Modelling invasive alien species distributions from digital biodiversity atlases. Model upscaling as a means of reconciling data at different scales

Arnald Marcer¹, Joan Pino¹, Xavier Pons² and Lluís Brotons¹³

¹CREAF (Centre de Recerca Ecològica i Aplicacions Forestals), Universitat Autònoma de Barcelona, E-08193, Bellaterra, Catalonia, Spain, ²Departament de Geografia, Universitat Autònoma de Barcelona, E-08193, Bellaterra, Catalonia, Spain, ³Grau d’Ecologia del Paisatge, Àrea de Biodiversitat, Centre Tecnològic Forestal de Catalunya, E-25280, Solsona, Catalonia, Spain

ABSTRACT

Aim There is a wealth of information on species occurrences in biodiversity data banks, albeit presence-only, biased and scarce at fine resolutions. Moreover, fine-resolution species maps are required in biodiversity conservation. New techniques for dealing with this kind of data have been reported to perform well. These fine-resolution maps would be more robust if they could explain data at coarser resolutions at which species distributions are well represented. We present a new methodology for testing this hypothesis and apply it to invasive alien species (IAS).

Location Catalonia, Spain.

Methods We used species presence records from the Biodiversity data bank of Catalonia to model the distribution of ten IAS which, according to some recent studies, achieve their maximum distribution in the study area. To overcome problems inherent with the data, we prepared different correction treatments: three for dealing with bias and five for autocorrelation. We used the MaxEnt algorithm to generate models at 1-km resolution for each species and treatment. Acceptable models were upscaled to 10 km and validated against independent 10 km occurrence data.

Results Of a total of 150 models, 20 gave acceptable results at 1-km resolution and 12 passed the cross-scale validation test. No apparent pattern emerged, which could serve as a guide on modelling. Only four species gave models that also explained the distribution at the coarser scale.

Main conclusions Although some techniques may apparently deliver good distribution maps for species with scarce and biased data, they need to be taken with caution. When good independent data at a coarser scale are available, cross-scale validation can help to produce more reliable and robust maps. When no independent data are available for validation, however, new data gathering field surveys may be the only option if reliable fine-scale resolution maps are needed.

Keywords Biodiversity databases, Catalonia, cross-scale validation, invasive alien species, MaxEnt, species distribution models.

INTRODUCTION

For centuries, species occurrences have been recorded in an ad hoc way by natural historians, museums, scientists and the like in the form of museum specimens, site inventories, citations in technical and scientific literature, etc. (Chapman & Busby, 1994; Chapman, 2005). In the last three decades, both governments and non-governmental organizations have invested considerable financial resources on the digitizing of these data into digital species distribution atlases and making them publicly available. Ideally, they should offer reliable high-quality digital data, which withstand public, scientific
and legal scrutiny (Robertson et al., 2010). Such databases represent a wealth of information on species distribution and an indispensable asset for science and conservation (Funk & Richardson, 2002; Graham et al., 2004; Suárez & Tsutsui, 2004; Franklin, 2009; Robertson et al., 2010). However, as these data usually come from opportunistic or ad hoc sources rather than well-planned surveys, they present some important drawbacks: they are presence-only in nature, are highly biased (not representing a random sample) and may show spatial aggregation derived from sampling biases. Also, because they are laboriously compiled from analogue sources, they are difficult and costly to georeference, and hence, coarser resolutions tend to dominate (Margules & Pressey, 2000; Pressey, 2004). As a result, most data in these biodiversity databases are often too coarse for use in conservation planning and management, where fine-grained maps (i.e. 1 km or better) are needed (McPherson et al., 2006; Guralnick & Hill, 2009; Niamir et al., 2011).

Planned systematic surveys of species presence, absence and abundance could provide the most precise, accurate and unbiased information on the spatial distribution of biodiversity. However, such surveys are expensive to conduct for large regions, even for a single species (Robertson et al., 2010). Given the current accelerated trend in world-wide biodiversity loss and the urge for addressing conservation problems, it becomes of utmost importance to find ways and methodologies to make the best use of this existing information (Newbold, 2010; Venette et al., 2010).

Such databases or atlases of species occurrences still represent a largely untapped potential of information that can play an important role in conservation biogeography (Elith & Leathwick, 2009; Richardson & Whittaker, 2010; Elith et al., 2011). Methods for dealing with biases and evaluating results need more development. Fortunately, in the last few years, the community of species distribution modelling has witnessed the appearance of new tools and methodologies from the fields of statistics and artificial intelligence which have the potential to address the problems inherent in these data. Some of these tools, if used properly (Segurado et al., 2006; Pearson et al., 2007; Phillips et al., 2008; Elith et al., 2011), can handle presence-only data even when only few occurrences are available, and provided attention is given to problems such as bias and autocorrelation (Elith et al., 2011). Despite that, in presence-only models, the lack of absence data compromises the model reliability (Aranda & Lobo, 2011) and that, some issues have been raised concerning the maximum entropy technique [e.g. (Peterson et al., 2007, 2008)], it has been judged among the best performers in distribution modelling for such kind of data (Elith et al., 2006) and is still widely used by the research community.

In many cases, applying these novel techniques to the scarce fine-resolution data can yield distribution maps with high validation scores. The question is whether we can take advantage of coarser data, which are relatively abundant and for some regions close to the species true distribution, to further validate these maps and assess their reliability. Having models that explain data at fine resolutions while being consistent with coarser resolution data is important (McPherson et al., 2006; Vallecillo et al., 2009; Niamir et al., 2011) as it can yield more robust and reliable distribution maps for conservation. The existence of databases with such information provides an opportunity to test this hypothesis. A specially relevant piece of information for conservation, which can potentially be derived from these costly data banks, is current and historical distributions of invasive alien species (IAS).

The spread of IAS, driven mainly by human activities, is increasing worldwide (Butchart et al., 2010) and poses potential problems not only to native biodiversity but also to economic development and human well-being (Vitousek et al., 1997; Taylor & Irwin, 2004; Pimentel et al., 2005; Chytrý et al., 2009; Pejchar & Mooney, 2009; Pyšek et al., 2010; Vila et al., 2011). Having risk maps representing the potential distribution of IAS is a necessary step towards effective management (Richardson & Whittaker, 2010; Jiménez-Valverde et al., 2011). Using all information in species inventories and atlases, coarse and fine-resolution records, is essential for maximizing limited financial resources (Nielsen et al., 2008).

We test whether fine-resolution maps of IAS can be obtained from existing biodiversity databases that are robust enough to explain occurrences at different scales held in the database. By using a probabilistic upscaling methodology, we translate 1-km resolution maps into 10-km resolution maps that are then used to further validate the models against well-known distributions at the coarser scale. With this approach, each model is evaluated with data at its own resolution. We use different bias and autocorrelation treatments to deal with problems inherent in the finer resolution data. Results suggest that, despite current modelling methodologies can deliver good fine-resolution models, these models cannot always explain well-known distributions at coarser scales.

METHODS

Area of study

Catalonia is a region of around 32,000 km² located in the northeastern part of the Iberian Peninsula (Fig. 1). It ranges in elevation from 0 to over 3000 m.a.s.l, from the Mediterranean coast up to the Pyrenees. Its environmental conditions are highly variable because of its location and complex topography. Although dominated by the Mediterranean climate, it also has continental and Atlantic influences. Mediterranean and Eurosiberian biogeographic regions dominate, while Subalpine and Alpine types appear in the upper zones of the Pyrenees. There is a trend of decreasing precipitation and increasing temperature towards the south (Ninyerola et al., 2000). It is a highly populated region, particularly around the Barcelona metropolitan area. The rest of the region is dominated by forests and agroforestry mosaics with
relatively large human influence although with lower intensity.

**Database**

As a case study, we used the Biodiversity Databank of Catalonia (BDBC, http://biodiver.bio.ub.es/biocat/) (Font et al., 2009). Catalonia is a region with a rich history in Botany which is reflected in the holdings of the database. For an area of around 32,000 km², the BDBC contains more than 1.5 million plant species citations from sources such as scientific articles, PhD theses and local floras. Most of its data are at a coarse resolution of 10 km for historical reasons – the recording tradition among botanists and the fact that, before the use of GPS, manually recording occurrences at finer resolutions was a time-consuming process. Catalonia is a well-surveyed region for vascular plants at 10-km resolution. The BDBC contains also more than 180,000 plant occurrence records at 1-km resolution. The geographical distribution of IAS as reflected by 10-km resolution occurrence records in the BDBC can be considered to approximate its true distribution for our study area (Pino et al., 2005). Therefore, when developing models at finer resolutions, their geographic distribution should be coherent with the distribution obtained from mapping the coarser data.

**Species data**

IAS may violate the assumption made in species distribution modelling that species have had the opportunity to realize their full potential range (Austin, 2002; Araújo & Pearson, 2005; Peterson, 2005; De Marco et al., 2008; Zimmermann et al., 2010; Václavík & Meentemeyer, 2012). According to some studies (Williamson et al., 2009; Gassó et al., 2010), neophytes reach their maximum range around 150 years after their introduction into the Iberian Peninsula. The 150-year period was hence taken as the main criterion for species selection, together with the availability of data at 1-km resolution. The 10 species selected are presented in Table 1. Four of them are considered to be highly invasive in Europe (DAISIE European Invasive Alien Species Gateway, 2011): Ailanthus altissima, Opuntia ficus-indica, Oxalis pes-caprae and Robinia pseudoacacia. See Appendix S1 in Supporting Information for maps of occurrences for each species.

**Data independence across scales**

Occurrence data at different resolutions in biodiversity atlases may not be independent; that is, occurrence records at coarser resolutions may have their origin in records at finer resolutions. To overcome this difficulty, we only

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbr</th>
<th>1 km</th>
<th>10 km</th>
<th>Intr</th>
<th>Yrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agave Americana L.</td>
<td>aga</td>
<td>20</td>
<td>124</td>
<td>XVIth</td>
<td>411</td>
</tr>
<tr>
<td>Ailanthus altissima (P.Mill) Swingle</td>
<td>aia</td>
<td>43</td>
<td>213</td>
<td>1818</td>
<td>192</td>
</tr>
<tr>
<td>Amaranthus albus L.</td>
<td>ama</td>
<td>29</td>
<td>194</td>
<td>1861</td>
<td>149</td>
</tr>
<tr>
<td>Conyza Canadensis (L.) Cronquist</td>
<td>coc</td>
<td>73</td>
<td>307</td>
<td>1784</td>
<td>226</td>
</tr>
<tr>
<td>Datura stramonium L.</td>
<td>das</td>
<td>31</td>
<td>230</td>
<td>XVIth</td>
<td>411</td>
</tr>
<tr>
<td>Oenothera biennis L.</td>
<td>oeb</td>
<td>55</td>
<td>80</td>
<td>1848</td>
<td>162</td>
</tr>
<tr>
<td>Opuntia ficus-indica (L.) Mill.</td>
<td>opf</td>
<td>13</td>
<td>102</td>
<td>XVIth</td>
<td>411</td>
</tr>
<tr>
<td>Oxalis pes-caprae L.</td>
<td>opx</td>
<td>12</td>
<td>41</td>
<td>1850</td>
<td>160</td>
</tr>
<tr>
<td>Robinia pseudoacacia L.</td>
<td>rop</td>
<td>66</td>
<td>257</td>
<td>XVIIIth</td>
<td>211</td>
</tr>
<tr>
<td>Xanthium spinosum L.</td>
<td>xas</td>
<td>56</td>
<td>252</td>
<td>XVIIIth</td>
<td>211</td>
</tr>
</tbody>
</table>

Abbr, species abbreviation; 1 km, number of 1 km occurrences; 10 km, number of 10 km occurrences; Intr, introduction date; Yrs, number of years since introduction (conservative estimate).
accepted 10 km squares which had at least one citation more per species than the sum of citations for the one hundred 1 km squares contained within the 10 km square; that is, there is at least one 10 km occurrence record which is independent from 1 km data. This procedure also allowed us to use all occurrences records at 1-km resolution.

Environmental data

We used 19 bioclimatic variables (Nix, 1986) (Table 2) that represent a combination of annual trends, seasonality and extreme conditions relevant to species physiological tolerances. We added two more variables regarding radiation (mean radiation of the least radiated quarter and mean radiation of the most radiated quarter) and three more variables that may partially explain the distribution of IAS (distance to main harbours, distance to the coast and degree of anthropization) (Brooks, 2007; Vicente et al., 2010) (See Table 2). We calculated the bioclimatic variables using the Digital Climatic Atlas of Catalonia (DCAC) (Ninyerola et al., 2000) that holds monthly data on temperature, precipitation and radiation for the whole of Catalonia. We calculated the degree of anthropization using the Land Cover Map of Catalonia (CREAF – Centre for Ecological Research & Forestry Applications, 2009; Ibáñez & Burriel, 2010). Each land cover category was assigned a value ranging from one (least anthropization) to five (most anthropization) (Table 2). Then, to represent the degree of anthropization, we calculated a weighted average scaled between 0 and 100 for each 1 km square grid.

As our goal is to predict species distributions rather than to understand which factors affect their distribution, all predictors were used for modelling each species. Extracting collinearity from the model was not necessary. Although collinearity can hinder the explanation of which variables affect species distribution, it does not affect MaxEnt predictive performance (Kuemmerle et al., 2010).

Species distribution modelling

Modelling involved a five-step process as shown in Fig. 2. In the first step, we modelled the species distribution at the finer resolution of 1 km following the methodology described in Case 1 of Elith et al. (2011). These authors use different alternative background scenarios to account for bias, and cross-validation techniques to validate models developed with presence-only data of Banksia prionotes from an atlas database. Accounting for bias and autocorrelation is an important issue in species distribution modelling, especially in presence-only models (Legendre, 1993; Legendre et al., 2002; Segurado et al., 2006; Phillips et al., 2009; Newbold, 2010; Merckx et al., 2011). As we expect fine-grained casually collected data to show a number of biases, we included three bias correction treatments and five spatial autocorrelation (SAC) correction treatments (see below) to evaluate the potential of these data to derive ecologically sound species distribution models.

This resulted in a total of 15 models per species. In a second step, only those models with an AUC (area under the curve) ≥ 0.7 not showing residual SAC were selected. In a third step, these selected models were upscaled to a coarser resolution of 10 km using a probabilistic model (see equation later). In step four, upscaled models were validated against the independent 10 km data set and only those with an AUC ≥ 0.7 at 10-km resolution (AUC10K) were selected. Therefore, the 1-km resolution models which, once upscaled, resulted in these selected 10 km models are the only ones which showed acceptable predictions at both scales. Finally, in step five, if more than one fine-scale model per species had been selected, we determined the best one by selecting that with the highest AUC at 1-km resolution (AUC1k).

Despite concerns about the use of AUC to compare species distribution models, this metric can safely be applied when evaluating model performance within species (Lobo et al., 2008; Blach-Overgaard et al., 2010) and when interpreting it as a measure of discrimination between presence and background rather than presence and absence (Phillips et al., 2006; Phillips & Dudík, 2008).

Table 2 Set of environmental predictors used in modelling

<table>
<thead>
<tr>
<th>Bioclimatic variables</th>
<th>Landscape and physical variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annual mean temperature</td>
<td>Anthropization degree</td>
</tr>
<tr>
<td>Minimum temperature of the coldest month</td>
<td>1 – Natural forests, shrublands,</td>
</tr>
<tr>
<td>Mean temperature of the coldest year quarter</td>
<td>wetlands, grasslands, rock</td>
</tr>
<tr>
<td>Mean temperature of the warmest year quarter</td>
<td>outcrops and screens, bare soil,</td>
</tr>
<tr>
<td>Mean temperature of the wettest year quarter</td>
<td>beaches, glaciers and snow cover</td>
</tr>
<tr>
<td>Mean temperature of the driest year quarter</td>
<td>and continental waters, 2 –</td>
</tr>
<tr>
<td>Maximum temperature of the warmest month</td>
<td>recently burnt areas and</td>
</tr>
<tr>
<td>Annual mean precipitation</td>
<td>reforestations, 3 – crops and</td>
</tr>
<tr>
<td>Precipitation of the coldest year quarter</td>
<td>tree plantations, 4 – agricultural</td>
</tr>
<tr>
<td>Precipitation of the driest month</td>
<td>water bodies and quarrying areas,</td>
</tr>
<tr>
<td>Precipitation of the warmest year quarter</td>
<td>5 – dense and sparse urban areas</td>
</tr>
<tr>
<td>Precipitation of the wettest month</td>
<td>and roads</td>
</tr>
<tr>
<td>Annual temperature range</td>
<td>Distance to coast</td>
</tr>
<tr>
<td>Mean temperature diurnal range</td>
<td>Distance to closest harbour</td>
</tr>
<tr>
<td>Isothermality</td>
<td></td>
</tr>
<tr>
<td>Temperature seasonality</td>
<td></td>
</tr>
<tr>
<td>Precipitation seasonality</td>
<td></td>
</tr>
<tr>
<td>Mean solar radiation of the least radiated quarter</td>
<td></td>
</tr>
<tr>
<td>Mean solar radiation of the most radiated quarter</td>
<td></td>
</tr>
</tbody>
</table>

Diversity and Distributions, 1–13, © 2012 Blackwell Publishing Ltd
Figure 2 Outline for the proposed modelling workflow. Step one corresponds to the modelling of each species at 1-km resolution with three different bias treatments and five different autocorrelation treatments, which gives a total of 15 models per species. In step two, we check for residual autocorrelation, calculate the AUC and select only those models with no residual spatial autocorrelation and with an $AUC \geq 0.7$. In step three, previously selected models are upscaled to 10-km resolution by probabilistic calculations. In step four, a ROC analysis is performed using independent data at 10-km resolution. Models with an $AUC \geq 0.7$ tell us which models at 1-km resolution are accepted. Finally, in step five, if more than one model per species at 1-km resolution has been accepted, we define the best model as the one which has the maximum $AUC_{1K}$. 
Bias correction treatments

Background samples should be chosen to reflect the spatial bias and thus to minimize the effects of bias in the data (Phillips et al., 2009; Veloz, 2009; Elith et al., 2011). We prepared three different background scenarios: (a) the entire study area (coded as ‘whole_area’), (b) 1 km squares with the presence of vascular plants citations (about 14.0% of the whole area, coded as ‘vasculars1k’) and (c) 1 km squares with the presence of IAS citations (about 2.3% of the whole area, coded as ‘invasive1k’). See Table 3.

Autocorrelation correction treatments

SAC may falsely inflate AUC measures for species distribution models with presence-only data (Segurado et al., 2006; Veloz, 2009) and environmental autocorrelation may have the same effects. There is no established methodology for accounting for SAC when dealing with presence-only data (Dormann et al., 2007; Elith & Leathwick, 2009). Autoregressive models are not applicable because both presence and absence data would be needed (Allouche et al., 2008). We took an a priori approach similar to (Segurado et al., 2006; Pearson et al., 2007) which consisted in filtering occurrences by setting a minimum spatial and environmental distance between them and then checking for residual autocorrelation.

We prepared five treatments for modelling each species. The first involved including all available presences without filtering them. The second and third involved randomly filtering and selecting occurrences so that any occurrence was at least at a spatial distance of 2830 m (two 1 km squares) and 4250 m (three 1 km squares) from each other, respectively. For the fourth and fifth treatment, we used a minimum multivariate environmental distance based on the Gower’s distance index with values 0.05 and 0.1, respectively (higher values resulted in an excessive reduction in occurrences). Models were then checked for significant residual autocorrelation [observed occurrence minus probability of occurrence as in (De Marco et al., 2008; Nuñez & Medley, 2011; Václavík & Meentemeyer, 2012)] by using Monte-Carlo simulation of Moran’s I autocorrelation coefficient using package spdep in R (Bivand, 2011). Only those models with a P-value ≥ 0.05 were accepted (as shown in Table 3).

Modelling and validation at 1-km resolution

We used MaxEnt software, version 3.3.3e, (Phillips et al., 2006; Phillips & Dudík, 2008). MaxEnt is a presence-background modelling tool based on the maximum entropy principle. There is wide agreement in the species distribution modelling community that it is the best available tool for presence-only data, even when only a limited number of

Table 3 Models with $AUC_{1k} \geq 0.7$ and no residual spatial autocorrelation at 1 km. Finally accepted models ($AUC_{10k}$ at 10 km with $\geq 0.7$) at 10-km resolution are indicated with a Y in column ‘Accepted’. In column ‘Best’, those with the highest $AUC_{1k}$ at 1 km from the accepted models are marked with an asterisk.

<table>
<thead>
<tr>
<th>Sp</th>
<th>Bias tr.</th>
<th>Aut. type</th>
<th>Min. dist.</th>
<th>$AUC_{1k}$</th>
<th>M P-value</th>
<th>$AUC_{10k}$</th>
<th>Accepted</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>aia</td>
<td>invasive1k</td>
<td>Spatial</td>
<td>2830 m</td>
<td>0.82</td>
<td>0.052</td>
<td>0.50</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>aia</td>
<td>vascular1k</td>
<td>Spatial</td>
<td>2830 m</td>
<td>0.78</td>
<td>0.076</td>
<td>0.65</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>aia</td>
<td>invasive1k</td>
<td>Spatial</td>
<td>4250 m</td>
<td>0.72</td>
<td>0.164</td>
<td>0.45</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>invasive1k</td>
<td>Spatial</td>
<td>0.10</td>
<td>0.72</td>
<td>0.124</td>
<td>0.67</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>opf</td>
<td>whole_area</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.75</td>
<td>0.086</td>
<td>0.57</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>whole_area</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.87</td>
<td>0.074</td>
<td>0.85</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>opf</td>
<td>whole_area</td>
<td>Spatial</td>
<td>2830 m</td>
<td>0.87</td>
<td>0.072</td>
<td>0.85</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>vascular1k</td>
<td>Spatial</td>
<td>2830 m</td>
<td>0.80</td>
<td>0.054</td>
<td>0.82</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>opf</td>
<td>whole_area</td>
<td>Spatial</td>
<td>4250 m</td>
<td>0.86</td>
<td>0.076</td>
<td>0.86</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>whole_area</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.70</td>
<td>0.078</td>
<td>0.78</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>vascular1k</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.87</td>
<td>0.136</td>
<td>0.89</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>whole_area</td>
<td>Spatial</td>
<td>0.00</td>
<td>0.94</td>
<td>0.054</td>
<td>0.93</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>oeb</td>
<td>vascular1k</td>
<td>Spatial</td>
<td>2830 m</td>
<td>0.86</td>
<td>0.056</td>
<td>0.90</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>rop</td>
<td>invasive1k</td>
<td>Environmental</td>
<td>0.05</td>
<td>0.83</td>
<td>0.252</td>
<td>0.52</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>rop</td>
<td>invasive1k</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.81</td>
<td>0.405</td>
<td>0.53</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>xas</td>
<td>vascular1k</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.75</td>
<td>0.150</td>
<td>0.69</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>xas</td>
<td>vascular1k</td>
<td>Environmental</td>
<td>0.10</td>
<td>0.76</td>
<td>0.577</td>
<td>0.70</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>

Sp, species abbreviation; Bias tr., bias treatment (whole_area, whole study area as background; vascular1k, UTM squares with citations of vascular plants as background; invasive1k, UTM squares with citations of invasive plants as background), Aut. Type, autocorrelation treatment type (spatial, based on spatial distance; environmental, based on environmental distance), Min. dist., autocorrelation minimum distance value; $AUC_{1k}$, AUC value for 1 km models; M P-value, Moran’s I P-value from Monte-Carlo simulation; $AUC_{10k}$, AUC value for 10 km models; Accepted, models accepted; Best, overall best models.
occurrence records are available (Elith et al., 2006; Hernandez et al., 2006; Phillips & Dudík, 2008; Wisz et al., 2008; Elith & Graham, 2009; Thorn et al., 2009; Costa et al., 2010) and with bias present (Rebelo & Jones, 2010). MaxEnt estimates the distribution of maximum entropy constrained in a way that expected values for predictor variables match their empirical average (Phillips et al., 2006). We used the logistic output of the model that indicates the relative environmental suitability of each pixel in relation to background for the study area (Phillips et al., 2006; Phillips & Dudík, 2008).

We ran the model for each species with default options using the whole set of environmental predictors (Table 2) and following the methodology explained in Case 1 of Elith et al., 2011. A total of 150 models were generated, which correspond to ten species times three bias scenarios times five autocorrelation correction treatments. When dealing with data from atlas databases, randomly partitioning occurrence data into training and test sets and using cross-validation techniques is often the only solution available to calibrate and test a model. We used 10-fold cross-validation and then used the average of all models as the final one. As a goodness-of-fit measure, we used the test AUC. As it is usually the norm in species distribution modelling, we accepted only models with an AUC $\geq 0.7$. Models with an AUC $\geq 0.9$ are considered excellent (Swets, 1988). As mentioned earlier, we only accepted models with no residual autocorrelation as tested by Moran’s I autocorrelation coefficient.

Upscaling and validation at a coarser scale

We assumed that habitat quality is related to probability of presence and upscaled each accepted model at 1-km resolution (AUC$_{1K}$) to 10-km resolution by a basic calculation of probabilities (see equation below). We computed the probability of presence for each $i_{th}$ 10 km square of the study area ($P_{10km,i}$), given the predicted probability of presence for each 1 km square contained within it ($P_{1km,i}$). If we subtract this probability from 1, we obtain the probability of absence for this $j_{th}$ 1 km square. For a given $i_{th}$ 10 km square to have an absence, all of its one hundred 1 km squares need also to be absences. Therefore, by multiplying the probabilities of absence for each $j_{th}$ 1 km square, we get the probability of absence for the $i_{th}$ 10 km square. Finally, by subtracting the probability of absence for an $i_{th}$ 10 km square from 1, we get its probability of presence ($P_{10km,i}$).

$$\forall i, P_{10km,j} = 1 - \prod_{j=1}^{100} (1 - P_{1km,j})$$

We then performed a receiver operating characteristic (ROC) analysis [ROCR package in R (Sing et al., 2009)] and computed the AUC$_{10K}$ value for each upscaled 10 km model using the independent data set at 10-km resolution. To ensure accurate prediction assessment, independent test sets should be available (Loiselle et al., 2008; Veloz, 2009). Again, those models with an AUC$_{10K}$ value $\geq 0.7$ were accepted. Finally, of all models accepted for each species, we selected the one with the highest AUC$_{1K}$ value at 1-km resolution as the best one. In summary, we obtained a set of distribution maps that perform well at the finer resolution and that also acceptably predict independent records at the coarser resolution. We think these models can be considered robust and reliable given the data available.

RESULTS

Overall, AUC test values at 1-km resolution (AUC$_{1K}$) ranged from as low as 0.37 to as high as 0.96 (including models with residual autocorrelation), while their corresponding upscaled models at 10-km resolution ranged from 0.45 to 0.93 (Table 3). Of 150 models, 101 (67%) had an AUC$_{1K}$ $\geq 0.7$. Of these, only 20 showed no significant residual SAC (Moran’s $P$-value from Monte-Carlo simulation $\geq 0.05$). The 20 that performed well at 1-km resolution are shown in Table 3. AUC$_{1K}$ test values for the accepted 20 models ranged from 0.7 to 0.94 and correspond to seven of the ten modelled species. The other three, Amaranthus albus, Conyza canadensis and Datura stramonium, did not perform well when modelling at 1-km resolution. Oxalis pes-caprae had the highest number of acceptable models at 1-km resolution but, nevertheless, unacceptable models predominated (10 of 15). The rest had between 11 and 14 unacceptable models. The worst models, those with an AUC$_{1K}$ $\leq 0.5$, were four models of Amaranthus albus and one of Datura stramonium. All of these models used the invasive1k bias treatment.

When evaluating the performance at 10-km resolution, 12 of these final 20 models (60%) had an AUC$_{10K}$ $\geq 0.7$ and were considered acceptable distribution models given the data available (see Table 3). Models marked with an asterisk correspond to our best models (see Table 3 and Fig. 3); that is, those with the maximum AUC$_{1K}$ value, when more than one model per species was accepted.

Half of the 12 models finally accepted required no bias treatment, while the other half performed better when a bias treatment was applied, although only one of them showed preference for the background offered by IAS citation areas. With respect to autocorrelation treatment, six performed better with some sort of SAC correction, while five did so with environmental autocorrelation correction. One model needed no autocorrelation correction, while none seemed to prefer the environmental correction with the shortest distance, and finally, only one model did not need either bias or autocorrelation treatment, which corresponded to Oxalis pes-caprae. This model also coincides with the best one of all, although care should be taken when comparing AUC values between species (Lobo et al., 2008; Blach-Overgaard et al., 2010). See Table 4 for a summary.

Three species, Ailanthus altissima, Oenothera biennis and Robinia pseudoacacia, did not pass the cross-scale validation cut (see Table 3). They had models that were acceptable at
1-km resolution but which, once scaled, did not offer acceptable predictive power at 10 km. Thus, their finer resolution models were discarded as not robust enough: that is, they could not explain the independent data set at 10-km resolution. As an example, Fig. 4 shows two models that, while having passed the cut at 1-km resolution modelling, show an AUC\textsubscript{10K} around 0.5 that is not better than random.

On a per species basis, Agave americana performed well under the whole\_area and vasculars1k bias treatments and for SAC correction with a minimum distance of 4250 m. Its best model was the one with the bias treatment whole\_area. Opuntia ficus-indica performed well under whole\_area and vasculars1k bias treatments and under both environmental and spatial occurrence filtering, its best model being the one with the whole\_area bias treatment and a SAC correction with a minimum distance of 2830 m. Oxalis pes-caprae performed well under all three bias treatments and under both spatial and environmental autocorrelation correction. Its best model required no bias or autocorrelation treatment at all. Finally, for Xanthium spinosum, the only successful treatment was the vasculars1k bias treatment and the environmental autocorrelation correction with a minimum distance of 0.1. Reductions in the number of available occurrences after autocorrelation correction for the final four best models were

![Figure 3](image1.png)

**Figure 3** Best models per species among all the accepted models. Only four species resulted in finally valid models at 1-km resolution. For these species, those shown in the figure are the ones with max (AUC\textsubscript{1K}). Legend scale ranges from 1.0 (maximum suitability) to 0.0 (no suitability). Black empty squares represent records of presence at 10-km resolution.

![Figure 4](image2.png)

**Figure 4** Examples of models that did not work. Even though these two models had a high AUC\textsubscript{1K} value and showed no residual autocorrelation, they had an AUC\textsubscript{10K} close to 0.5 and are thus not better than random. Legend scale ranges from 1.0 (maximum suitability) to 0.0 (no suitability). Black empty squares represent records of presence at 10-km resolution.
as follows: *Agave americana* from 20 to 12, *Opuntia ficus-indica* from 13 to 10, *Xanthium spinosum* from 56 to 22 and no reduction for *Oxalis pes-caprae* because its best model was the one without autocorrelation correction.

Table 4 presents a summary of accepted and discarded models according to treatments. For SAC treatments, the option of no treatment was the worst (93.3% of these models showed residual SAC), while the best was the treatment corresponding to an environmental distance (Gower’s index) of 0.1 (less than half 43.3%) of the models showed residual SAC). The rest of SAC treatments had similar results: only a quarter to a fifth of the models showed no residual SAC. For bias treatments, using some kind of treatment worked better (68% for *invasive1k* and 64% for *vasculars1k*) than no treatment (86%). The best treatments for removing residual SAC were the combination of an environmental distance of 0.1 with some bias treatment (*invasive1k* or *vasculars1k*). As expected, the number of models with residual SAC is inversely proportional to the intensity of the SAC treatment applied.

**DISCUSSION**

Our results show that species distribution maps derived from presence-only records held in biodiversity databases or atlases should be used with caution. Apparently, high scores in predictive power from species distributions can be obtained from scarce, biased and autocorrelated presence records using modern tools such as MaxEnt. However, our work shows that these results can be misleading when confronted with independent data at different scales. Other authors have reached similar conclusions (Wisz et al., 2008). If the distribution of a species was well-known at two different scales, these should necessarily be coherent with one another. To generate reliable fine-resolution distribution maps, these need to be in accordance across scales (Niamir et al., 2011). For a given species, its real distribution map at a fine scale should match its real distribution map at a coarser scale once upscaled. This seems not to be the case for some species, indicating that either the modelled distributions at fine resolution are wrong or that the known distributions at the coarser scale are, in fact, incomplete (unlikely for a well-surveyed region for vascular plants such as our study area). Therefore, if one accepts this assumption, our results suggest that distribution maps at the finer scales are not as good as they appear to be. However, this could also be due to the fact that our models have been built without explanatory variables which can account for other environmental factors and biotic interactions (e.g. interspecific competition), thus not reflecting the realized niche of the species. Although at macroecological scales climate is the main factor affecting species distributions, biotic interactions may also play a role (Araújo & Luoto, 2007; Heikkinen et al., 2007; Kissling et al., 2010). Such variables, if available, could positively affect our models and make them more in accordance with well-known distributions at coarser scales. Unless this problem can be solved, if these models are used for decision-making in conservation, they may not always accomplish the objectives for which they are meant.

Of the 20 models that performed well at 1-km resolution, only 12 were coherent with data at 10-km resolution. Species-wise, it might seem that good fine-scale predictive maps could be derived from the biodiversity database for seven species. However, fine-scale distribution maps were in accordance with their coarser scale data for only four of them (Table 3). We can thus consider the fine-resolution maps for these four species to be sufficiently reliable for biodiversity conservation. Coarse resolution data do not often match the requirements of conservation planning (Araújo et al., 2005), but, when these data are assumed to reflect the distribution of the species at the coarse scale, they can be used to make a cross-scale validation of modelled fine-scaled distribution maps, even if high predictive scores had been obtained. The resulting maps will be much more reliable and robust and will help decision-makers to better meet their conservation goals.
Atlas data commonly suffer from bias and autocorrelation problems (Robertson et al., 2010). Treating both hindrances is paramount to developing robust and reliable species distribution models (Segurado et al., 2006; Merckx et al., 2011). We tried several bias and autocorrelation correction scenarios but did not find any particular pattern in our results that can help in establishing protocols for distribution modelling. Although the overall outcome of the modelling process can only be known on a case-by-case basis, with respect to the correction of residual SAC, our results suggest that the use of data filtering (either by setting minimum environmental or spatial distances) helps to remove residual SAC: the larger the minimum distance, the better. Also, to a lesser extent, bias treatments help to remove it. Except for Oxalis pes-caprae, for which the environmental predictors used were able to remove residual SAC without applying any treatment to the occurrence data, the rest of the species needed some sort of occurrence filtering to remove it. This suggests that, for the other species in this study, some missing explanatory variables (environmental or biotic), if available, may have resulted in models with a better fit and no residual SAC without having to discard occurrences to build it.

The lesson that can be learnt from this study is to be sceptical of fine-resolution maps obtained when modelling species distributions from scarce and biased data, even though they score high when measuring their predictive power. Given that current modelling techniques and computer power allow us to run many models per species taking into account different scenarios of bias and autocorrelation correction treatments, it is always sensible to do so and check, on a case-by-case basis, which one works best. Having a set of independent data to validate the model seems indispensable. Cross-scale validation, when possible, is a good solution to produce reliable and robust maps which can then be used to make better conservation decisions.

CONCLUSIONS

Casual observations at fine resolution in biodiversity atlases or databases have the potential to generate continuous species distribution maps through species distribution modelling, providing powerful tools for conservation management and planning. However, it may be impossible to obtain them for many species because of scarcity and strong biases in the data. Although high validation scores can be obtained when modelling such data, there is a risk that the distribution maps reflect the data distribution rather than the true species distribution. Cross-scale validation of the data with species distribution information at a coarser scale appears to be a consistent protocol to test the validity and robustness of fine-resolution models and thus to make them much more reliable for decision-makers in conservation.

Fine-resolution maps can be derived from biodiversity atlases with data scarcity, bias and autocorrelation. Bias, and specially autocorrelation, treatments can help to remove residual SAC in models. However, if no independent set of data is available to further validate them, results should be viewed with caution. When good coarser scale data are available, cross-scale validation appears to be effective in assessing the robustness of the models. When these options are not available, new field surveys may be the only option if reliable fine-scale maps are needed.

ACKNOWLEDGEMENTS

The authors would like to thank three anonymous reviewers for their valuable comments that helped to improve the final manuscript. We also thank Jane Elith for her insightful comments on a late version of the manuscript, Roberto Molowny for discussion on the upscaling method, Xavier Font for his insights on the data, and Arthur Chapman and Robert Guralnick for their early comments and encouragement to carry out this work. This study was supported by project CONSOLIDER-MONTES (CSD2008-00040) of the Ministerio de Ciencia e Innovación of the Spanish government. The study is also a contribution to the FP7-226852 EU-project SCALES. Xavier Pons is recipient of an ICREA Academia Excellence in research grant 2011–2015.

REFERENCES


**SUPPORTING INFORMATION**

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

**Appendix S1 Maps of occurrences.**

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

**BIOSKETCHES**

Arnald Marcer is a researcher at the Center for Ecological Research and Forestry Applications (CREAF), Autonomous University of Barcelona. His research interests are in the fields of biodiversity conservation and modelling, and environmental information handling and management through the development of information systems.

Author contributions: A.M. conceived the ideas, analysed the data and wrote the manuscript. L.B., J.P. and X.P helped to refine the ideas and methodology and to revise the final manuscript. The authors’ research groups and interests can be found at the following hyperlinks:

http://www.grumets.uab.cat/index_en.htm

http://biodiversitylandscapeecologylab.blogspot.com/

Editor: Mark Robertson