

**Predicting regional densities from bird occurrence data: validation and effects of
species traits in a Macaronesian Island**

Luis M. Carrascal^{1*}, Pedro Aragón¹, David Palomino², Jorge M. Lobo¹

¹ *Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales, CSIC. C/ José Gutiérrez Abascal 2, 28006 Madrid, Spain.*

² *Wildlife Consultor, Guadarrama, Spain.*

*Correspondence: Luis M. Carrascal, Department of Biogeography and Global Change, Museo Nacional de Ciencias Naturales, CSIC. C/ José Gutiérrez Abascal 2, 28006 Madrid, Spain. E-mail: lmcarrascal@mncn.csic.es

Running title: Predicting regional bird densities from occurrence data

1 **ABSTRACT**

2 **Aim** Quantifying species abundances is costly, especially when many species are
3 involved. To overcome this problem several studies have predicted local abundances (at
4 the sample unit level) from species occurrence distribution models (SODM), with
5 differences in predictive performance among studies. Surprisingly, the ability of SODM
6 to predict regional abundances of an entire area of interest has never been tested, despite
7 the fact that it is an essential parameter for species conservation and management. We
8 tested whether local and regional abundances of 21 terrestrial bird species could be
9 predicted from SODMs in an exhaustively surveyed island, and examined the variation
10 explained by species-specific traits.

11 **Location** La Palma Island, Canary Islands.

12 **Methods** We firstly assessed two types of algorithms representing the two main
13 families of SODMs. We built models using presence/absence (with boosted
14 classification trees) and presence/background (with MaxEnt) data as a function of
15 relevant environmental predictors, and tested their ability to predict the observed local
16 abundances. The predicted probabilities of occurrence (P_i) were translated into animal
17 numbers (n') using the revisited equation $n_i' = - \ln (1-P_i)$, and we obtained regional
18 abundances (for the whole island).

19 **Results** Predictive ability of presence/absence models was superior than that of MaxEnt.
20 At the regional level, the observed average densities of all species were highly
21 predictable from occurrence probabilities ($R^2=93.5\%$), without overall overestimation or
22 underestimation. Interspecific variation in the accuracy of predicted regional density
23 was largely explained ($R^2=73\%$); with habitat breadth and variation in local abundance
24 being the traits of greatest importance.

25 **Main conclusions** Despite uncertainties associated with local predictions and the
26 idiosyncrasies of each species, our procedures enabled us to predict regional abundances
27 in an unbiased way. Our approach provides a cost-effective tool when a large number of
28 species is involved. Furthermore, the influence of species-specific traits on the
29 prediction accuracy provides insights into sampling designs for focal species.

30

31 **Keywords**

32 Biodiversity monitoring, species abundance, species distribution modelling, boosted
33 classification trees, MaxEnt, birds, island biogeography.

34

35 **INTRODUCTION**

36 Organism abundance and richness are recognised as two of the most important
37 components of biological diversity. Measures of species abundance for biodiversity
38 assessments provide useful information, from aspects of population dynamics and biotic
39 interactions to ecosystem functioning (e.g. Estes *et al.*, 1998; Yamamoto *et al.*, 2007).
40 Moreover, human-mediated changes in biodiversity are detected more quickly using
41 abundance measurements than accounting for other biodiversity components (Chapin *et*
42 *al.*, 2000). However, quantifying species abundances is challenging because it is costly
43 in terms of time, and human and economic resources. In contrast, the number of studies
44 analyzing variables derived from species presences is becoming disproportionately
45 higher than those making use of measures of abundance (Guisan & Thuiller, 2005;
46 Rodríguez *et al.*, 2007). To overcome this problem several studies aimed to predict
47 species abundance from Species Occurrence Distribution Models (SODMs; e.g.,
48 Conlisk *et al.*, 2009 and references therein). Thus, linking successfully distributional
49 occurrence data with abundance through relevant factors should provide a useful tool
50 since species presence data are easier to obtain, which opens the possibility of
51 coordinating volunteer programs in field survey designs.

52 However, the extent to which SODM outputs are able to precisely predict local
53 abundances or densities remains controversial (Pearce & Ferrier, 2001; Nielsen *et al.*,
54 2005; Jiménez-Valverde *et al.*, 2009; Estrada & Arroyo, 2012; Van Couwenberghe *et*
55 *al.*, 2013; Thuiller *et al.*, 2014; Bean *et al.*, 2014; Yañez *et al.*, 2014; Russell *et al.*,
56 2015). Despite potential limitations of SODMs to account for local abundances (Pearce
57 & Ferrier, 2001; Nielsen *et al.*, 2005), their ability to predict the total count of
58 individuals in the whole study area (hereafter regional abundance) is unknown. When

59 the central limit theorem holds (Grinstead & Snell, 1997) local overpredictions and
60 underpredictions can be counteracted because they are randomly, and equally
61 distributed. In this case, regional abundances could be accurately predicted even in
62 cases of moderate ability of SODMs to predict local abundances.

63 Among the SODM types there are also differences regarding the difficulty of
64 obtaining distributional data, mainly depending on whether they makes use of
65 presence/absence or only true presences. Recording presence/absence data requires
66 greater survey effort than presence-only data since uncertainties associated with
67 absences are greater (Jiménez-Valverde *et al.*, 2008). Moreover, part of the variability
68 regarding the ability of SODM to predict abundances might be influenced by whether or
69 not true absences are assumed (Nielsen *et al.*, 2005, VanDerWal *et al.*, 2009).
70 Therefore, elucidating the extent to which obtaining absence data merits additional
71 survey efforts needs to solve the trade-off between feasibility and effectiveness when
72 predicting abundances from SODMs. Comparisons between SODM outputs,
73 considering they include or not reliable absence data, may help to examine the
74 variability in the relationships between probability/suitability values and abundance
75 estimations.

76 In the same way, species-specific traits linked with natural history are also sources
77 of variability in model accuracy to predict species' distributions and abundances. This
78 interspecific variability limits the predictive power of modelling exercises, a limitation
79 that cannot be always overcome by mere statistical refinements (Seoane *et al.*, 2005).
80 Several studies have shown that ecological and natural history traits of species may
81 predict the errors in SODMs (Boone & Krohn, 1999; Kadmon *et al.*, 2003; Carrascal *et*
82 *al.*, 2006). For example, modelling success is inversely related to spatial variability

83 (mobility and nomadism) and niche breadth, although the observed patterns are not
84 consistent across all biological groups (Pearce & Ferrier, 2000; Pearce *et al.*, 2001).
85 Similar species-specific variations in modelling success have been found considering
86 the positive effects of commonness, abundance and detectability (Boone & Krohn,
87 1999; Kadmon *et al.*, 2003). Therefore, the analysis of the association between species'
88 biological traits and model accuracy is useful because if we know the effect of specific
89 traits on modelling results, we can improve the sampling design for multi-species
90 studies (Seoane *et al.*, 2005).

91 In this study we examined whether local and regional abundances of a group of
92 terrestrial bird species can be predicted from SODMs in La Palma, a Macaronesian
93 island in the Canary archipelago. An exhaustive field survey was carried out to record
94 presence/absence data and abundances of twenty one bird species throughout a
95 representative sample of transects encompassing the spatial and environmental range of
96 the island. First, for each species we built distribution models for La Palma Island using
97 presence/absence or presence/background data as a function of relevant environmental
98 predictors. Second, we compared the ability of these two types of models to predict the
99 observed local abundances of the studied bird species. Third, we used the type of
100 SODM that derived better local predictions to obtain estimations of regional
101 abundances. For this purpose SODM outputs were converted to abundances by means
102 of a previously proposed and well founded procedure in the early seventies, the
103 binomial sampling to estimate average densities (Gerrard & Chiang, 1970). This
104 conversion has been rarely applied for organisms other than arthropods but merits
105 further evaluation, because it does not require complex parameterizations. Our
106 predictions of regional abundances were then evaluated using total number of birds

107 recorded in the field. Fourth, we performed an analysis including all species to elucidate
108 species-specific traits that can potentially explain an inter-specific variation in the
109 regional abundance estimations. To our knowledge this is the first time that the ability
110 of SODM to predict species regional abundances has been examined.

111

112 **METHODS**

113 **Study area**

114 The study area is located in La Palma (28°42' N, 17°50' W; 706 km²) a young (1-2 My)
115 oceanic island of the Canary archipelago located 417 km from the African coast. It is a
116 high island (2,426 m a.s.l.), with extensive areas with annual precipitation higher than
117 600 mm, and with a widespread representation of native shrublands and pine and
118 evergreen '*laurisilva*' forests (although natural cover has been much reduced since
119 humans occupied the islands: de Nascimento *et al.*, 2009). A considerable proportion of
120 island area below 1,100 m a.s.l. has been highly transformed by agricultural activities
121 and urban sprawl. See Juan *et al.* (2000) and Fernández-Palacios and Martín-Esquivel
122 (2001) for more details on island characteristics.

123

124 **Abundance estimations**

125 Bird censuses, devoted to record presence/absence and abundance data, were carried out
126 during the breeding season (April 2007). The survey method was the line transect,
127 frequently used in extensive assessments of abundance, general distribution patterns and
128 habitat preferences of birds (Bibby *et al.*, 2000). Fieldwork was designed as a broad-
129 scale sampling for land birds. Thus, censuses were carried out across the whole island in

130 an attempt to sample the total range of vegetation types, land-use types and degrees of
131 slope (see Seoane *et al.*, 2011 for a detailed description on the sampling protocol). We
132 recorded all birds heard or seen without a detection limit distance, distinguishing
133 between those registered inside and outside the survey belt of 25 m at each side of the
134 progression line, in order to estimate a measurement of detectability (see below). All
135 censuses were carried out on windless and rainless days, at a low speed (ca. 1–3 km/h),
136 early in the morning (7:00–11:00 GMT) and late in the evening (16:00–17:30 GMT).

137 Transects were 0.5-km sample units of homogeneous habitat structure. They were
138 measured and georeferenced with portable GPS (precision of ± 2 m by means of the
139 average location function). The starting point of transects were randomly determined
140 and then the rest of 0.5-km samples were performed one after the other ($n = 437$
141 transects). We feel confident in assuming that these transects provide a representative
142 sample of broad habitat classes present in La Palma island (see Fig. 1).

143 A surrogate of detectability was built as the ratio between the birds belonging to
144 each species observed inside the transect belt of 25 m at both sides of the observer, and
145 the total number of birds detected (i.e., the ratio p of main belt to total belt
146 observations). This index reflects important species characteristics related to the
147 interaction with the observer, such as song or call intensity and audibility,
148 conspicuousness and mobility (Järvinen & Väisänen, 1975). Density estimations,
149 accounting for species-specific detectability, were calculated using the following
150 equation (Järvinen and Väisänen, 1975; Järvinen, 1978):

$$151 \quad D = (N * k) / L$$

$$152 \quad \text{being } k = (1 - (1 - p)^{0.5}) / 0.025$$

153 where D is the density in birds/km², N is the number of detected birds, k is a
154 detectability coefficient, L is the transect length in km, and p is the ratio of main belt to
155 total belt observations of each bird species (0.025 is transect belt of 25 m expressed in
156 km). This is a convenient approach to account for differences in detection probabilities
157 among species in highly vegetated environments, when measuring exact distances to
158 each individual bird is not feasible because devices such as laser range-finders cannot
159 be applied precisely to birds heard but not seen in densely vegetated habitats.

160

161 **Environmental predictors**

162 Models were built with environmental predictors that have been shown to play a role in
163 shaping the distributions and/or abundances of birds at our spatial resolution, such as
164 those expressing vegetation structure, primary productivity, topography and human
165 impact (Seoane et al. 2005; Mcfarland *et al.* 2012). The vegetation structure categories
166 were assigned to each transect based on an existing map of plant communities in the
167 Canary Islands (Del Arco *et al.*, 2003). The following ten broad classes were identified:
168 volcanic fields (*'malpaises'*), pasturelands, *Euphorbia* shrublands, scrublands, tall
169 heathlands (*'fayal-brezal'*), evergreen forests (*'laurisilva'*), pine forests of *Pinus*
170 *canariensis*, rocky slopes with scattered plants (*'cerrillar'*), agricultural habitats and
171 urban areas. For each transect, we also measured the minimum distance to these habitats
172 using ArcGis. The altitude, cardinal direction and the terrain slope in the centre of each
173 transect were obtained from a digital model (100 m spatial resolution). As an indicator
174 of primary productivity, we quantified photosynthetic activity using a normalized
175 difference vegetation index (NDVI). Raw NDVI data were ten-day synthesis obtained
176 from the sensor VEGETATION onboard the SPOT satellite, averaging data for March

177 to June of the sampling year, and discarding cloudy pixels. Additionally, in order to
178 increase prediction capacity of models we also used UTM latitude and longitude in
179 meters to absorb potential remaining spatial variation not explained by vegetation and
180 topography. All these data were also obtained for the centre of all UTM 500 m x 500 m
181 squares of La Palma Island ($n = 3263$).

182

183 **Species occurrence distribution models**

184 Boosted Classification Trees (BCT) were employed to assess the probability of
185 occurrence (presence-1/absence-0) of each species in the sample of 437 transects of 0.5-
186 km using the 16 formerly mentioned predictor variables. The BCT algorithm builds a
187 number of regression trees (typically hundreds) in a stage-wise fashion on randomly
188 selected subsets of data and combines them to improve predictive performance (see for
189 details: De'Ath, 2007; Elith *et al.*, 2008). We used a five-fold approach in order to test
190 the accuracy of predictions of BCT models. As outputs from boosting are not well
191 calibrated, posterior probabilities predictions of BCT models were calibrated applying a
192 logit function to transform boosting predictions with a sigmoid function (Niculescu-
193 Mizil & Caruana, 2007).

194 To compare BCT predictions with those provided when accurate absence data does
195 not exist, occurrences for each species were also modelled using the MaxEnt algorithm
196 (Phillips *et al.*, 2006; Phillips & Dudik, 2008). We selected this modelling technique
197 because it is a widely used procedure when only presences are available, and is also a
198 machine learning method. As in the classic resource selection functions of use-
199 availability designs (Manly *et al.*, 2002), MaxEnt generates suitability outputs from
200 presence data and a pool of background absences selected at random from the study area

201 using a maximum entropy approach (Phillips *et al.*, 2006; Pearce & Boyce, 2006). In
202 our case, these background absences were selected out of the UTM squares in which
203 transects occur and equal in numbers to those used in BCT models for each species
204 (range: 44 to 420; average=335). This approach has been chosen to i) avoid the use of
205 true absences as background absences, and ii) to ease the comparison of model outputs
206 using an identical number of true absences (in BCT) and background absences (in
207 MaxEnt). Moreover, for MaxEnt the 5-fold data split into training and testing subsets
208 was the same as for the BCT models within each species. Thus, our data arrangement
209 will enable more direct inferences regarding the use of reliable absences in models
210 while keeping other sources of inter-model variability as fixed as possible.

211 The discrimination ability of BCT and MaxEnt models to predict each species'
212 distribution was compared through the area under the curve (AUC) of the receiver
213 operating characteristic (ROC) plot of sensitivity against 1-specificity (Fielding & Bell,
214 1997). AUC values should not be interpreted uncritically, and one of the major misuses
215 is relying on absolute values to compare among species with different prevalences
216 (Lobo *et al.*, 2008). In spite of this, its use in a relative way may be useful to compare
217 among modelling techniques within species with identical prevalences (Aragón &
218 Sánchez-Fernández, 2013).

219

220 **Predicting local and regional abundances from occurrence distribution models**

221 Firstly, we aimed to assess the general ability of presence/absence models (BCT) and
222 presence/background absence models (MaxEnt) to predict local abundances at the
223 transect level. For this purpose, we estimated separately for each species the Pearson
224 correlations of the relationships between observed abundances in transects and SODM

225 outputs (the predicted habitat suitabilities using MaxEnt or probabilities of occurrence
226 using BCT). Sequential Bonferroni adjustment was applied to these analyses to control
227 for type I errors (Benjamini & Hochberg 1995). Then, we used a paired *t*-test to
228 compare between the Pearson correlation coefficients obtained for each species
229 separately with BCT outputs and those obtained with MaxEnt outputs. In addition, we
230 assessed the degree of triangularity in the relationships between observed local
231 abundances and model outputs separately for each species (see Appendix S1 in
232 Supporting Information).

233 As use-availability models, such as MaxEnt, are unable to predict the probability of
234 occurrence (Hastie & Fithiam, 2013), BCT probabilities of occurrence were
235 subsequently used to obtain regional abundance estimations. For this purpose, we firstly
236 converted the probabilities of occurrence to bird numbers applying a procedure that has
237 been shown to be appropriate for the case of outputs from presence/absence models.
238 The predicted probabilities of occurrence for each transect (P_i) derived from BCT
239 models were converted to predicted bird numbers for each species (n_i') using the
240 following expression under the assumption of random distributions with Poisson
241 distributed populations (Gerrard & Chiang 1970, Gerrard & Cook 1982):

$$242 \quad n_i' = - \ln (1-P_i)$$

243 The summation of the predicted n_i' figures for each species ($\sum n_i'$) was used to
244 estimate its resemblance to the true number of birds counted in the whole sample ($\sum n_i$)
245 of 437 transects that equal 218.5 km. These numbers were transformed in regional
246 densities (DENREG; birds / km²) considering the above mentioned formula by Järvinen
247 and Väisänen (1975). Finally, we performed a Pearson correlation to estimate the
248 relationship between predicted and observed regional densities for the 21 bird species

249 recorded. Additionally, we used t-tests to assess whether this predicted regression line
250 deviated significantly from the equality between the observed and predicted densities.

251

252 **Interspecific variation in prediction accuracy of regional density**

253 Interspecific variation in the prediction accuracy of densities using BCT models was
254 characterized by calculating the percentage difference between predicted and observed
255 regional densities in relation to observed regional density (hereafter % change). The
256 thus obtained % change was then related to several autoecological traits of the species:
257 species prevalence in the whole sample of transects (range: 0.04 – 0.90), coefficient of
258 variation in bird numbers when each species was present (30 – 167 %), a surrogate of
259 detectability (as measured by the ratio p of main belt to total belt observations of each
260 bird species -see above-; range: 0.22 – 0.89), body mass (5.8 – 480 g; obtained from
261 Perrins 1998 as the mean weight of males and females, or as the average value of body
262 weight range in spring and summer), and habitat breadth (0.14 – 0.74) and ecological
263 density (3.5 – 248.1 birds/km²) estimated for the most preferred habitat (these two last
264 variables obtained from Appendix B of Seoane *et al.*, 2011).

265 All possible subsets of the predictors using General Linear Models were estimated
266 (64 models), and were compared with Akaike's second-order AIC corrected for small
267 sample sizes (AICc; Burnham & Anderson, 2002) to assess their weights of evidence.
268 The strength of evidence of models was obtained using weights (W_i) derived from *AICc*
269 figures, using all possible models (R package *glmulti*). Parameter estimates
270 (standardised regression coefficients, β ; R^2 of models) were averaged using model
271 weights (W_i ; Arnold, 2010).

272

273

274

275 **RESULTS**

276 **Accuracy of species distribution models**

277 Since AUC values were obtained by five-fold cross-validation, predictions of bird
278 distributions from both BCT and MaxEnt models can be considered excellent or good
279 according to usual performance criteria (Swets 1988) ($n = 21$; mean AUCs \pm SD: BCT
280 = 0.835 ± 0.118 ; MaxEnt = 0.792 ± 0.128 ; Table 1). AUCs for BCT and MaxEnt
281 models were significantly and positively correlated (Pearson's correlation: $r = 0.693$; P
282 = 0.0005) although BCT figures were slightly higher than those obtained with MaxEnt
283 (paired t -test = 2.073 , $P = 0.051$, Table 1).

284

285 **Predicting bird local and regional abundances from distribution models**

286 Probabilities of occurrence (from BCT) and habitat suitability values (from MaxEnt)
287 were positively and significantly associated with their corresponding observed
288 abundances for nearly all species using transects as sample units (see Pearson's
289 correlation coefficients in Table 1). The exceptions were *Phylloscopus canariensis* and
290 *Streptopelia turtur* for MaxEnt outputs, where relationships with abundance were not
291 significant after sequential Bonferroni corrections. The strength of association between
292 model predictions and observed abundances was considerably higher for BCT than for
293 MaxEnt models (paired t -test = 10.792 ; $P < 0.001$; $n = 21$ species). On the other hand,
294 the triangular relationship assessed with quantile regressions was always present and
295 was not different between BCT and MaxEnt results (see Appendix S1).

296 At the regional level (i.e., using the whole sample of transects in the island), the
297 observed average densities per species were highly correlated with those predicted by
298 BCT occurrence probabilities (P_i) when converted to regional densities (i.e. $\sum -\ln[1-P_i]$
299 for transects $i = 1$ to $i = 437$; $r = 0.967$; $P \ll 0.001$; Table 1; Fig. 2). Coefficients a and
300 b in the equation $\text{OBSERVED} = a + b \cdot \text{PREDICTED}$ did not significantly differ from
301 zero and one respectively ($a = 3.5$, $\text{SE} = 3.25$; $b = 1.014$, $\text{SE} = 0.061$; $P > 0.2$ in both t-
302 tests; Fig. 2). Therefore, the observed and predicted regional densities are operatively
303 interchangeable. Moreover, % of difference between predicted and observed regional
304 density was close to zero (mean % difference = -0.076 , $\text{SE} = 7.97$). Thus, there was no
305 overall bias toward either overestimation or underestimation of bird abundance at the
306 regional level.

307

308 **Interspecific variation in prediction accuracy of regional density**

309 Interspecific variation in the accuracy of predicted average density at regional scale
310 (i.e., the average density in the whole sample of transects) was explained to a great
311 amount (73% of variance) by a weighted average model. The variability in bird counts
312 when the species was present, habitat breadth, prevalence in the sample of transects and
313 regional maximum density were the most influential variables ($\sum W_i \geq 0.4$; Table 2). The
314 variable most affecting the accuracy of predicted regional density was the variability in
315 bird counts measured by the coefficient of variation ($\text{CV}\%$; $\sum W_i = 1$, with the largest
316 absolute value of the standardized regression coefficient; Fig. 3a). Habitat breadth had
317 also a similarly high importance, although its magnitude effect was lower (β coefficients
318 in Table 4; see Fig. 3b). Summarizing, predicted regional abundance tended to be
319 underestimated in those species which occupy a narrow range of habitats and show a

320 large variability in numbers when present. High prevalence in the sample and high
321 density in the most preferred habitat also tended to underestimate regional estimates.

322

323 **DISCUSSION**

324 In this study we examined the extent to which the continuous predictions obtained from
325 species' presence/absence (probabilities of occurrence from boosted classification trees)
326 or presence/background (suitabilities from MaxEnt) models can predict species
327 abundances, either at local (sampling units) or at a regional level (La Palma Island). To
328 allow comparisons between presence/absence and presences/background models,
329 prevalences and 5-fold partitions were kept identical in both modelling procedures for
330 each species, while the only difference was the use of observed absences *vs.* background
331 random data. Although the accuracy of presence/absence models was only slightly
332 higher than that of presence/background models in predicting the occurrence of species,
333 the ability to predict observed local abundances was clearly superior for
334 presence/absence models using BCT. Since we designed an experimental protocol to
335 rule out the influence of differences in the prevalence of training data, our results reveal
336 that the differences found between MaxEnt and BCT were due to the algorithm used
337 and/or to the nature of the non-presence data (absences/background) independently of
338 the prevalence. Our results are robust since predicted occurrence probabilities derived
339 from SODMs were obtained from five-fold cross-validations with data not used to build
340 models.

341 An important difference between modelling with presence/absence and with
342 presence-only data is that the latter operates with background data (a mixture of
343 unrecorded absences and presences), inflating thus the number of false absences to an

344 unknown degree (Lobo *et al.*, 2008). As a consequence, the use of background data is
345 less appropriate to estimate abundances from occurrence data. Our results support the
346 use of presence/absence sampling protocols to predict animal abundance even at the
347 local scale, although better results were obtained by combining these predictions to infer
348 regional abundances (*i.e.* the total number of individuals per species recorded in the
349 whole sample of line transects carried out in La Palma Island). Despite the fact that the
350 relationships between SODM outputs and observed local abundances tended to be
351 triangular (see Appendix S1), our estimations at the regional level turned out to be
352 highly precise after applying the simplest transformation, assuming Poisson distributed
353 populations, proposed by Gerrard and Chiang (1970). This is clearly shown by the fact
354 that the regression line nearly represents the perfect equivalency between predicted and
355 observed average bird densities in La Palma island (Table 1, Fig. 2).

356 Although several studies have focused on the ability of SODMs to predict local
357 abundances (Nielsen *et al.*, 2005; Seoane *et al.*, 2005; Jiménez-Valverde *et al.*, 2009;
358 Estrada & Arroyo 2012; Thuiller *et al.*, 2014; Bean *et al.*, 2014; Yañez *et al.*, 2014;
359 Russell *et al.*, 2015), showing generally that they only allow for the demarcation of the
360 upper-limit of the observed abundances (e.g. VanDerWal *et al.*, 2009; Tôrres *et al.*,
361 2012), little is known about the usefulness of occurrence data to predict regional
362 abundances (*i.e.*, number of individuals or densities). The advantage of the approach
363 applied here at the regional level is that the same transformation is applied for all
364 species, and hence it can be used as an alternative to specific parameterizations
365 proposed in other studies for each species separately (*e.g.* VanDerWal *et al.*, 2009). We
366 propose that this procedure is especially appropriate and cost-effective when the aim is
367 to infer regional abundances of large sets of species under sampling restrictions, as often

368 occur in biodiversity studies. Thus, our procedure to predict average regional densities
369 can be a powerful tool in cases of biodiversity assessment in poorly known regions or
370 remote areas. Furthermore, we may be interested in examining the potential effect of an
371 ecological perturbation by comparing species abundances in the target area before and
372 after the perturbation occurred, or between the disturbed and other neighbouring areas.
373 In the same vein, this procedure can provide insights in the context of reserve design;
374 comparing predicted regional densities among contiguous areas with different
375 protection status would help to make decisions when reviewing their protection
376 capacity. It is remarkable that studies on reserve design selection are often based on
377 species representation (Araújo *et al.*, 2007), analogous procedures based on
378 probabilities of occurrence are scarce (Cabeza *et al.*, 2004), and there is a general lack
379 of approaches dealing with abundances in many organisms (apart from birds,
380 considering their attractiveness for citizen science projects). The high accuracy of the
381 procedure used here to predict regional densities from SODM outputs with true
382 presence/absences suggests its potential value when working with organisms for which
383 census programs dealing with abundances are not the norm or are not feasible.

384 At the regional scale, we found that the interspecific variation in prediction accuracy
385 of regional abundance can be explained by species-specific traits related to distribution
386 patterns and habitat preferences. This is in line with previous studies showing that
387 autoecological traits may affect model performance in predicting species distributions
388 from observed presences/absences (Hernandez *et al.*, 2006), abundances from observed
389 abundances (Seoane *et al.*, 2005; Carrascal *et al.*, 2006), and abundances from
390 occurrence probabilities (Nielsen *et al.*, 2005; Jiménez-Valverde *et al.*, 2009; Estrada &
391 Arroyo 2012; Russell *et al.*, 2015). Habitat breadth and the coefficient of variation in bird

392 numbers were specific traits with higher relative importance in explaining the
393 interspecific variation in predicting regional densities. Bird species with a greater
394 habitat breadth, such as *Falco tinnunculus* and *Streptopelia turtur*, tended to be
395 overestimated (Fig. 4b, see Appendix S2). Species inhabiting a greater number of
396 habitat types can be associated to a greater range of environmental variation and hence
397 predictions might be closer to the upper part of their potential. It is also plausible that
398 species with broad niches are at lower numbers than the expected potential simply
399 because other biologically relevant factors not included in the models might be also
400 shaping subtle variations in their abundances. Thus, species with larger habitat breadths
401 may be more sensitive to the exclusion of unknown relevant factors in models, which
402 result in a greater mismatch between observed and predicted abundances. Whatever the
403 processes involved, it appears that environmental tolerance governs both species
404 occurrence distributions and abundances, since it has been shown to affect the accuracy
405 of SODM and abundance models (Carrascal *et al.*, 2006; Hernandez *et al.*, 2006;
406 Seoane *et al.*, 2005).

407 Species with higher coefficients of variation of local abundance when present, such
408 as *Pyrrhocorax pyrrhocorax*, *Carduelis cannabina* and *Columba livia* (e.g., from 1 to
409 30 individuals as opposed to ranges of 1 to 3 individuals), tended to be underestimated.
410 The coefficient of variation may be linked to the within-species variation regarding
411 grouping behaviour or environmental fine-grained variables affecting animal abundance
412 not included in the models (e.g., habitat structure, food availability, substrata for
413 nesting). Estrada and Arroyo (2012) found that differences between two harrier species
414 regarding the degree of association between SODM outputs and abundances could be
415 explained by the degree of gregariousness and by the interspecific variation in the use of

416 social information for site selection. Thus, it is possible that the within and among-
417 species variation in grouping behaviour affects abundance predictions intra- and inter-
418 specifically. Finally, our results show that among the species traits considered,
419 detectability had the lowest relative importance in explaining deviations from the
420 observed regional density. In fact, it has been argued that presence/absence models are
421 less affected by this trait than models built with presence-only data (Pearce & Ferrier,
422 2001).

423 To conclude, our results show that when predicting species abundances from
424 occurrence data, presence/absence models outperformed presence/background models.
425 If abundance or density information is essential to advise conservation decisions, such
426 information should not be derived when reliable absences are lacking. The use of
427 presence-only models with background data does not allow good predictions of local
428 abundances. Moreover, the impossibility of estimating the probability of occurrence
429 from these only-presence designs (Hastie & Fithian, 2013) hinders the estimation of
430 abundances by the conversion of probabilities to animal numbers. Our study shows that
431 despite limitations of occurrence binary data (presence/absence) to predict precise local
432 abundances, these local predictions may be combined to predict unbiased average
433 regional abundance. This is because, although accuracies are not similar across species,
434 overestimations and underestimations compensate each other within each species.

435 It is highly surprising that the procedure revisited here designed by Gerrard &
436 Chiang (1970) to convert local probabilities of occurrence into numbers of individuals
437 has rarely been used with vertebrates (but see Tellería & Sáez-Royuela 1986),
438 considering that the accuracy of the predictions are very high as it has been
439 demonstrated in this paper and previously with arthropods (e.g., Gerrard & Chiang

440 1970, Badenhausser *et al.*, 2007, Hall *et al.*, 2007). The only concern is to avoid the
441 “dangerous zone” where the probability of occurrence (P_i) is higher than ca. 0.9. Over
442 this probability, the observed and predicted abundances grow exponentially, so very
443 small changes in P_i generate very large variations in abundance. Therefore, the obvious
444 advice is to define sampling protocols where the size of the sampling unit (i.e., 0.5-km
445 length transects in our study) produces probabilities or frequencies of occurrence below
446 the “saturation point” of 0.9 (see also Gerrard & Chiang 1970). Further studies with
447 heterogeneous taxa, scales and situations will likely reinforce the generality of this
448 procedure.

449 Although obtaining good species’ absences in a random sampling protocol is
450 economically costly and time consuming, the costs associated to measure species’
451 abundances are considerably higher and not always feasible. This study highlights the
452 usefulness of surrogate measures of species abundances derived from distribution
453 models built with presence/absence data. This approach can be a useful tool in applied
454 ecology, especially when working in remote areas, under budget restrictions or with
455 limited qualified personnel. Since the accuracies of predicted regional densities are
456 similar across species, the approach is highly valuable in studies of biodiversity that
457 deal with a large number of species. Moreover, analyses testing the potential influence
458 of species-specific traits on prediction accuracy should be viewed as a valuable
459 complement to gather further insights on the processes involved in the interaction
460 between the sampling method and focus species.

461

462 **ACKNOWLEDGMENTS**

463 This paper is a contribution to the projects CGL2011-28177/BOS and CGL2014-56416-
464 P of the Spanish Ministry of Education and Science and Spanish Ministry of Economy
465 and Competitiveness, respectively. P.A. was supported by a "Ramón y Cajal" contract
466 (RYC-2011-07670) from the Spanish Ministry of Economy and Competitiveness. We
467 thank A. Jiménez-Valverde for helpful comments on the subject, and C. Jasinski for
468 improving the English of the manuscript.

469

470 REFERENCES

- 471 Aragón, P. & Sánchez-Fernández, D. (2013) Can we disentangle predator-prey
472 interactions from species distributions at a macro-scale? A case study with a
473 raptor species. *Oikos*, **122**, 64-72.
- 474 Araújo, M.B., Lobo, J.B. & Moreno, J.C. (2007) The effectiveness of Iberian protected
475 areas in conserving terrestrial biodiversity. *Conservation Biology*, **21**, 1423-
476 1432.
- 477 Arnold, T.W. (2010) Uninformative Parameters and Model Selection Using Akaike's
478 Information Criterion. *Journal of Wildlife Management*, **74**, 1175-1178.
- 479 Badenhausser, I. Amouroux, P. & Bretagnolle, V. (2007) Estimating acridid densities in
480 grassland habitats: a comparison between presence-absence and abundance
481 sampling designs. *Environmental Entomology*, **36**, 1494-1503.
- 482 Bean, W.T., Prugh, L.R., Stafford, R., Butterfield, H.S., Westphal, M. & Brashares, J.S.
483 (2014) Species distribution models of an endangered rodent offer conflicting
484 measures of habitat quality at multiple scales. *Journal of Applied Ecology*, **51**,
485 1116-1125.

- 486 Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical
487 and powerful approach to multiple testing. *Journal Royal Statistical Society Ser.*
488 *B*, **57**, 289–300.
- 489 Bibby, C.J., Burgess, N.D., Hill, D.A. & Mustoe, S.H. (2000) *Bird census techniques*,
490 2nd edn. Academic Press, London.
- 491 Boone, R.B. & Krohn, W.B. (1999) Modeling the occurrence of bird species: are the
492 errors predictable? *Ecological Applications*, **9**, 835-848.
- 493 Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference: a*
494 *practical information–theoretic approach*, 2nd edn. Springer-Verlag, New York.
- 495 Cabeza, M., Araújo, M.B., Wilson, R.J., Thomas, C.D., Cowley, M.J.R. & Moilanen, A.
496 (2004) Combining probabilities of occurrence with spatial reserve design.
497 *Journal of Applied Ecology*, **41**, 252-262.
- 498 Carrascal, L.M., Seoane, J., Palomino, D., Alonso, C.L. & Lobo, J.M. (2006) Species-
499 specific features affect the ability of census derived models to map winter avian
500 distribution. *Ecological Research*, **21**, 681-691.
- 501 Chapin III, F.S., Zavaleta, E.S., Eviner, V.T., Naylor, R.L., Vitousek, P.M., Reynolds,
502 H.L., Hooper, D.U., Lavorel, S., Sala, O.E., Hobbie, S.E., Mack, M.C. & Díaz,
503 S. (2000) Consequences of changing biodiversity. *Nature*, **405**, 234-242.
- 504 Conlisk, E., Conlisk, J, Enquist, B., Thompson, J. & Harte J. (2009) Improved
505 abundance prediction from presence–absence data. *Global Ecology and*
506 *Biogeography*, **18**, 1-10.
- 507 De' Ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology and*
508 *Society*, **88**, 243-251.

509 Del Arco, M., Wildpret, W., Pérez de Paz, P.L., Rodríguez, O., Acebes, J.R., García, A.,
510 Martín, V.E., Reyes, J.A., Salas, M., Díaz, M.A., Bermejo, J.A., González, R.,
511 Cabrera, M.V. & García, S. (2003) *Cartografía 1:25.000 de la Vegetación*
512 *Canaria*. GRAFCANS. A., Santa Cruz de Tenerife.

513 de Nascimento, L., Willis, K.J., Fernández-Palacios, J.M., Criado, C. & Whittaker, R.J.
514 (2009) The long-term ecology of the forest of La Laguna, Tenerife (Canary
515 Islands). *Journal of Biogeography*, **36**, 499-514.

516 Elith, J., Leathwick, J. & Hastie T. (2008) A working guide to boosted regression trees.
517 *Journal of Animal Ecology*, **77**, 802-813.

518 Estes, J.A., Tinker, M.T., Williams, T.M. & Doak, D.F. (1998) Killer whale predation
519 on sea otters linking oceanic and nearshore ecosystems. *Science*, **282**, 473-476.

520 Estrada, A. & Arroyo, B. (2012) Occurrence vs abundance models: Differences between
521 species with varying aggregation patterns. *Biological Conservation*, **152**, 37-45.

522 Fernández-Palacios, J.M. & Martín-Esquivel, J.L. (2001) *Naturaleza de las Islas*
523 *Canarias: ecología y conservación*. Turquesa, Santa Cruz de Tenerife.

524 Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction
525 errors in conservation presence/absence models. *Environmental Conservation*,
526 **24**, 38-49.

527 Gerrard, D.J. & Chiang, H.C. (1970) Density estimation of corn rootworm egg
528 populations based upon frequency of occurrence. *Ecology*, **51**, 235-243.

529 Gerrard, D.J. & Cook, R.D. (1972) Inverse binomial sampling as a basis for estimating
530 negative binomial population densities. *Biometrics*, **28**, 971-980.

531 Grinstead, C.M. & Snell, J.L. (1997) Central Limit Theorem. *Introduction to*
532 *Probability* (2nd ed.). pp. 325-360. American Mathematical Society.

- 533 Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than
534 simple habitat models. *Ecology Letters*, **8**, 993-1009.
- 535 Hall, D.G., Childers, C.C. & Eger, J.E. (2007) Binomial sampling to estimate rust mite
536 (Acari: Eriophyidae) densities on orange fruit. *Journal of Economic Entomology*,
537 **100**, 233-240.
- 538 Hastie, T. & Fithian, W. (2013) Inference from presence-only data; the ongoing
539 controversy. *Ecography*, **36**, 864-867.
- 540 Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of
541 sample size and species characteristics on performance of different species
542 distribution modeling methods. *Ecography*, **29**, 773-785.
- 543 Järvinen, O. (1978) Species-specific census efficiency in line transects. *Ornis*
544 *Scandinavica*, **9**, 164-167.
- 545 Järvinen, O. & Väisänen, R.A. (1975) Estimating relative densities of breeding birds by
546 line transect method. *Oikos*, **26**, 316-322.
- 547 Järvinen, O. & Väisänen, R.A. (1976) Estimating relative densities of breeding birds by
548 the line transect method. IV. Geographical constancy of the proportion of main
549 belt observations. *Ornis Fennica*, **53**, 87-91.
- 550 Jiménez-Valverde, A., Diniz, F., de Azevedo, E.B. & Borges, P.A.V. (2009) Species
551 distribution models do not account for abundance: the case of arthropods on
552 Terceira Island. *Annales Zoologici Fennici*, **46**, 451-464.
- 553 Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2008) Not as good as they seem: the
554 importance of concepts in species distribution modelling. *Diversity and*
555 *Distributions*, **14**, 885-890.

556 Juan, C., Emerson, B.C., Oromí, P. & Hewitt, G.M. (2000) Colonization and
557 diversification: towards a phylogeographic synthesis for the Canary Islands.
558 *Trends in Ecology and Evolution*, **15**, 104-109.

559 Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting
560 the performance of climatic envelope models. *Ecological Applications*, **13**, 853-
561 867.

562 Lobo, J.M., Jimenez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the
563 performance of predictive distribution models. *Global Ecology and*
564 *Biogeography*, **17**, 145-151.

565 Manly, B.F.J., McDonald, L.L., Thomas, D.L., McDonald, T.L. & Erickson, W.P.
566 (2002) *Resource Selection by Animals: Statistical Design and Analysis for Field*
567 *Studies*. Kluwer Academic Publishes, Dordrecht, The Netherlands.

568 Mcfarland, T.M., Van Riper III, C. & Johnsona, G.E. (2012) Evaluation of NDVI to
569 assess avian abundance and richness along the upper San Pedro River. *Journal*
570 *of Arid Environments* **77**, 45-53.

571 Niculescu-Mizil, A. & Caruana, R. (2005) Obtaining Calibrated Probabilities from
572 Boosting. Proc. 21st Conference on Uncertainty in Artificial Intelligence AUAI
573 Pres. <http://arxiv.org/ftp/arxiv/papers/1207/1207.1403.pdf>

574 Nielsen, S.E., Johnson, C.J., Heard, D.C. & Boyce, M.S. (2005) Can models of
575 presence– absence be used to scale abundance? Two case studies considering
576 extremes in life history. *Ecography*, **28**, 197-208.

577 Pearce, J. & Ferrier, S. (2000) An evaluation of alternative algorithms for fitting species
578 distribution models using logistic regression. *Ecological Modelling*, **128**, 127-
579 147.

- 580 Pearce, J. & Ferrier, S. (2001) The practical value of modelling relative abundance of
581 species for regional conservation planning: a case study. *Biological*
582 *Conservation*, **98**, 33-43.
- 583 Pearce, J., Ferrier, S. & Scotts, D. (2001) An evaluation of the predictive performance
584 of distributional models for flora and fauna in northeast New South Wales.
585 *Journal of Environmental Management*, **62**, 171-184.
- 586 Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with
587 presence-only data. *Journal of Applied Ecology*, **43**, 405-412.
- 588 Perrins, C. (1998) *The complete birds of the Western Palearctic on CD-ROM*. Oxford
589 University Press, Oxford.
- 590 Phillips, S.J. & Dudik, M. (2008) Modeling of species distributions with Maxent: new
591 extensions and a comprehensive evaluation. *Ecography*, **3**, 161-175.
- 592 Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modelling of
593 species geographic distributions. *Ecological Modelling*, **190**, 231-259.
- 594 Rodríguez, J.P., Brotons, L., Bustamante, J. & Seoane, J. (2007) The application of
595 predictive modelling of species distribution to biodiversity conservation.
596 *Diversity and Distributions*, **13**, 243-251.
- 597 Russell, D.J.F., Wanless, S., Collingham, Y.C., Anderson, B.J., Beale, C., Reid, J.B.,
598 Huntley, B. & Hamer, K.C. (2015) Beyond climate envelopes: bio-climate
599 modelling accords with observed 25-year changes in seabird populations of the
600 British Isles. *Diversity and Distributions*, **21**, 211-222.
- 601 Seoane, J., Carrascal, L.M., Alonso, C.L. & Palomino, D. (2005) Species-specific traits
602 associated to prediction errors in bird habitat suitability modelling. *Ecological*
603 *Modelling*, **185**, 299-308.

- 604 Seoane, J., Carrascal, L.M. & Palomino, D. (2011) Assessing the ecological basis of
605 conservation priority lists for bird species in an island scenario. *Journal for*
606 *Nature Conservation*, **19**, 103-115.
- 607 Tellería, J.L. & Sáez-Royuela, C. (1986) The use of the frequency in the study of large
608 mammals abundance. *Acta Oecologica*, **7**, 69-75.
- 609 Thuiller, W., Münkemüller, T., Schiffrers, K.H., Georges, D., Dullinger, S., Eckhart,
610 V.C., Edwards, T.C., Gravel, J.D., Kunstler, G., Merow, C., Moore, K., Piedallu,
611 C., Vissault, S., Zimmermann, N.E., Zurrell, D. & Schurr, F.M. (2014) Does
612 probability of occurrence relate to population dynamics? *Ecography*, **37**, 1155-
613 1166.
- 614 Tôrres, N.M., De Marco, P., Santos, T., Silveira, L., de Almeida Jácomo, A.T. & Diniz-
615 Filho, J.A.F. (2012) Can species distribution modelling provide estimates of
616 population densities? A case study with jaguars in the Neotropics. *Diversity and*
617 *Distributions*, **18**, 615-627.
- 618 Van Couwenberghe, R., Collet, C., Pierrat, J.C., Verheyen, K. & Gégout, J.C. (2013)
619 Can species distribution models be used to describe plant abundance patterns?
620 *Ecography*, **36**, 665-674.
- 621 VanDerWal, J., Shoo, L.P., Johnson, C.N. & Williams, S.E. (2009) Abundance and the
622 environmental niche: environmental suitability estimated from niche models
623 predicts the upper limit of local abundance. *The American Naturalist*, **174**, 282-
624 291.
- 625 Yamamoto, N., Yokoyama, J. & Kawata, M. (2007) Relative resource abundance
626 explains butterfly biodiversity in island communities. *Proceedings of the*

627 *National Academy of Sciences of the United States of America*, **104**, 10524-
628 10529.

629 Yañez-Arenas, C., Guevara, S., Martínez-Meyer, E., Mandujano, S. & Lobo, J.M.
630 (2014) Predicting species' abundances from occurrence data: effects of sample
631 size and bias. *Ecological Modelling*, **294**, 36-41.

632

633

634 **SUPPORTING INFORMATION**

635 Additional Supporting Information may be found in the online version of this article:

636

637 **Appendix S1** Degree of triangularity in the relationships between local observed and
638 predicted abundances.

639

640 **Appendix S2** Species-specific characteristics describing the distribution-abundance
641 patterns of bird species.

642

643

644 **Biosketch**

645 Luis M. Carrascal is a research professor at the Museo Nacional de Ciencias Naturales
646 (CSIC, Spain). His current research interests are focused on macroecology, the
647 biogeographical ecology of the avifauna of the South-western Palaearctic and on the
648 study of habitat selection in birds for modelling patterns of species
649 abundance/occurrence.

Table 1 Summary of model results for 21 bird species in La Palma Island (Canary Islands, Spain). **MaxEnt AUC**: AUC values from MaxEnt models; **BCT AUC**: AUC values from Boosted Classification Trees models; **r MaxEnt**: correlation coefficients from Pearson's correlations between MaxEnt outputs and estimated specie's abundances; **r BCT**: coefficients from Pearson's correlations between BCT outputs and estimated specie's abundances (significant correlations at $P < 0.05$ after sequential Bonferroni correction are shown in bold type); **DENREG pred**: average regional density (birds / km²) predicted from transformed BCT probabilities in all transects; **DENREG est**: estimated regional density (birds / km²) derived from all transects; **% change**: % difference between predicted and estimated regional densities in relation to estimated regional density. Predictions were obtained from five-fold cross-validations. Data on species presences/absences were obtained from 437 transects covering all habitats of the island.

Species	MaxEnt AUC	BCT AUC	r MaxEnt	r BCT	DENREG pred	DENREG est	% change
<i>Alectoris barbara</i>	0.796	0.709	0.287	0.633	2.5	2.6	-2.9
<i>Anthus berthelotii</i>	0.892	0.895	0.624	0.818	14.5	13.4	7.7
<i>Carduelis cannabina</i>	0.683	0.680	0.141	0.714	1.4	2.6	-45.9
<i>Columba bolli</i>	0.985	0.944	0.754	0.853	6.7	7.9	-16.2
<i>Columba junoniae</i>	0.954	0.868	0.634	0.863	14.0	12.0	16.5
<i>Columba livia</i>	0.707	0.780	0.332	0.682	23.7	74.1	-68.0
<i>Erithacus rubecula</i>	0.872	0.921	0.550	0.864	25.3	25.9	-2.3
<i>Falco tinnunculus</i>	0.594	0.647	0.119	0.731	6.4	3.8	68.1
<i>Fringilla coelebs</i>	0.862	0.922	0.551	0.806	31.6	37.4	-15.6
<i>Motacilla cinerea</i>	0.908	0.901	0.472	0.847	8.9	5.3	67.3
<i>Parus caeruleus</i>	0.757	0.815	0.349	0.759	25.5	22.9	11.2
<i>Phylloscopus canariensis</i>	0.580	0.932	0.091	0.635	188.1	186.5	0.9
<i>Pyrrhocorax pyrrhocorax</i>	0.599	0.643	0.154	0.644	7.3	15.4	-52.7
<i>Regulus regulus</i>	0.866	0.941	0.576	0.823	81.2	93.5	-13.1
<i>Serinus canaries</i>	0.668	0.890	0.344	0.799	78.5	89.4	-12.3
<i>Streptopelia decaocto</i>	0.940	0.965	0.611	0.803	11.5	17.7	-35.1
<i>Streptopelia turtur</i>	0.610	0.559	0.083	0.751	7.7	4.6	68.3
<i>Sylvia atricapilla</i>	0.864	0.904	0.575	0.818	36.5	40.3	-9.5
<i>Sylvia conspicillata</i>	0.789	0.847	0.268	0.736	5.4	4.4	22.2
<i>Sylvia melanocephala</i>	0.871	0.882	0.561	0.831	23.2	20.3	14.7
<i>Turdus merula</i>	0.834	0.908	0.536	0.810	70.4	74.1	-4.9

Table 2 Alternative models for interspecific variation in large-scale prediction accuracy of bird density in 21 species inhabiting La Palma island (Canary Islands, Spain). Accuracy is measured as the percentage of variation of predicted average densities with respect to estimated average densities of birds in the whole sample of the 437 0.5-km line transects (see % change in Table 1). Only models with $\Delta AICc < 2$ are shown for brevity. Multimodel inference (lower part of the table) has been obtained considering all the possible combinations of predictors (64 models), averaging the results according model weights (W_i). Figures for each variable are standardized regression coefficients (β) obtained in general linear models. For each variable, ΣW_i is the sum of weights of the models in which the variable appears, weighted average β is the weighed average of standardized regression coefficients, and $se \beta$ the unconditional standard errors. AICc: AIC corrected for small sample sizes. R^2 : variance explained by each model (in %). CV%: coefficient of variation in bird numbers in transects where each species occurred; HB: habitat breadth considering 11 different habitats; PREV: prevalence of each species in the sample of 437 0.5-km line transects; DETECT: ratio of main belt (25 m) to total belt observations of each bird species (larger figures correspond to less detectable species); MASS: body mass of species (in log); DMAX: maximum density recorded in 11 different habitats. See Appendix S2 for more details on species characteristics. Models 1 to 5 are highly significant ($P < 0.001$) using the classical frequentist approach.

large-scale accuracy	standardized regression coefficients (β)						R^2 (%)	W_i	AICc
	PREV	CV%	DETECT	HB	DMAX	MASS			
model 1		-0.718		0.280	-0.351		75.0	0.166	194.6
model 2	-0.450	-0.726		0.463			74.9	0.155	194.7
model 3	-0.607	-0.674		0.568		-0.224	78.6	0.111	195.3
model 4		-0.765			-0.242		68.5	0.081	196.0
model 5		-0.792					62.7	0.066	196.4
MULTIMODEL INFERENCE									
	ΣW_i	0.438	1.000	0.185	0.632	0.483	0.268		
	weighted average β	-0.150	-0.737	-0.019	0.246	-0.165	-0.046	72.7	
	se β	0.265	0.139	0.054	0.242	0.212	0.086		

Figure legends

Figure 1 Location of 437 0.5-km transects in La Palma island. Each dot represents the center of the 0.5-km transects. The background map shows the topography of the island.

Figure 2 Linear relationship between predicted and estimated average regional densities for 21 species in La Palma Island (Canary Islands, Spain). Predictions were obtained from five-fold cross-validated boosted classification trees, whose outputs were converted to regional densities (through $-\ln [1-f]$; see methods). Solid line denotes the regression line and dashed line denotes equality between the estimated and predicted densities.

Figure 3 Partial residual plots illustrating the influence of the coefficient of variation in bird numbers where they occurred (a), and habitat breadth (b) on the accuracy of predicted average regional densities measured as the percentage difference between predicted and estimated regional density respect to estimated regional density (% change in Table 1). $N = 21$ bird species from La Palma island (Canary Islands, Spain). Residual plots show the relationship between a given independent variable and the response given that the other independent variables in Table 2 are also in the model, therefore partialling out their effects.

Figure 1

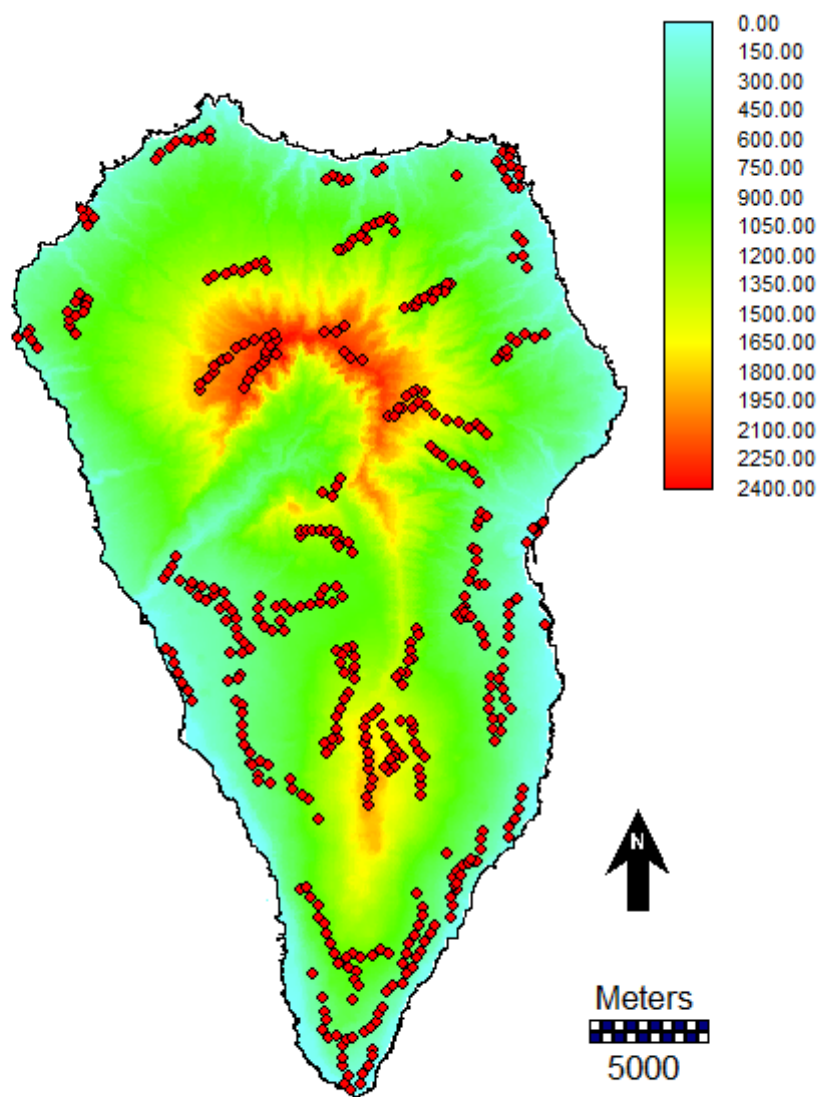


Figure 2

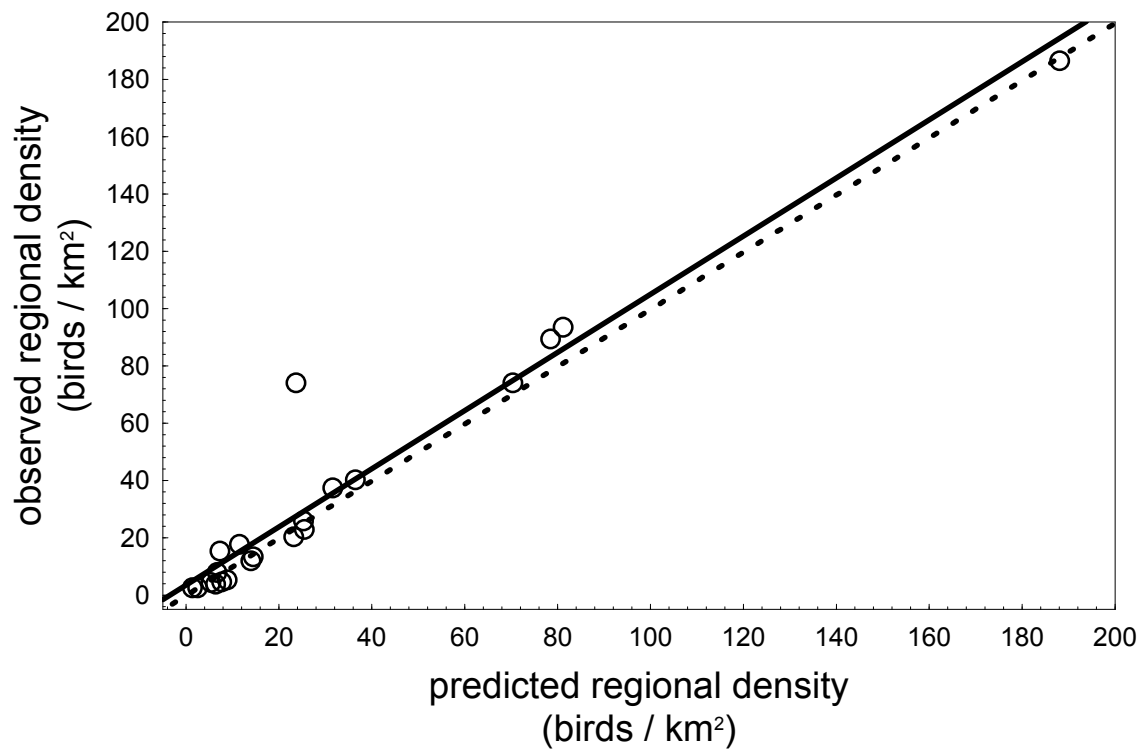
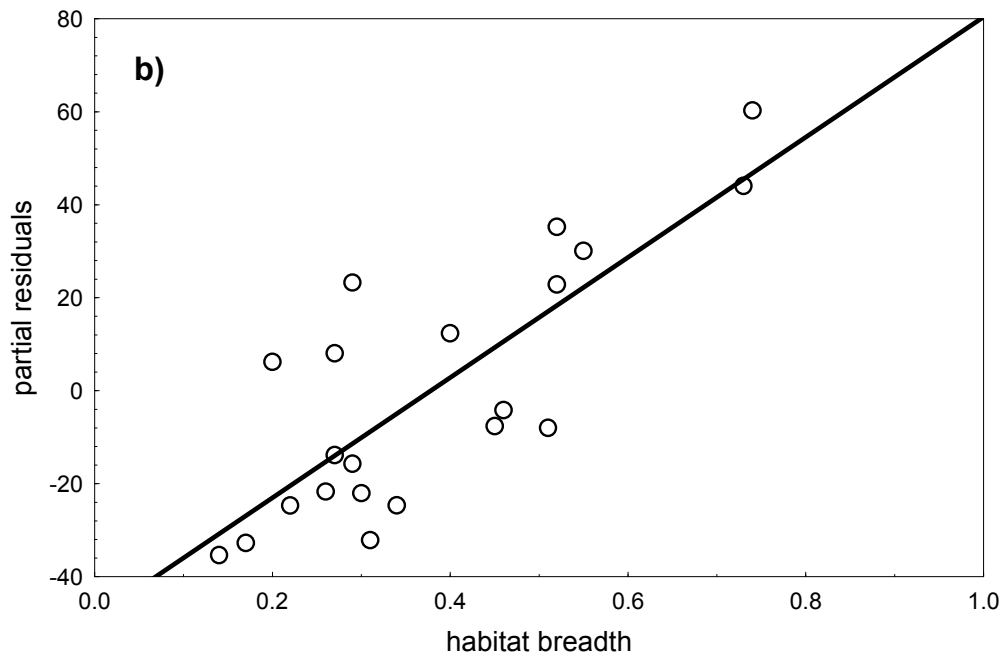
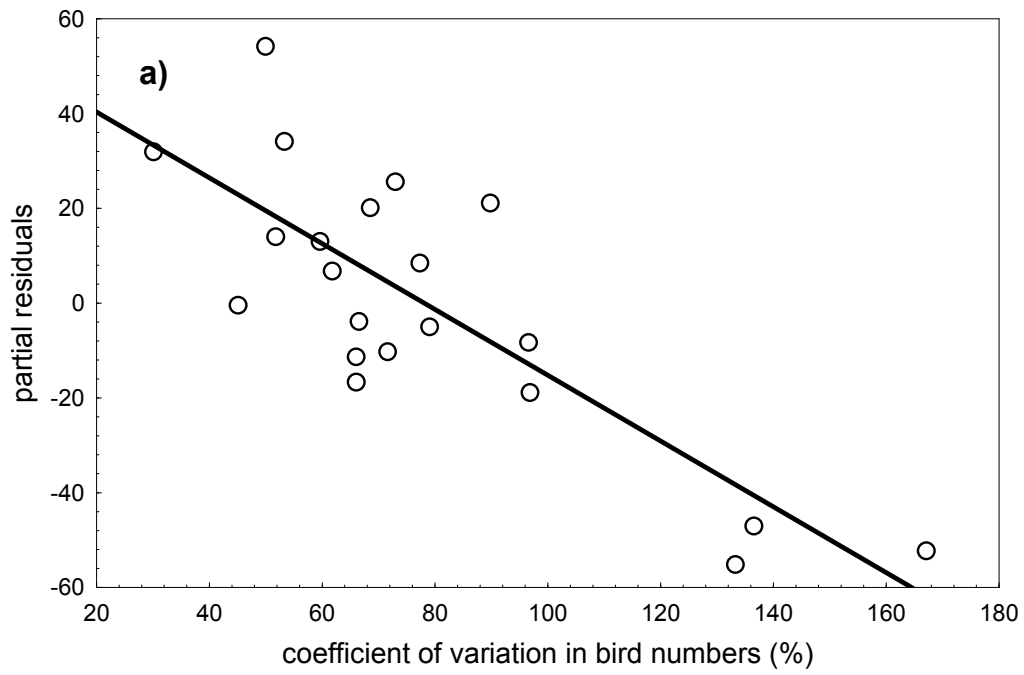


Figure 3



Appendix S1 Assessment of the degree of triangularity in the relationships between estimated local abundances and model outputs

The shape of the relationships between the estimated local abundances at the transect level and the SODM outputs were also analysed using quantile regression models (Cade & Noon, 2003), in order to test for changes in the slope of local abundance vs. suitability or probability derived from models in different subsets of data. We estimated the slopes of local abundances – SODM outputs at percentiles 50% (i.e., median regression, $\tau = 0.5$) and 90% (i.e., the maximum response of organisms attaining maximum ecological abundances; $\tau = 0.9$). The changes of the slopes between percentiles 50% and 90% ($b_{\tau=0.9} - b_{\tau=0.5}$) measure the unequal variation of local abundance with suitability or probability derived from models, indicating complex interactions between these two parameters that show solid, triangular, patterns instead of clearly linear relationships. Sequential Bonferroni adjustments were also applied to estimate the significance of slopes at the two selected percentiles, and the “triangularity” degree of the relationship between local abundances – SODM outputs, using BCT and MaxEnt, was tested by means of paired t -tests of the differences ($b_{\tau=0.9} - b_{\tau=0.5}$).

We found that patterns of the relationship between estimated local abundances and predictions of bird distributions from both BCT and MaxEnt models were triangular (see Table S1 and an example with the endemic subspecies *Regulus regulus ellenthalerae* in Fig. S1): quantile slopes for percentile 90% were significantly higher than those for 50% both for BCT probabilities (paired t -test comparing slopes at $\tau = 0.9$ vs. $\tau = 0.5$: $t = 7.10$, 21 species, $P \ll 0.001$) and MaxEnt suitabilities ($t = 6.22$, 21 species, $P \ll 0.001$). Twenty out of 21 bird species have significant 0.9-quantile slopes

relating estimated local abundance to BCT predictions of probability of occurrence (established after sequential Bonferroni's correction for multiple P estimates); nevertheless, 0.9-quantile slopes for MaxEnt predictions attained the significance level for only 16 bird species. The triangularity of the relationship (estimated abundances – SDOM outputs), measured by the difference in the quantile slopes at $\tau = 0.9$ and 0.5 ($b_{\tau=0.9} - b_{\tau=0.5}$), was not different comparing BCT and MaxEnt models (paired t -test: $t = -0.372$, 21 species, $P = 0.714$).

The shape of the distribution (estimated local abundance – predicted probability or suitability) is triangular, in such a way that lower predicted probabilities remain associated to lower estimated abundances, whereas higher predicted probabilities remain associated to a higher variation in estimated abundances (see also VanDerWal *et al.*, 2009; Gutiérrez *et al.*, 2013). Several non-exclusive potential explanations underline these triangular distributions. First, it may be simply the asymmetric meaning of presence/absence data regarding animal abundance. The absence of a species in the area covered by the sampling unit, if true, has a unique possible value of zero individuals; but the presence of a species may have a very large span of figures ranging from one to many individuals (Comte & Grenouillet, 2013). This concern has been previously acknowledged in the analysis of spatial variation of binomial response variables, with an overall higher variability and bias of results for binary data (McCullagh & Nelder, 1989; Guisan & Zimmermann, 2000; Cushman & McGarigal, 2004). Second, the tendency of presence-absence data to derive triangular relationships with abundance might depend on the used resolution (see Bean *et al.*, 2014) and the aggregation of the focus species. Third, the local abundance of a species cannot change above some upper limit set by the measured environmental predictors included in the modelling tools

(BCT and MaxEnt in this paper), but might change below that upper limit according to some limiting unmeasured variables (Cade & Noon, 2003). Moreover, multiplicative interactions among unmeasured ecological factors might contribute to the residual variation in the estimated abundance when it is predicted from SDOM. Finally, there are limits to prediction accuracy unbeatable by methodological refinements (Seoane *et al.*, 2005), which are rooted on stochastic phenomena due to natural or anthropogenic factors (e.g., harsh weather, wildfires, hunting, poisoning), or to endogenous metapopulations' cycles unreachable by coarse grained environmental predictors obtained from GIS. In spite of this, the averaging of local abundance estimates over larger spatial scales compensates those components of random variation and generates precise projections of animal numbers at the regional level (i.e. La Palma island; see Table 1 and Fig. 2 in the main text).

Additional references not included in the main text

- Cade, B.S. & Noon B.R. (2003) A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, **1**, 412-420.
- Comte, L. & Grenouillet, G. (2013) Species distribution modelling and imperfect detection: comparing occupancy versus consensus methods. *Diversity and Distributions*, **19**, 996–1007.
- Cushman, S.A. & McGarigal, K. (2004) Patterns in the species-environment relationship depend on both scale and choice of response variables. *Oikos*, **105**, 117-124.
- Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147-186.

- Gutiérrez, D., Harcourt, J., Díez, S.B., Gutiérrez-Illán, J. & Wilson, R.J. (2013) Models of presence–absence estimate abundance as well as (or even better than) models of abundance: the case of the butterfly *Parnassius apollo*. *Landscape Ecology*, **28**, 401-413.
- McCullagh, P. & Nelder, J.A. (1989) *Generalized linear models*. Chapman and Hall, London.

Table S1 Slopes of quantile regressions of estimated local abundances of species in 0.5-km length transects and outputs derived from BCT (boosted classification trees; probability) and MaxEnt (suitability) models at percentiles 50% ($\tau = 0.5$) and 90% ($\tau = 0.9$). Significant slopes at $P < 0.05$ after sequential Bonferroni correction are shown in bold type.

Species	BCT		MaxEnt	
	$\tau = 0.5$	$\tau = 0.9$	$\tau = 0.5$	$\tau = 0.9$
<i>Alectoris barbara</i>	0.00	10.72	0.00	5.23
<i>Anthus berthelotii</i>	1.16	3.72	0.99	4.90
<i>Carduelis cannabina</i>	1.09	7.30	0.00	0.00
<i>Columba bolli</i>	3.29	6.54	3.21	9.92
<i>Columba junoniae</i>	2.33	5.92	1.20	8.63
<i>Columba livia</i>	0.50	2.87	0.00	2.61
<i>Erithacus rubecula</i>	1.72	4.25	0.00	6.32
<i>Falco tinnunculus</i>	2.37	5.03	0.00	0.00
<i>Fringilla coelebs</i>	1.36	4.13	0.77	5.74
<i>Motacilla cinerea</i>	2.55	5.23	0.00	8.67
<i>Parus caeruleus</i>	1.11	4.00	0.00	3.47
<i>Phylloscopus canariensis</i>	1.17	2.73	0.00	0.00
<i>Pyrrhocorax pyrrhocorax</i>	0.41	2.69	0.00	1.06
<i>Regulus regulus</i>	1.36	3.71	1.22	5.20
<i>Serinus canaries</i>	0.90	2.52	0.68	3.02
<i>Streptopelia decaocto</i>	0.92	8.73	0.00	9.34
<i>Streptopelia turtur</i>	2.69	8.14	0.00	0.00
<i>Sylvia atricapilla</i>	1.18	3.34	1.30	4.24
<i>Sylvia conspicillata</i>	2.77	8.30	0.00	3.56
<i>Sylvia melanocephala</i>	2.18	4.45	0.00	5.75
<i>Turdus merula</i>	1.20	3.17	1.35	3.87

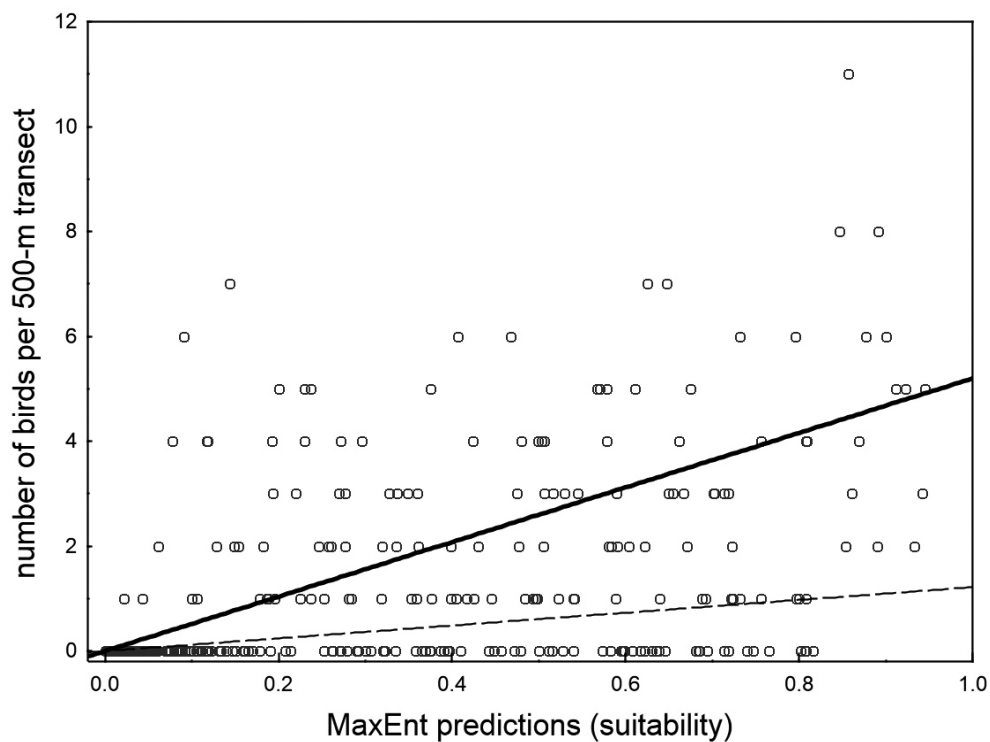
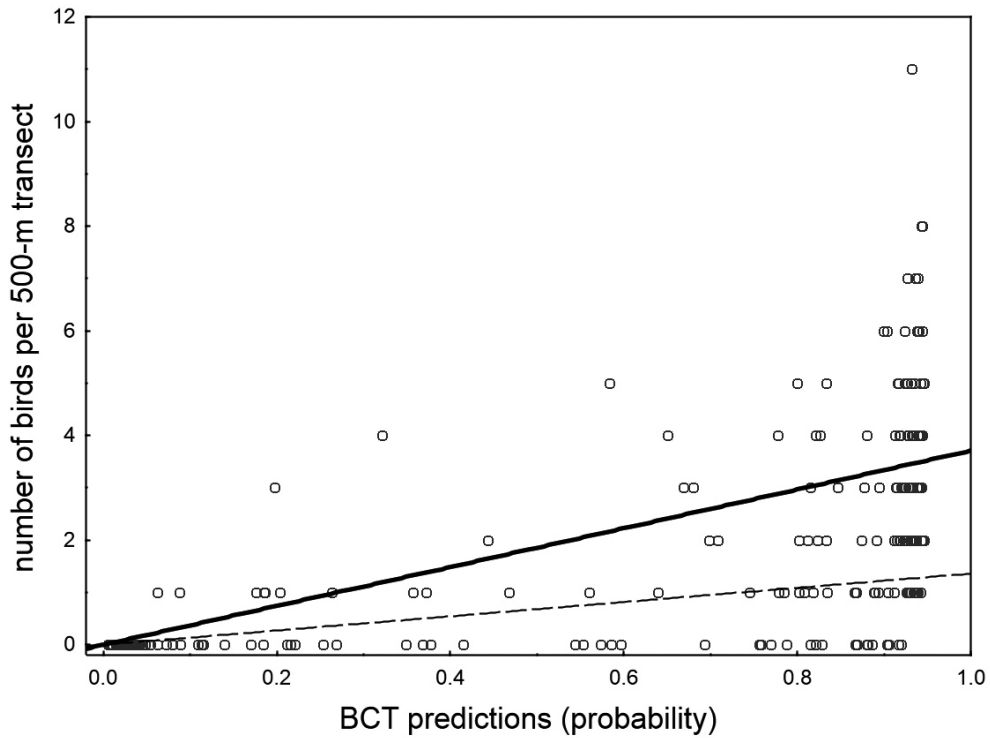


Figure S1 Shape of the relationship between estimated local abundance and predictions of probability of occurrence and suitability derived, respectively, from BCT and MaxEnt models. The panels show the relationships for the endemic subspecies *Regulus regulus ellenthalerae*. Regression lines show the quantile regressions for $\tau = 0.5$ and $\tau = 0.9$. $n = 437$ sample units (0.5-km length transects).

Appendix S2

Table S2 Species-specific characteristics describing the distribution-abundance patterns of 21 terrestrial bird species in La Palma island. PREV: prevalence of each species in the sample of 437 line transects; CV%: coefficient of variation in bird numbers in transects where each species occurred; p: ratio of main belt (25 m) to total belt observations of each bird species (larger figures correspond to less detectable species); HB: habitat breadth considering 11 different habitats; MASS: body mass of species (in log); DMAX: maximum density recorded in 11 different habitats. Data for DMAX and HB obtained from Appendix B of Seoane *et al.* (2011).

	PREV	CV%	p	HB	DMAX	MASS
<i>Alectoris barbara</i>	0.05	45.1	0.39	0.31	3.5	480.0
<i>Anthus berthelotii</i>	0.17	89.8	0.42	0.20	64.5	16.5
<i>Carduelis cannabina</i>	0.04	136.5	0.23	0.26	16.3	17.6
<i>Columba bolli</i>	0.06	61.8	0.59	0.14	58.3	286.0
<i>Columba junoniae</i>	0.10	73.0	0.59	0.27	42.9	328.7
<i>Columba livia</i>	0.33	133.2	0.45	0.45	117.9	216.0
<i>Cyanistes caeruleus</i>	0.25	71.6	0.55	0.46	29.2	11.3
<i>Erithacus rubecula</i>	0.19	66.5	0.61	0.30	60.6	16.7
<i>Falco tinnunculus</i>	0.19	30.1	0.22	0.73	3.6	174.5
<i>Fringilla coelebs</i>	0.25	79.1	0.63	0.29	112.0	23.0
<i>Motacilla cinerea</i>	0.08	50.0	0.53	0.29	12.5	18.0
<i>Phylloscopus canariensis</i>	0.90	68.5	0.47	0.74	248.1	7.7
<i>Pyrhacorax pyrrhacorax</i>	0.14	167.1	0.32	0.40	21.6	321.5
<i>Regulus regulus</i>	0.31	66.0	0.89	0.34	146.5	5.8
<i>Serinus canarius</i>	0.50	96.6	0.52	0.52	124.8	15.3
<i>Streptopelia decaocto</i>	0.08	96.8	0.73	0.17	54.3	196.0
<i>Streptopelia turtur</i>	0.07	53.3	0.57	0.52	6.4	125.0
<i>Sylvia atricapilla</i>	0.39	66.1	0.42	0.51	48.4	22.3
<i>Sylvia conspicillata</i>	0.07	51.8	0.51	0.22	11.6	9.5
<i>Sylvia melanocephala</i>	0.19	59.6	0.60	0.27	52.5	11.2
<i>Turdus merula</i>	0.54	77.3	0.49	0.55	130.2	86.1