quad (4d)$$

Equations (4a–d) have no intercept, so that effort is 0 when all effort covariates are 0. In addition, we fix p at a value of .5, so that the $\alpha_1 - \alpha_7$ coefficients express the relationship between covariates and the effort necessary to reach a detection probability of .5 per unit of effort. Without fixing p, Equation (3) becomes over-parameterized. Coefficients $\alpha_1 - \alpha_7$ of the effort functions also show the relative contribution of each covariate to the total estimated effort per dataset (see code in Appendix S1). Finally, our detection/non-detection histories Y_{ij} in each dataset follow the Bernoulli distribution:

$$Y_{ij} \sim \mathrm{Bernoulli}(z_i \times p_j^*). \quad (5)$$

We first fitted a Full Model accounting for the effects of all effort metrics, all site covariates and spatial autocorrelation. Subsequently, we evaluated the impact of different modelling decisions on predicted accuracy by fitting 11 additional models listed in Table 2. We fitted all the models using a Bayesian estimator coded in the BUGS language and run on WinBugs software (Lunn et al., 2000), which includes predefined model structures for CAR random effects. Inference was based on draws from the posterior distribution of model parameters using an MCMC algorithm with three chains, 200,000 iterations, and a burn-in phase of 100,000. We considered parameters with an R-hat lower than 1.1 to have converged and used results to draw parameter posterior distributions.

We assessed model fit by excluding all the detection non-detection data from a randomly selected set of 650 municipalities (20% of the total), fitting the models to the training dataset (i.e. remaining data) and then predicting the validation dataset (excluded data) based on the estimated parameters. In this cross-validation approach, our prediction accuracy measures a model's ability to predict excluded data as expressed by the likelihood-based Deviance:

$$D = -2 \sum \log(\mathscr{L}), \quad (6)$$

where the likelihood $\mathscr{L}$ equals $\hat{y}^{y} * (1-\hat{y})^{1-y}$ for each site and visit in the validation dataset (Hooten & Hobbs, 2015). We use y and $\hat{y}$ to represent, respectively, the observed, binary data and the predicted probability of detecting VBPs for each site and visit based on estimates from the training data. The lowest deviance values indicate the best fit. We examined overall model deviance, summed across data sources, as well as individual deviance values for each data source to look at source-specific predictive performance. Comparisons among values also revealed the impact of site covariates, detection covariates and the CAR component on the predictive performance of our models.

To determine whether each of the individual datasets improved the predictive ability of our model, we fit the model to four truncated datasets, including all covariates and the CAR random effect, but excluding one data source at a time (Models 5–8, Table 2). Such rotating exclusion made it possible to examine whether the addition of a data source to the mix improves the model's ability to predict the validation set from other sources. Specifically, we asked whether predictions of validation data from a training data source were more or less accurate when each of the other data sources were excluded.

For example, if eBird does contribute to improving the overall model, then including eBird data should lead to better predictions of Xeno-canto, WikiAves and roost count data. This is a measure of overall prediction consistency among data sources. To better assess the usefulness of data integration, we also fit four models that retain the site covariate and CAR components of the Full Model, but include only one data source at a time (Models 9–12, Table 2).

Finally, we represent the VBP geographic range using two estimates of site occupancy. The first, "realized" occupancy, is conditional on the observations; it equals 1 in all municipalities where VBP was seen at least once, and is the expected value of the latent occupancy state (z_i) where it was not seen. As effort increases and VBPs are not observed, z converges towards 0, and so does realized occupancy. Even though z_i can only be 0 or 1, "realized" occupancy, the expected value of z_i, obtained by averaging the MCMC chain for z in site i can take values between 0 and 1. This metric provides a measure of local uncertainty about species presence given all available data and, unlike typical predictions by distribution models, accurately expresses local certainty of occurrence by adjusting predictions to actual observation. The second estimate, "predicted" occupancy, offers estimates of ψ_i, which express occupancy probability for a statistical population of municipalities with the same site covariates and neighbourhood of municipality i (Figure 1c). Predicted occupancy is not conditioned on the actual data for a municipality: unlike z_i, which always equals 1 if the species was detected at site i, ψ_i can be smaller than 1 in municipalities where the species was detected. Predicted occupancies are typically visualized in distribution models, expressing how estimated environmental relationships affect the local probability of occurrence across a species range.

3 | RESULTS

We draw on 1,007 VBP detections from 47,240 samples in four datasets collected across the 3,402 municipalities within our study area (Table 1). While the roost count data contains 40% of all detections, roost counts covered only 2.2% of the municipalities in our study area. The highest detection rate—given by the ratio of n_{det} to Sample size, in Table 1—appears in the roost count dataset (88%), as expected, because roost counts were only carried out in locations where VBPs were known to occur. This resulted in the highest detection probability per sample among all datasets ($p = .87 \pm .144$; Table 1). The 596 detections jointly returned by the three citizen-science platforms, on the other hand, come from 3,401 municipalities, 92% of the number of municipalities in the study area. WikiAves had the widest coverage, with data for 3,190 municipalities, and VBP detections for 191 of them. One WikiAves sample comprises all the photographs and recordings submitted for one municipality, a large amount of effort per sample, so WikiAves had the highest detection rate and, naturally, the highest detection probability per sample of all citizen-science sources. eBird had smaller coverage than WikiAves but had the largest number of VBP detections of all sources: 388 from 71 municipalities. Differently from WikiAves (and Xeno-canto), one eBird sample is not the set of all records in one municipality, but one birding list. The number of eBird samples varied substantially across municipalities, ranging from 1 to 3,244 (São Paulo, SP, Brazil) with a mean of 42. With so many samples and relatively little effort per sample, eBird had the lowest detection rate, of 1%, and the lowest estimated detection probability of all platforms ($p = .06 \pm .003$; Table 1). Xeno-canto, with the smallest coverage and number of VBP detections had intermediate values of both detection rate and estimated p.

Table 2 shows model predictive ability based on cross-validation. The Full Model had the best predictive ability. Exclusion of detection covariates (Model 3) had the greatest negative impact on predictive ability, with estimated deviance being 2.15 times higher for this model than for the Full Model (Table 2). Removal of the CAR component (Model 2) had an intermediate but measurable effect on deviance, with residual spatial structure (Figure S1) visibly influencing the distribution map (Figure S2). The values in Table 2 result from one trial of data exclusion and prediction. We performed another two trials of this procedure for the first four models in Table 2 with consistent results for total deviance. The ranking of models with respect to specific dataset deviances changed between trials, but it showed a tendency for better prediction with the Full Model and worse prediction when detection covariates are excluded (Table S1).

TABLE 1 Sample size, spatial coverage and number of Vinaceous-breasted Parrot detections from roost counts (RC), eBird (EB), WikiAves (WA) and Xeno-canto (XC)

Datasets	Sample size	Coverage	n_{det}	n_{muni}	Sampling unit	p
RC	466	74	411	60	Count	.87 ± .144
EB	42,855	1,274	388	71	List	.06 ± .003
WA	3,190	3,190	191	191	Municipality	.25 ± .011
XC	729	729	17	17	Municipality	.08 ± .015
Total	47,240	3,402	1,007	339		

Note: Sample size is number of samples, following each database's sampling unit definition. Spatial coverage is the number of municipalities sampled, with total smaller than the sum across databases because some municipalities are included in more than one database. Labels n_{det} and n_{muni} show, respectively, the number of parrot detections and the number of municipalities with at least one detection. The sampling unit is the data category considered as a replicate; p is estimated detection probability per sampling unit at average effort for each dataset, under the Full Model.

TABLE 2 Deviance for each site-occupancy model in this study

		Deviance in each dataset			
Models	Total deviance	RC	EB	WA	XC
1. Full Model	**440.85**	**28.84**	**281.19**	**103.35**	**27.46**
2. No CAR	581.32	50.97	362.58	139.56	28.20
3. No detection covs.	952.84	57.21	735.61	133.60	26.41
4. No occupancy covs.	477.06	26.04	315.34	107.79	27.87
5. All data but RC	—	—	301.88	108.78	27.56
6. All data but EB	—	28.53	—	110.26	25.55
7. All data but WA	—	35.27	326.01	—	28.61
8. All data but XC	—	34.04	309.29	116.18	—
9. Only RC	—	23.22	—	—	—
10. Only EB	—	—	314.78	—	—
11. Only WA	—	—	—	107.75	—
12. Only XC	—	—	—	—	28.77

Note: Model 1, designated as "Full Model," includes detection as well as occupancy covariates and was fitted to data from all datasets: roost counts (RC), eBird (EB), WikiAves (WA) and Xeno-canto (XC). Model 2 equals model 1 without spatial autocorrelation. Models 3 and 4 are variants of model 1 without, respectively, detection and occupancy covariates. Models 5–8 differ from the Full Model by the exclusion of one dataset each, as shown. Models 9–12 are each fitted to an individual dataset alone. As models 5–12 do not use the same data, their Total Deviance is not comparable and is omitted from the table. Bold font highlights the model with the best fit by Total Deviance. Contrast the values on line 1 with those on lines 5–12 to see sixteen possible comparisons between the Full Model fit to all four datasets (line 1) and the same model fit to different combinations of datasets (lines 5–12).

Our results reveal that effort-based modelling of detection, inclusion of spatial autocorrelation in occupancy, and consideration of occupancy covariates improved the predictive ability of our species distribution models.

Models 5–8 assess whether individual datasets improve overall predictive ability. We compare dataset-specific deviances from the validation data for each of the four models to that of the Full Model (where no data were excluded). Including the four datasets in the analysis (i.e. using the Full Model) improved fit in all but two cases. Dataset-specific deviances of Models 5, 7 and 8 were all higher—indicating lower prediction power— than those of the Full Model (Table 2). Removal of eBird data (Model 6) slightly improved the prediction of Xeno-canto data, but clearly worsened the fit to WikiAves data, leaving that of the roost count data virtually unchanged. The Full Model fit to all data sources did a better job of predicting EB, WA, and XC data than the individual-dataset Models 10–12 themselves. Model 9, which was fit to RC data alone, predicted RC validation data better than the Full Model, but it produced an incongruous realized range map (Figure 2a), with mean z values in excess of 0.9 for hundreds of municipalities where other datasets produced mean z smaller than 0.3 (Figure 2b–d). The realized range map obtained under Model 10, of the EB data alone, missed a large part of the northern VBP distribution (Figure 2b). Data integration under the Full Model improved the prediction of EB validation data more that the prediction for any other dataset.

The sum of municipality areas weighted by the Full Model realized occupancy estimates returned a realized VBP range of 434,670 km^2, which is three times larger than the IUCN Red List "Extant" area (BirdLife International & Handbook of the Birds of the World, 2016). Both the realized and the predicted ranges appear split in two large patches (Figure 1b,c). The southern patch covers parts of Argentina, Paraguay and the Brazilian states of Rio Grande do Sul, Santa Catarina and Paraná; the northern patch overlaps the Brazilian states of Minas Gerais, Espírito Santo and Bahia. The realized range also includes small areas between the two large patches, mainly in the Campos do Jordão region, near the border between São Paulo and Minas Gerais (Figure 1b). Uncertainty about the VBP range is greatest around high-occupancy patch edges, as shown by intermediate values of realized occupancy (Figure 1b) and high standard deviation of predicted occupancy (Figure 1d). As expected, municipalities with the most extreme occupancy values (close to 0 or 1) returned the lowest standard deviation values (Figure 1d).

Araucaria and Atlantic forest cover had strong (and positive) effects on occupancy probability (Table 3). Altitude had a weaker, positive, but more precise effect on ψ, when compared with the two forest covariates. The different effort covariates on the bottom part of Table 3 had varying, though always positive, effects on detection probability. Among these, time spent observing showed the highest effect, both as α_1, which measures the duration of a roost count, and as α_3, the time spent collecting an eBird list. The number of audio recordings uploaded in WikiAves was a stronger predictor of survey effort (α_6), and thus overall detection probability for a municipality, than the number of photos (with effect α_5).

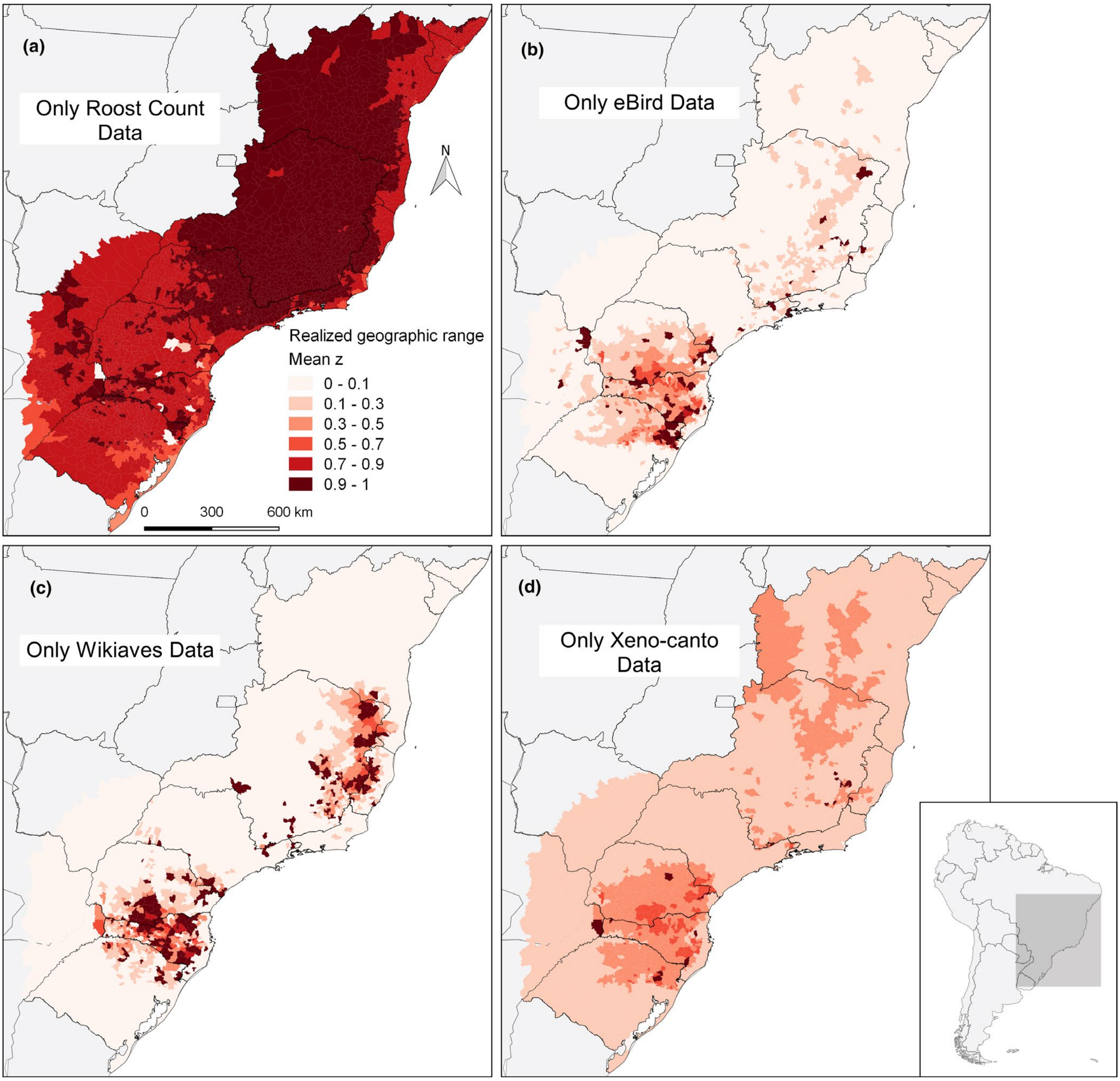

FIGURE 2 Vinaceous-breasted Parrot realized geographic range (mean *z*) based on separate analyses of each dataset. The panels show results based on roost count (a), eBird (b), Wikiaves (c) and Xeno-canto (d) data. Spatial units correspond to municipalities, with darker tones of red representing higher mean *z*; intermediate values—of *z* ~ 0.5—indicate the highest uncertainty about occupancy

4 | DISCUSSION

The Vinaceous-breasted Parrot geographic range covers approximately 434 thousand square kilometres subdivided into two large patches, one centred in the southern Brazilian state of Santa Catarina and another to the north, centred in eastern Minas Gerais state, also in Brazil. A third, much smaller area of occupancy comprises a group of relatively high-altitude municipalities near Campos do Jordão, in São Paulo and Minas Gerais states, approximately 100 km west of the Rio de Janeiro border. Our two-patch range contrasts with the five patches represented in the IUCN "resident" range. The "possibly resident" IUCN range, which encloses all of the "resident" patches, conveys uncertainty about the subdivision in five areas (BirdLife International & Handbook of the Birds of the World, 2016). Our study provides evidence for redrawing the VBP range while quantifying uncertainty associated with the new map. We look forward to seeing population genetic studies that elucidate the extent of reproductive isolation between the two large patches, as well as between the small Campos do Jordão area and the northern, Minas Gerais patch. A comparison between realized and predicted ranges

TABLE 3 Estimated mean, standard deviation (*SD*) and 95% credible intervals (CI) for the posterior distribution of Full Model coefficients

Parameter	Mean ± *SD*	95% CI
Biological process		
β_1 (Atlantic forest cover)	2.110 ± 0.8684	0.379–3.792
β_2 (Araucaria forest cover)	2.133 ± 0.9806	0.296–4.104
β_3 (Altitude)	0.852 ± 0.1205	0.579–1.055
Sampling process		
α_1 (RC: Time observing)	1.814 ± 0.1110	1.613–2.043
α_2 (EB: Species seen)	0.002 ± 0.0002	0.001–0.002
α_3 (EB: Time observing)	0.008 ± 0.0027	0.003–0.013
α_4 (EB: Route length)	0.005 ± 0.0019	0.002–0.009
α_5 (WA: Photos)	0.001 ± 0.0002	0.001–0.002
α_6 (WA: Audio recordings)	0.006 ± 0.0021	0.003–0.011
α_7 (XC: Audio recordings)	0.007 ± 0.0017	0.004–0.011

Note: Occupancy function coefficients (β_1 to σ) specify the biological process, while detection coefficients (α_1–α_7) specify the sampling process. The covariates corresponding to each coefficient appear in parentheses in front of its name; σ measures the magnitude of spatial autocorrelation in site occupancy. Coefficients α_1, α_2–α_4, α_5–α_6 and α_7 correspond, respectively, to metrics of effort per municipality in roost counts (RC), eBird (EB), WikiAves (WA) and Xeno-canto (XC) databases. Each metric is indicated in parentheses in front of the coefficient name.

shows that some municipalities with high mean *z* have relatively low predicted occupancy probability (ψ). We trust the WikiAves moderation system, have no doubts about VBP identification in the roost counts, and manually checked every VBP record from Xeno-canto; but still, we cannot rule out the possibility of some false positive observations in these municipalities. Occasional discrepancy between mean *z* and ψ could also derive from the observation of animals released or escaped from captivity. These municipalities deserve further investigation, particularly those in south-west Minas Gerais and south-west São Paulo, to exclude the possibility of there being unknown isolated populations. Intermediate values of realized occupancy and high standard deviation of the posterior distribution of predicted occupancy reveal areas with high uncertainty about VBP presence, which, like the isolated high-*z* municipalities, ought to be targeted by future field searches. Three regions stand out for high uncertainty about VBP presence: northeastern Minas Gerais, central Paraná, and northern Rio Grande do Sul, in Brazil, together with a few municipalities in eastern Paraguay. These are the regions that could contribute most to further improvement of knowledge about the VBP geographic range.

Our estimated VBP range exceeds the area of past Araucaria forest mapped by Hueck (1966) and includes vast areas of the Atlantic forest biome that have been cleared. Nonetheless, both vegetation site covariates—Araucaria and Atlantic forest cover—had strong positive effects on site-occupancy probability. The Paraná Pine plays an important role in the VBP natural history, at least in part of its range, offering roost sites (Prestes et al., 2014), nesting cavities (Cockle et al., 2007) and nutrition during the coldest months of the year (Collar et al., 2017; Kilpp et al., 2015; Prestes et al., 2014; Tella et al., 2016). Nevertheless, as Araucaria forests only extend as far north as the Campos do Jordão region, parrots from the northern patch must rely on other plant species to obtain whatever resources their southern counterparts get from the Paraná Pine. Living at a lower latitude, they may also escape the harshness of cold winter weeks, when Araucaria seeds are a unique source of energy for several species of the southern fauna (Dénes et al., 2018). Indeed, Carrara et al. (2008) registered foraging and roosting in different trees between northern and southern locations. Likewise, Cockle et al. (2007), as well as Prestes et al. (2014), document foraging and cavity nesting in non-Araucaria Atlantic Forest trees of the southern part of the range. The effect of altitude on site occupancy was smaller and more uncertain than the effects of forest cover, but still indisputably positive. Thus, environmental consequences of altitude are not limiting the VBP distribution.

The increasing availability of citizen-science datasets offers a great opportunity to improve species distribution maps. In our study, eBird, WikiAves and Xeno-canto jointly produced 1.45 times more VBP detections, from samples that covered 45 times more municipalities, than the researcher-led counts. Comparison of the realized geographic range produced by the Full Model (Figure 1b) with equivalent maps produced by separate analysis of each dataset (Figure 2) suggests that the former is more accurate. Even though roost counts had reliable identifications based on the most standardized samples in our data, analysis of roost count data alone produces severe overestimation of occupancy in areas where the species is well known to be absent. Such overestimation, and the high predictive power of the roost count's Model 9, stem from the deliberate sampling bias of roost counting, which is targeted to sites where the species is known to be present. Conversely, Xeno-canto data underestimate occupancy in places where the species was recorded by other datasets. Analysed in isolation, eBird data miss information about the northern part of the VBP distribution; WikiAves, in turn, misses the presence of the species in Paraguay altogether, because it only accepts records from Brazil. The assessment of predictive accuracy enabled us to measure the contribution of each dataset for the final estimates. Excluding one dataset at a time from the analyses, or analysing only one dataset at a time, resulted in worse prediction by the truncated analyses than by the joint analysis of all datasets. Only three out of sixteen possible comparisons resulted in lower deviance for the truncated data; all three corresponding to prediction of roost count or Xeno-canto data, the smallest of the four datasets (Table 2). Exclusion of WikiAves data had the highest impact on predictive power, increasing Deviance for the other datasets between 4% and 25%. WikiAves still lacks an automated data download tool, but it is currently the best source of bird species distribution information in Brazil because of its high coverage and number of records. Xeno-canto has the fewest records and smallest spatial coverage, but it still produced a measurable improvement of predictive power when added to the other datasets. Roost counts and eBird had the least consistent impact on prediction power but still produced an average decrease in deviance across datasets. These two datasets also contributed with sampling replication, essential for the quantification of false negative results. While there are limits to the usefulness of data integration (Simmonds et al., 2020), in our case, integration clearly improved the fit of models, suggesting that different datasets are capturing similar realities of parrot distribution; otherwise, their combination should make it more, not less difficult to predict excluded data.

Comparisons across datasets were only possible thanks to a methodology that explicitly accounts for differences in data collection among data sources. Model 3 (Tables 2 and S1), which did not account for the variation of detection probability with respect to effort covariates, consistently showed the largest increase in total deviance relative to the Full Model. Exclusion of the occupancy covariates (Model 2) and the spatial autocorrelation component (Model 4) caused an intermediate but measurable decrease in predictive power. The effect of spatial autocorrelation on deviance signals a spatially structured geographic distribution. Such residual structure was evidently not captured by the occupancy covariates in our models. It remains evident after our accounting of environmental factors, either due to endogenous movement of animals between adjacent sites irrespective of the local environment, or due to exogenous environmental factors that are themselves spatially structured and are missing from, or mis-specified in our models (Legendre, 1993). Further interpretation of the spatial structure should clarify the relative importance of endogenous versus exogenous processes, but for now we emphasize that residual structure is still present and should be accounted for in a distribution map of the species. Neglecting spatial contagion easily leads to biased parameter estimates, potentially resulting in erroneous maps (Guélat & Kéry, 2018; Johnson et al., 2013).

The term “citizen science” covers a wide variety of collaborative arrangements that involve people from outside the scientific community in scientific research (Heigl et al., 2019; Tulloch, 2013; Wiggins & Crowston, 2011). When it comes to collaborative recording of wildlife sightings, however, most citizen-science initiatives compile presence-only information from opportunistic samples. Our analysis employs presence–absence (roost counts, eBird) and presence-only (WikiAves, Xeno-canto) data, as well as a planned survey (roost counts) and opportunistic sampling (WikiAves, eBird, Xeno-canto). While integrating planned-survey with opportunistic sampling data, we account for spatial bias in citizen-science data via estimation of effort per sample, based on covariates obtained from the citizen-science datasets themselves. This approach is synthesized in Equations (3) and (4a–c), which express detection probability conditional on species presence. Other studies develop models with more explicit descriptions of the complex variation of sampling effort that is characteristic of citizen-science datasets (e.g. August et al., 2020; Johnston et al., 2021). We opted for a more general approach that, for example, carries no information about individual observer behaviour. There certainly are biases that were not or cannot be accounted for within our approach, especially when analysing one dataset at a time. Nonetheless, our integration of four datasets did increase spatial cover (relative to each dataset) and captured the substantial importance of accounting for spatial bias in sampling effort. Total deviance more than doubled when effort covariates were removed from the analysis, but it increased only up to 7% (for the eBird data) when we removed the planned-survey roost count data. These results are in agreement with the usefulness of integrating citizen-science with planned-survey data without any particular data source being regarded as a gold standard. They also strengthen our confidence in the contribution of large, multi-species citizen-science datasets for improving knowledge about species distributions.

ACKNOWLEDGEMENTS

This paper owes a great deal to Reinaldo Guedes who volunteered his free time over the last twelve years to developing and administrating WikiAves, the most successful citizen-science initiative in Brazil. Roost counts were supported in large part by *Projeto Charão*, in Brazil, *Proyecto Selva Pino Paraná*, in Argentina, and *Guyra Paraguay*, in Paraguay. Last, but not least, we are indebted to the thousands of bird observers who uploaded photographs, audiorecordings and birding lists to eBird, WikiAves and Xeno-canto.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/ddi.13416.

DATA AVAILABILITY STATEMENT

The R codes and datasets used for the analyses are openly available at GitHub in the address: https://github.com/vivizulian/DataIntegrationModels.

ORCID

Viviane Zulian https://orcid.org/0000-0003-1723-6995
David A. W. Miller https://orcid.org/0000-0002-3011-3677
Gonçalo Ferraz https://orcid.org/0000-0001-8748-0462

REFERENCES

Altwegg, R., & Nichols, J. D. (2018). Occupancy models for citizen-science data. *Methods in Ecology and Evolution*, *10*(1), 8–21. https://doi.org/10.1111/2041-210X.13090

Araya-Salas, M., & Smith-Vidaurre, G. (2017). warbleR: an r package to streamline analysis of animal acoustic signals. *Methods in Ecology and Evolution*, *8*(2), 184–191. http://dx.doi.org/10.1111/2041-210x.12624

August, T., Fox, R., Roy, D. B., & Pocock, M. J. O. (2020). Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. *Scientific Reports*, *10*, 11009. https://doi.org/10.1038/s41598-020-67658-3

BirdLife International (2017). *Amazona vinacea. The IUCN Red List of Threatened Species*. http://dx.doi.org/ https://doi.org/10.2305/IUCN.UK.2016-3.RLTS.T22686374A93109194.en

BirdLife International, & Handbook of the Birds of the World (2016). *Amazona vinacea* (3rd ed.) [Map]. The IUCN Red List of Threatened Species. http://www.iucnredlist.org

Carrara, L. A., Faria, L. C., Matos, J. R., & de Antas, P. T. Z. (2008). Papagaio-de-peito-roxo *Amazona vinacea* (Kuhl) (Aves: Psittacidae) no norte do Espírito Santo: Redescoberta e conservação. *Revista Brasileira de Zoologia*, *25*(1), 154–158. https://doi.org/10.1590/S0101-81752008000100021

Cockle, K., Capuzzi, G., Bodrati, A., Clay, R., del Castillo, H., Velázquez, M., Areta, J. I., Fariña, N., & Fariña, R. (2007). Distribution, abundance, and conservation of Vinaceous Amazons (*Amazona vinacea*) in Argentina and Paraguay. *Journal of Field Ornithology*, *78*(1), 21–39. https://doi.org/10.1111/j.1557-9263.2006.00082.x

Cockle, K. L., Ibarra, J. T., Altamirano, T. A., & Martin, K. (2019). Interspecific networks of cavity-nesting vertebrates reveal a critical role of broadleaf trees in endangered Araucaria mixed forests of South America. *Biodiversity and Conservation*, *28*(12), 3371–3386. https://doi.org/10.1007/s10531-019-01826-4

Collar, N., Boesman, P., & Juana, E. (2017). Vinaceous-breasted Amazon (*Amazona vinacea*). In J. del Hoyo, A. Elliott, J. Sargatal, D. A. Christie, & E. de Juana (Eds.), *Handbook of the birds of the world alive*. Lynx Edicions. http://www.hbw.com/node/54755

Coron, C., Calenge, C., Giraud, C., & Julliard, R. (2018). Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. *Environmental and Ecological Statistics*, *25*(1), 71–93. https://doi.org/10.1007/s10651-018-0398-2

Dénes, F. V., Tella, J. L., Zulian, V., Prestes, N. P., Martínez, J., & Hiraldo, F. (2018). Combined impacts of multiple non-native mammals on two life stages of a critically endangered Neotropical tree. *Biological Invasions*, *20*, 3055–3068. https://doi.org/10.1007/s10530-018-1758-4

DIVA-GIS (2018). *Elevation*. Free spatial data. https://diva-gis.org/gdata

Dorazio, R. M. (2014). Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. *Global Ecology and Biogeography*, *23*(12), 1472–1484. https://doi.org/10.1111/geb.12216

eBird (2019). *EBird*. An online database of bird distribution and abundance. https://ebird.org/home

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, *6*(4), 424–438. https://doi.org/10.1111/2041-210X.12242

Fletcher, R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., & Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, *100*, e02710.

Giraud, C., Calenge, C., Coron, C., & Julliard, R. (2016). Capitalizing on opportunistic data for monitoring relative abundances of species. *Biometrics*, *72*(2), 649–658. https://doi.org/10.1111/biom.12431

Gregory, R. D., van Strien, A., Vorisek, P., Gmelig Meyling, A. W., Noble, D. G., Foppen, R. P. B., & Gibbons, D. W. (2005). Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1454), 269–288. https://doi.org/10.1098/rstb.2004.1602

Guélat, J., & Kéry, M. (2018). Effects of spatial autocorrelation and imperfect detection on species distribution models. *Methods in Ecology and Evolution*, *9*(6), 1614–1625. https://doi.org/10.1111/2041-210X.12983

Heigl, F., Kieslinger, B., Paul, K. T., Uhlik, J., & Dörler, D. (2019). Opinion: Toward an international definition of citizen science. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(17), 8089–8092. https://doi.org/10.1073/pnas.1903393116

Hill, M. O. (2012). Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in Ecology and Evolution*, *3*(1), 195–205. https://doi.org/10.1111/j.2041-210X.2011.00146.x

Hooten, M. B., & Hobbs, N. T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, *85*(1), 3–28. https://doi.org/10.1890/14-0661.1

Hudson, M.-A.-R., Francis, C. M., Campbell, K. J., Downes, C. M., Smith, A. C., & Pardieck, K. L. (2017). The role of the North American Breeding Bird Survey in conservation. *The Condor*, *119*(3), 526–545. https://doi.org/10.1650/CONDOR-17-62.1

Hueck, K. (1966). *Die Wälder Südamerikas*. Ökologie, Zusammensetzung un wirtschaftliche Bedeutung.

Isaac, N. J. B., Jarzyna, M. A., Keil, P., Dambly, L. I., Boersch-Supan, P. H., Browning, E., Freeman, S. N., Golding, N., Guillera-Arroita, G., Henrys, P. A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O. L., Schmucki, R., Simmonds, E. G., & O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, *35*(1), 56–67. https://doi.org/10.1016/j.tree.2019.08.006

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution*, *27*(3), 151–159. https://doi.org/10.1016/j.tree.2011.09.007

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C., & Pond, B. A. (2013). Spatial occupancy models for large datasets. *Ecology*, *94*(4), 801–808. https://doi.org/10.1890/12-0564.1

Johnston, A., Hochachka, W. M., Strimas-Mackey, M., Gutierrez, V. R., Robinson, O. J., Miller, E. T., Auer, T., Kelling, S. T., & Fink, D. (2021). Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions. *Diversity and Distributions*, *27*(7), 1265–1277. https://doi.org/10.1111/ddi.13271

Kéry, M. (2011). Towards the modelling of true species distributions: Commentary. *Journal of Biogeography*, *38*(4), 617–618. https://doi.org/10.1111/j.1365-2699.2011.02487.x

Kéry, M., Gardner, B., & Monnerat, C. (2010). Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, *37*, 1851–1862. https://doi.org/10.1111/j.1365-2699.2010.02345.x

Kilpp, J. C., Prestes, N. P., Pizzol, G. E. D., & Martinez, J. (2015). Dieta alimentar de *Amazona vinacea* no sul e sudeste de Santa Catarina, Brasil. *Atualidades Ornitológicas*, *183*, 6.

La Sorte, F. A., & Somveille, M. (2020). Survey completeness of a global citizen-science database of bird occurrence. *Ecography*, *43*, 34–43. https://doi.org/10.1111/ecog.04632

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology*, *74*(6), 1659–1673. https://doi.org/10.2307/1939924

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). *Statistics and Computing*, *10*(4), 325–337. http://dx.doi.org/10.1023/a:1008929526011

Mace, G. M., Collar, N. J., Gaston, K. J., Hilton-Taylor, C., AkçAkaya, H. R., Leader-Williams, N., Milner-Gulland, E. J., & Stuart, S. N. (2008). Quantification of extinction risk: IUCN's system for classifying threatened species. *Conservation Biology*, *22*(6), 1424–1442. https://doi.org/10.1111/j.1523-1739.2008.01044.x

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, *83*(8), 2248–2255.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., & Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, *10*(1), 22–37. https://doi.org/10.1111/2041-210X.13110

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, *51*(11), 933–938. http://dx.doi.org/10.1641/0006-3568(2001)051[0933:teotwa]2.0.co;2

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., & Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, *98*(3), 840–850. https://doi.org/10.1002/ecy.1710

Peel, S. L., Hill, N. A., Foster, S. D., Wotherspoon, S. J., Ghiglione, C., & Schiaparelli, S. (2019). Reliable species distributions are obtainable with sparse, patchy and biased data by leveraging over species and data types. *Methods in Ecology and Evolution*, *10*(7), 1002–1014. https://doi.org/10.1111/2041-210X.13196

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., & Ferrier, S. (2009). Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. *Ecological Applications*, *19*(1), 181–197. https://doi.org/10.1890/07-2153.1

Prestes, N. P., Martinez, J., Kilpp, J. C., Batistela, T., Turkievicz, A., Rezende, É., & Gaboardi, V. T. R. (2014). Ecologia e conservação de *Amazona vinacea* em áreas simpátricas com *Amazona pretrei*. *Ornithologia*, *6*(2), 109–120.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., & Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography*, *35*(2), 211–226. https://doi.org/10.1177/0309133311399491

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B., & O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*, *43*(10), 1413–1422. https://doi.org/10.1111/ecog.05146

Stauffer, G. E., Miller, D. A. W., Williams, L. M., & Brown, J. (2018). Ruffed grouse population declines after introduction of West Nile virus. *The Journal of Wildlife Management*, *82*(1), 165–172. https://doi.org/10.1002/jwmg.21347

Sullivan, B. L., Phillips, T., Dayer, A. A., Wood, C. L., Farnsworth, A., Iliff, M. J., Davies, I. J., Wiggins, A., Fink, D., Hochachka, W. M., Rodewald, A. D., Rosenberg, K. V., Bonney, R., & Kelling, S. (2017). Using open access observational data for conservation action: A case study for birds. *Biological Conservation*, *208*, 5–14. https://doi.org/10.1016/j.biocon.2016.04.031

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Tella, J. L., Dénes, F. V., Zulian, V., Prestes, N. P., Martínez, J., Blanco, G., & Hiraldo, F. (2016). Endangered plant-parrot mutualisms: Seed tolerance to predation makes parrots pervasive dispersers of the Parana pine. *Scientific Reports*, *6*, 31709. https://doi.org/10.1038/srep31709

Tulloch, A. I. T., Possingham, H. P., Joseph, L. N., Szabo, J., & Martin, T. G. (2013). Realising the full potential of citizen science monitoring programs. *Biological Conservation*, *165*, 128–138. https://doi.org/10.1016/j.biocon.2013.05.025

Wiggins, A., & Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. In *2011 44th Hawaii International Conference on System Sciences* (pp. 1–10).

WikiAves (2019). *WikiAves, a Enciclopédia das Aves do Brasil*. http://www.wikiaves.com.br

Xeno-canto (2019). *Xeno-Canto: Bird Sounds from around the World*. https://www.xeno-canto.org/

Zulian, V., Müller, E. S., Cockle, K. L., Lesterhuis, A., Tomasi Júnior, R., Prestes, N. P., Martinez, J., Kéry, M., & Ferraz, G. (2020). Addressing multiple sources of uncertainty in the estimation of global parrot abundance from roost counts: A case study with the Vinaceous-breasted Parrot (*Amazona vinacea*). *Biological Conservation*, *248*, 108672. https://doi.org/10.1016/j.biocon.2020.108672

BIOSKETCHES

Viviane Zulian is a PhD candidate at Universidade Federal do Rio Grande do Sul, Brazil. Her research is focused on species distribution models and abundance estimates with data integration to better understand the population biology of species of high conservation importance. More information at: http://ferrazlab.org/.

David A. W. Miller is an Associate Professor at Pennsylvania State University, United States. His research focuses on matching quantitative tools and ecological understanding to better understand how species will respond to environmental change. More information at: https://www.appliedpopecol.org/.

Gonçalo Ferraz is an Associate Professor at Universidade Federal do Rio Grande do Sul, Brazil. His current work centers on demography of non-human populations, with an emphasis on studying animal population responses to landscape change. More information at: http://ferrazlab.org/

Author contributions: VZ: Conceptualization, investigation, data curation, formal analysis, funding acquisition, project administration, writing. DAWM: Metodology, formal analysis, funding acquisition, supervision, writing. GF: Conceptualization, investigation, resources, funding acquisition, project administration, supervision, writing.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zulian, V., Miller, D. A. W., & Ferraz, G. (2021). Integrating citizen-science and planned-survey data improves species distribution estimates. *Diversity and Distributions*, 27, 2498–2509. https://doi.org/10.1111/ddi.13416