Title: A habitat class to land cover translation model for mapping Area of Habitat of terrestrial vertebrates

Authors: Maria Lumbierres[1,3], Prabhat Raj Dahal[1,3], Moreno Di Marco[2], Stuart H. M. Butchart[3,4], Paul F. Donald[3,4], Carlo Rondinini[1]

Author's address

1- Global Mammal Assessment Program, Department of Biology and Biotechnologies, Sapienza University of Rome, Viale dell'Università 32, 00185 Rome, Italy

2- Department of Biology and Biotechnologies, Sapienza University of Rome, Viale dell'Università 32, 00185 Rome, Italy

3- BirdLife International, David Attenborough Building, Pembroke Street, Cambridge CB2 3QZ, UK

4- Conservation Science Group, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

Abstract: Area of Habitat (AOH) is defined as the 'habitat available to a species, that is, habitat within its range' and is produced by subtracting areas of unsuitable land cover and elevation from the range. Habitat associations are documented using the IUCN Habitats Classification Scheme, and unvalidated expert opinion has been used so far to match habitat to land-cover classes generating a source of uncertainty in AOH maps. We develop a data-driven method to translate IUCN habitat classes to land-cover based on point locality data for 6,986 species of terrestrial mammals, birds, amphibians and reptiles. We extracted the land-cover class at each point locality and matched it to the IUCN habitat class(es) assigned to each species occurring there. Then we modelled each land cover class as a function of IUCN habitat using logistic regression models. The resulting odds ratios were used to assess the strength of the association of each habitat land-cover class. We then compared the performance of our data-driven model with those from a published expert knowledge translation table. The results show that some habitats (e.g. forest and desert) could be more confidently assigned to land-cover classes than others (e.g. wetlands and artificial). We calculated the association between habitat classes and land-cover classes as a continuous variable, but to map AOH, which is in the form of a binary presence/absence , it is necessary to apply a threshold of association. This can be chosen by the user according to the required balance between omission and commission errors. We demonstrate that a data-driven translation model and expert knowledge perform equally well, but the model provides greater standardization, objectivity and repeatability. Furthermore, this approach allows greater flexibility in the use of the results and allows uncertainty to be quantified. Our model can be developed regionally or for different taxonomic groups.

Keywords: Habitat suitability models, commission and omission errors, Copernicus Global Land Service Land Cover (CGLS-LC100), ESA Climate Change Initiative (ESA-CCI), IUCN habitat, IUCN Red List.

INTRODUCTION

Habitat loss is the most important driver of biodiversity decline (Díaz et al.,, 2019). Therefore, there is an urgent need to determine where habitat is located within each species' distribution (Pimm et al., 2014; Brooks et al., 2019). Several approaches have been developed to map global species' distributions, but accurate spatial data are only available for a limited number of species (Rondinini et al., 2005; Rondinini & Boitani 2012).

The most complete dataset of maps of species' ranges is that available in the International Union for Conservation of Nature (IUCN) Red List (www.iucnredlist.org). The IUCN Red List has assessed more than 134,400 species, and species groups, including mammals, amphibians, and birds, have been comprehensively assessed. The IUCN range maps are generally drawn to minimize errors of omission (i.e. false absence), with the result that they often contain substantial areas that are not occupied by the species, and so suffer from errors of commission (i.e. false presence) (Ficetola et al., 2014; Di Marco et al., 2017).

Area of Habitat (AOH; previously known as extent of suitable habitat, or ESH) is the ‘habitat available to a species, that is, habitat within its range’ (Brooks et al., 2019). AOH maps are produced by subtracting unsuitable areas from range maps, using data on each species’ associations with land cover and altitude (Beresford et al., 2011; Rondinini et al., 2011; Ficetola et al., 2015), and attempts to reduce commission errors in range maps. Therefore, the production of AOH maps requires an understanding of which habitats a species occurs in and where those habitats are located within its range.

Information on habitat preferences is documented for each species assessed on the IUCN Red List (IUCN 2013) following the IUCN Habitats Classification Scheme (IUCN habitat; IUCN, 2012), a classification and coding system of habitats that ensures global consistency. IUCN standardized habitat definitions independently of taxonomy or geography. However, IUCN habitat classes are not spatially explicit, although recent efforts have attempted to delimit them (Jung et al., 2020). Land-cover classes derived from remote sensing have been widely used as a surrogate of habitat (e.g. Buchanan et al., 2008; Beresford et al., 2011; Rondinini et al., 2011; Tomaselli et al., 2013; Montesino Pouzols et al., 2014; Corbane et al., 2015; Santini et al., 2019), although habitat is a complex multi-dimensional concept that is difficult to simplify into land-cover classes.

A translation table between habitat and land-cover classes is typically used to represent IUCN habitat classes spatially and to produce AOH maps. This is a table that shows which habitat classes map onto which land-cover classes. Previous versions have been based solely on expert knowledge, raising concerns about the accuracy and objectivity of the resulting associations, as the assumptions generated in the translation process are rarely considered in detail and the errors are difficult or impossible to quantify (Bradley et al., 2012). Furthermore, there is a lack of standardization in the procedure (Seoane et al., 2005), which is subject to variability in expert opinion (Johnson & Gillingham, 2004).

Repositories of point locality data (i.e. locational records where particular species have been recorded (Rondinini et al., 2006)) primarily from citizen science have been successfully used in habitat suitability models (e.g. Gueta & Carmel, 2016; Bradter et al.,, 2018; Crawford, Olds, Maerz, & Moore, 2020). The potential, therefore, exists to use such data also to develop an objective data-driven translation table between habitat and land-cover classes by extracting information on land cover from point localities of species with different habitat associations.

Here we propose a standardized, data-driven methodology to produce a translation table between IUCN habitat classes and two widely used global land-cover maps, the Copernicus Global Land Service Land Cover (CGLS-LC100; Copernicus Global Land Operations “Vegetation and Energy”, 2018a; Buchhorn et al., 2019) and the European Space Agency Climate Change Initiative land cover 2015 (ESA-CCI; ESA, 2017) using point locality data for mammals, birds, amphibians and reptiles (the best-documented groups of species). The aim of this analysis was to develop a translation table that quantifies the power of association between land cover and habitat classes. In doing so, we aim to illustrate a method that improves on expert opinion by (a) quantifying errors in associations between habitat and land cover classes, (b) being flexible to the needs of the user in terms of the required trade-off between reducing commission errors and increasing omission errors, and (c) can be developed at different spatial scales, for different taxa, using any set of habitat or land-cover classes.

METHODS

Data cleaning and preparation

We downloaded point locality data for mammals (GBIF, 2019; GBIF, 2020), amphibians (GBIF, 2020) and reptiles (GBIF, 2020) from the Global Biodiversity Information Facility (GBIF) and for birds from GBIF (GBIF, 2019; GBIF, 2020) and eBird (eBird Basic Dataset, 2019). The data were restricted to point localities dated from January 2005 to December 2018 for the model building (70% training and 30% test), and from January 2019 to December 2020 for the evaluation of the model. For eBird data, we selected only stationary point localities with a coordinate uncertainty of less than 30 m. To minimize errors, and uncertainties inherent to repositories of point locality data, we included only the most precisely georeferenced points (Rondinini et al., 2006; Meyer 2012) and applied a set of filters following the guidelines of Boitani et al., (2011). The main attributes considered were currency, spatial accuracy and spatial coverage (Fig. 1).

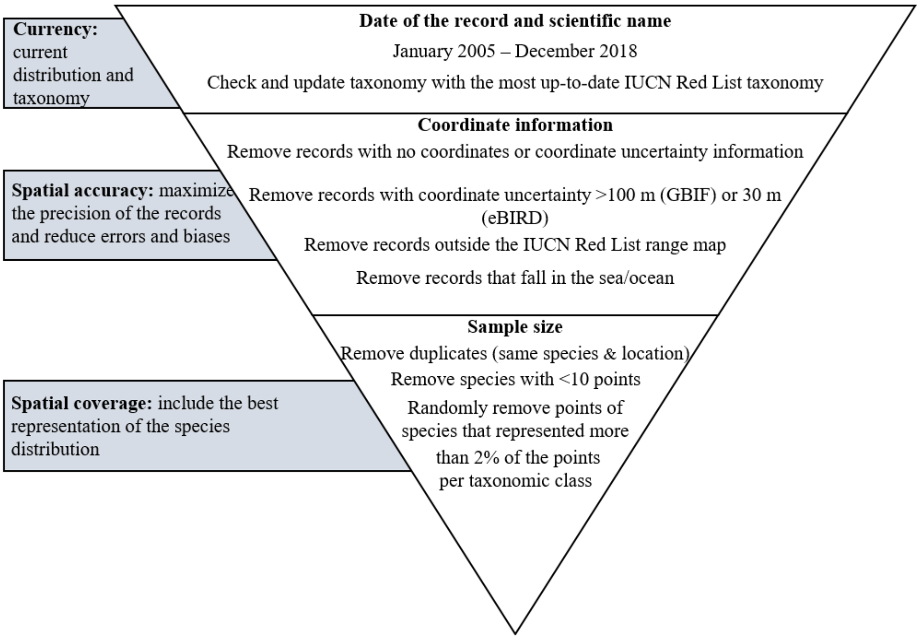

Currency: current distribution and taxonomy
Date of the record and scientific name
January 2005 – December 2018
Check and update taxonomy with the most up-to-date IUCN Red List taxonomy
Coordinate information
Remove records with no coordinates or coordinate uncertainty information
Spatial accuracy: maximize the precision of the records and reduce errors and biases
Remove records with coordinate uncertainty >100 m (GBIF) or 30 m (eBIRD)
Remove records outside the IUCN Red List range map
Remove records that fall in the sea/ocean
Sample size
Remove duplicates (same species & location)
Remove species with <10 points
Randomly remove points of species that represented more than 2% of the points per taxonomic class
Spatial coverage: include the best representation of the species distribution

Figure 1 . Description of the point locality cleaning process, following Boitani et al., (2011). The factors considered were currency, spatial accuracy and spatial coverage. The filters were applied from top to bottom.

To make it clear where we are referring to explicit classes, we present land-cover class names in quotation marks and IUCN habitat class names in italics.

The habitat class(es) association of each species were extracted from the IUCN habitat (IUCN, 2020). These follow a hierarchical classification of habitat with three levels. The definitions consider biogeography, latitudinal zonation and depth in marine systems. In this analysis, we used Level-1 habitat classes for all habitats except for artificial terrestrial, for which we used a modification of Level-2 (Appendix S1). We subdivided artificial terrestrial into three subclasses because in terms of land cover these are distinct habitat classes that could aggregate different species (Ducatez et al., 2018).

Because the land-cover classes from the two remote sensing products are exclusively terrestrial, we limited the analysis to species coded only to terrestrial habitat classes, thus excluding species coded to one or more IUCN marine habitats. We also excluded species coded to more than five Level-1 habitat classes, because habitat generalists are likely to add little information to the habitat-land cover relationship. In contrast, specialist species coded to only one habitat class provide more insight into the relationship between habitat and land cover. For that reason, for each taxonomic class, we randomly subsampled point records from species coded to more than one habitat class to match the number of points of species coded to one habitat and thereby gave equal weight to habitat specialists even when they had fewer points.

We developed models for two different global land-cover products derived from remote sensing: CGLS-LC100 and ESA-CCI. The CGLS-LC100 has a 100-m spatial resolution and a global classification accuracy of 80.2% (Copernicus Global Land Operations "Vegetation and Energy", 2018b). The ESA-CCI has a 300-m spatial resolution and a global classification accuracy of 71.1% (ESA 2017). It is part of a time series from 1992 to 2015, of which we used the 2015 map. Both products use the United Nations Food and Agriculture

Organization Land Cover Classification System (UN-LCCS), although they have different legends. CGLS-LC100 has 12 land-cover classes at Level-1 and 23 classes at Level-3 (Level-2 is not used by CLGS-LC100), and we used Level-3. ESA-CCI has 22 land-cover classes at Level-1 and 38 classes at Level-2. In this analysis, we only used Level-1 because Level-2 is only available for some regions of the globe.

To prepare the data for the model, we extracted the land-cover class at the coordinates of each point locality. Some land-cover classes did not have enough point localities falling within them to be modeled, although in all cases these were land-cover classes with very low global coverage. For CGLS-LC100, the under-represented land-cover classes were “open forest deciduous needle leaf” (10 points, 0.03% of global land surface), “snow and ice” (108 points, 3.1% of global land surface), “moss and lichen” (124 points, 2.3% of global land surface) and “closed forest deciduous needle leaf” (383 points, 3.0% of global land surface). For ESA-CCI, the only class represented too infrequently for analysis was “lichens and mosses” (713 points, 2.2% of global land surface).

Modeling of habitat-land cover associations

To quantify the relationship between IUCN habitat classes and land-cover classes, we modeled the presence or absence of each land-cover class as a function of the IUCN habitat class(es) of the species whose point localities fell within it. An important consideration for modeling was that the number of habitat classes per species varied from one to five. Therefore, it was impossible to model land-cover class as a one-to-one relationship with habitat classes, as each point location was associated with one or multiple habitats. This consideration restricted the number of models we could use for our analysis. We required a flexible model that allowed a many-to-many match between habitat classes and land-cover classes to model this matrix of habitat vs land-cover class relations. In multinomial logistic regression models, the data and the computational power requirements increase exponentially with the number of response categories. In our case, with more than 20 land-cover categories, this option was not feasible. Therefore, we modeled each land-cover class separately, transforming it into a binary variable of 1 or 0 (land cover present/not present in a point locality). Then, we used logistic regressions to model the binary land-cover class variable as a function of the different habitat classes:

$$[1] log\frac{p_{lc}}{1-p_{lc}} = \beta_0 + \beta_1 H_{Forest} + \beta_2 H_{Savanna} + \beta_3 H_{Shrubland} + \beta_4 H_{Grassland} + \beta_5 H_{Wetlands} + \beta_6 H_{RockyAreas} + \beta_7 H_{Desert} + \beta_8 H_{Artificial1} + \beta_9 H_{Artificial2} + \beta_{10} H_{Artificial3} + \beta_{11} H_{Artificial4}$$

where ($p_{lc}/(1-p_{lc})$) is the land cover odds ratio and β_x are the model parameters for each of the habitats H_x.

The transformation of the land-cover class into a binary form for each of the models generated a highly unbalanced variable, with many more zeroes than ones. In a logistic regression model, unbalanced data underestimate the probability of an event so it is recommended to adjust the number of 1s and 0s (King & Zeng 2001; Pozzolo et al., 2015). We therefore randomly subsampled the 0s in the training set before running the model. The assumption behind this is that in the majority class there are many redundant observations and randomly removing some of them does not change the estimation of the within-class distribution (Pozzolo et al., 2015).

To reduce the intrinsic spatial and taxonomic bias point locality data (Boitani et al., 2011; Meyer et al., 2016), and to account for multiple but varying numbers of point localities per species, we added taxonomic and spatial variables as random effects in the model (Bird et al., 2014). As taxonomic variables, we added species nested within taxonomic class (Amphibia, Reptilia, Aves, Mammalia). Adding intermediate taxonomic groupings (e.g. family or genus) in the nesting would result in many factor levels with single or very few replicates. To test whether there was any bias between taxonomic classes, we first produced separate models for each class, and found that the association between land cover and habitat classes from the different translation tables were very similar; therefore, we decided to model all classes together. As a spatial variable, we added the country of the point record as a random effect.

We used the coefficients of the models to evaluate the association between land-cover class and habitat classes. The intercept did not provide any information on the relationship between land-cover class and habitat class as it represents the odds of a point locality falling within a particular land-cover class after the subsampling of the data set, independently of the habitat (Ranganathan et al., 2017). The coefficients represent the odds ratio, in other words, the odds of a point locality falling in a particular land-cover class (when the species to which the point locality relates is coded for a particular habitat class) divided by the odds of the species occurring in that land-cover class when it is not coded for that habitat class. The ratio, therefore, indicates the extent to which being coded to a particular habitat class increases or decreases the odds of a species being found in a particular land-cover class. The units of the logit function are log(odds ratio), but for easier interpretation, we exponentiated them and present the results as odds ratios.

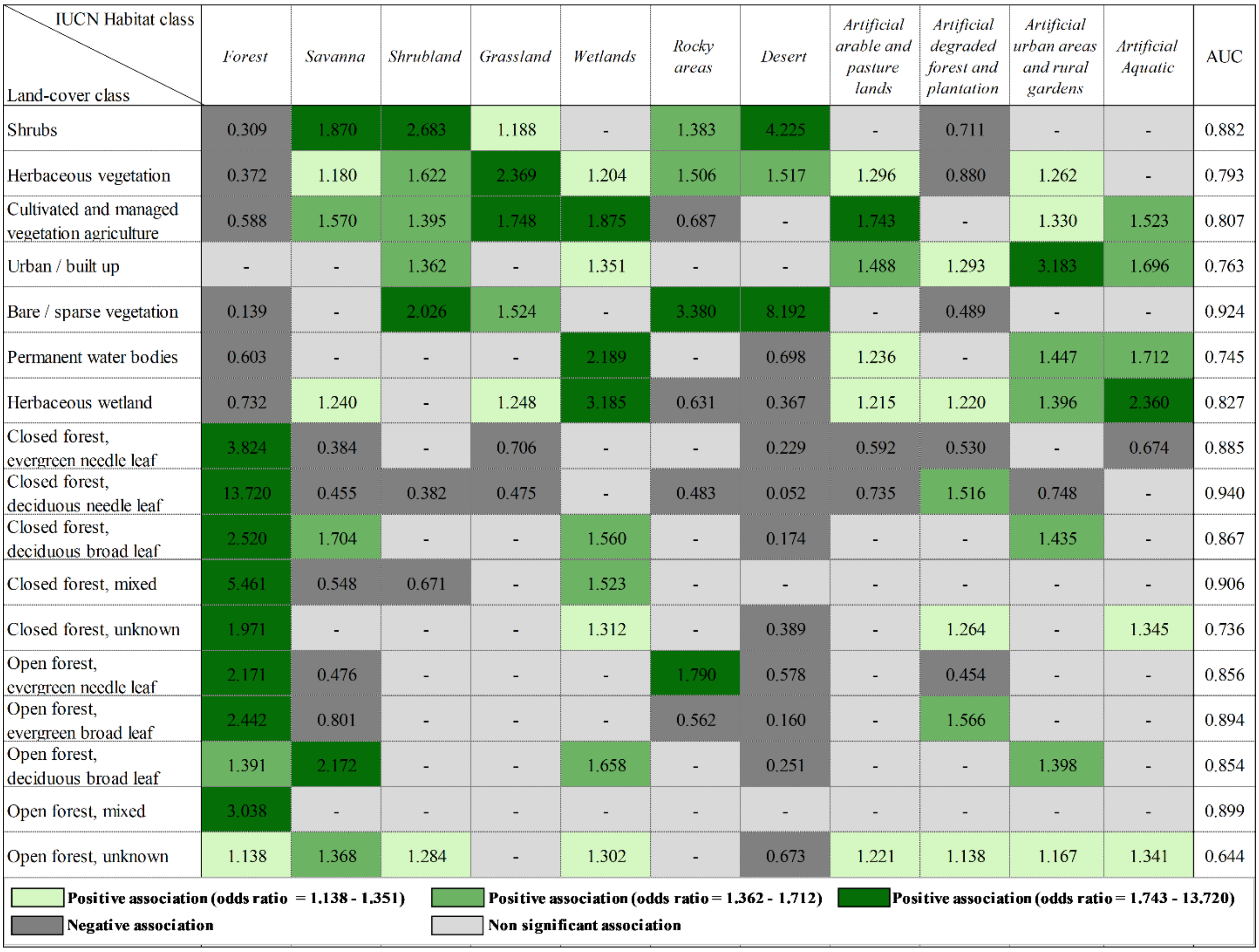
Odds ratio values below 1 indicate a negative association between land cover and habitat classes, while those above 1 indicate a positive association. As the odds ratio is a continuous variable, it is necessary to set a threshold to transform the results into a binary translation table that can be used to assign, or not, a particular habitat class to a particular land-cover class. The threshold can be modified according to the needs of the user based on the required balance between minimizing commission errors (land-cover classes incorrectly associated with a habitat class) and increasing omission errors (land-cover classes incorrectly omitted from a habitat class). Coefficients that had *p*-values higher than 0.05 were considered to indicate a lack of association between land cover and habitat classes. To adjust the significance threshold of the *p*-values for multivariable analysis, we used Bonferroni corrections.

To validate the models, we set aside 30% of the point occurrence data for testing, leaving 70% to train the model. As a validation test, we used the Area Under the Curve (AUC) from a Receiver Operating Characteristic (ROC) curve. AUC is a model accuracy measure that provides information on how well a model can distinguish among classes. In our case, we used it to test how well the models predicted the presence/absence of a point locally in a given land cover class. AUC values range from 0 to 1, a value of 0.5 means that the model does not performs better than random, while a value of 1 indicates that the model can perfectly separate the two groups.

We then compared the performance of the data-driven translation table with that of an expert knowledge translation table (Santini et al., 2019) based on the same ESA-CCI land cover classification used here. We did not find any published translation table that used CGLS-LC100. Santini et al., matched the ESA CCI land cover classes against Level-2 IUCN habitat classes, so we aggregated the habitat classes to Level-1 IUCN habitat classes to make the two translation tables comparable. We limited the comparison to birds and mammals because they were the taxonomic groups considered by Santini et al., For each species we mapped suitable habitat based on both tables. We assessed the proportion of points located in the suitable habitat (point prevalence) and compared it with the proportion of suitable habitat inside the species' range (model prevalence) to determine whether the results were better than a randomly assigned set of points (Rondinini et al., 2011).

RESULTS

The number of point localities and species available for this analysis was 200,683 and 455 respectively for mammals, 4,083,510 and 5,154 for birds, 92,327 and 479 for amphibians, and 131,077 and 898 for reptiles. For the CGLS-LC100 land-cover product, 71 coefficients showed a significantly positive association (odds ratio >1) and 38 coefficients showed a significantly negative association (odds ratio <1) between land-cover classes and habitat classes (Fig. 2). For the ESA-CCI land-cover product, 101 coefficients showed a significantly positive association and 40 coefficients showed a significantly negative association (Fig. 3).



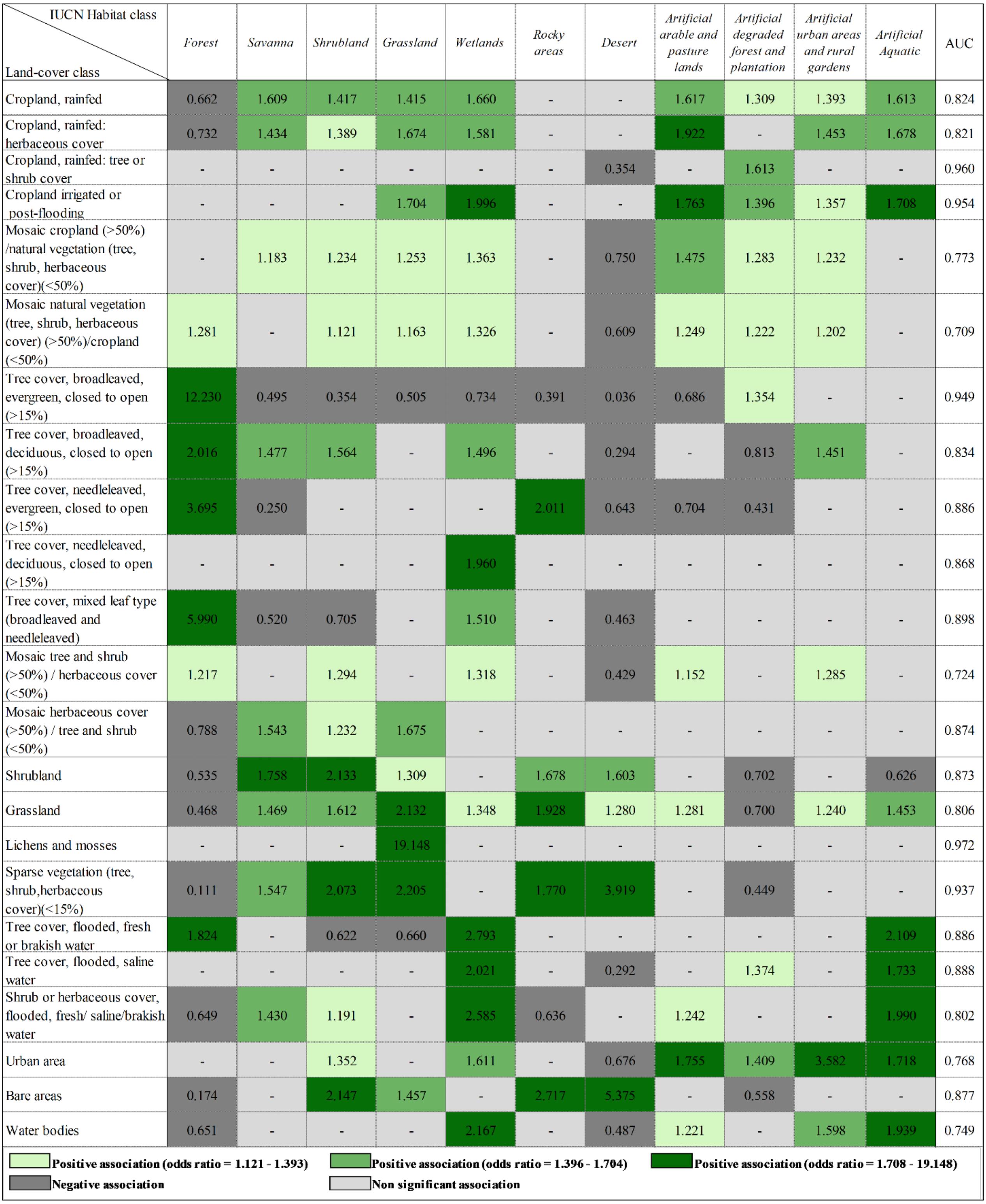
IUCN Habitat class / Land-cover class	*Forest*	*Savanna*	*Shrubland*	*Grassland*	*Wetlands*	*Rocky areas*	*Desert*	*Artificial arable and pasture lands*	*Artificial degraded forest and plantation*	*Artificial urban areas and rural gardens*	*Artificial Aquatic*	AUC
Shrubs	0.309	1.870	2.683	1.188	-	1.383	4.225	-	0.711	-	-	0.882
Herbaceous vegetation	0.372	1.180	1.622	2.369	1.204	1.506	1.517	1.296	0.880	1.262	-	0.793
Cultivated and managed vegetation agriculture	0.588	1.570	1.395	1.748	1.875	0.687	-	1.743	-	1.330	1.523	0.807
Urban / built up	-	-	1.362	-	1.351	-	-	1.488	1.293	3.183	1.696	0.763
Bare / sparse vegetation	0.139	-	2.026	1.524	-	3.380	8.192	-	0.489	-	-	0.924
Permanent water bodies	0.603	-	-	-	2.189	-	0.698	1.236	-	1.447	1.712	0.745
Herbaceous wetland	0.732	1.240	-	1.248	3.185	0.631	0.367	1.215	1.220	1.396	2.360	0.827
Closed forest, evergreen needle leaf	3.824	0.384	-	0.706	-	-	0.229	0.592	0.530	-	0.674	0.885
Closed forest, deciduous needle leaf	13.720	0.455	0.382	0.475	-	0.483	0.052	0.735	1.516	0.748	-	0.940
Closed forest, deciduous broad leaf	2.520	1.704	-	-	1.560	-	0.174	-	-	1.435	-	0.867
Closed forest, mixed	5.461	0.548	0.671	-	1.523	-	-	-	-	-	-	0.906
Closed forest, unknown	1.971	-	-	-	1.312	-	0.389	-	1.264	-	1.345	0.736
Open forest, evergreen needle leaf	2.171	0.476	-	-	-	1.790	0.578	-	0.454	-	-	0.856
Open forest, evergreen broad leaf	2.442	0.801	-	-	-	0.562	0.160	-	1.566	-	-	0.894
Open forest, deciduous broad leaf	1.391	2.172	-	-	1.658	-	0.251	-	-	1.398	-	0.854
Open forest, mixed	3.038	-	-	-	-	-	-	-	-	-	-	0.899
Open forest, unknown	1.138	1.368	1.284	-	1.302	-	0.673	1.221	1.138	1.167	1.341	0.644

🟩 **Positive association (odds ratio = 1.138 - 1.351)** 🟩 **Positive association (odds ratio = 1.362 - 1.712)** 🟩 **Positive association (odds ratio = 1.743 - 13.720)**
■ **Negative association** □ **Non significant association**

Figure 2. Odds ratio values describing the association between CGLS-LC100 land-cover classes and IUCN habitat classes. Odds ratio values significantly < 1 indicate a negative association, and values significantly > 1 indicate a positive association. The significantly positive associations are divided into tertiles (shown in shades of green), indicating three possible options for setting a threshold to convert continuous variables into a binary association/non-association variable for creating AOH maps. AUC indicates the values of Area Under the Curve from a ROC, a measure of accuracy of a classification model.

Higher odds ratios (>1) indicated stronger positive associations between land cover and habitat classes, and lower odds ratios (nearer to zero) indicated stronger negative associations. We divided the significantly positive values into tertiles to identify three potential thresholds for creating a table of binary association/non-association variables for producing AOH maps: 1.138-1.351, 1.362-1.712 and 1.743-13.720 for CGLS-LC100, and 1.121-1.393, 1.396-1.704 and 1.708-19.148 for ESA-CCI.

Forest and *Desert* had the strongest positive associations between land cover and habitat classes. The *Forest* habitat class was associated with almost all the forest and tree cover land-cover classes (CGLS-LC100 average positive odds ratio = 3.8; ESA-CCI average positive odds ratio = 4.0) and with no other land-cover classes. The *desert* habitat class was also strongly associated with particular land-cover classes: “shrubs”, “herbaceous vegetation”, and “bare/sparse vegetation” in CGLS-LC100 (average positive odds ratio = 4.6) and “shrubland”, “grassland”, “sparse vegetation (tree, shrub, herbaceous cover < 15%)” and “bare areas” in ESA-CCI (average positive odds ratio = 3.0). *Rocky areas* were associated with almost the same land-cover classes as *Desert* but with lower odds ratios.



IUCN Habitat class Land-cover class	Forest	Savanna	Shrubland	Grassland	Wetlands	Rocky areas	Desert	Artificial arable and pasture lands	Artificial degraded forest and plantation	Artificial urban areas and rural gardens	Artificial Aquatic	AUC
Cropland, rainfed	0.662	1.609	1.417	1.415	1.660	-	-	1.617	1.309	1.393	1.613	0.824
Cropland, rainfed: herbaceous cover	0.732	1.434	1.389	1.674	1.581	-	-	1.922	-	1.453	1.678	0.821
Cropland, rainfed: tree or shrub cover	-	-	-	-	-	-	0.354	-	1.613	-	-	0.960
Cropland irrigated or post-flooding	-	-	-	1.704	1.996	-	-	1.763	1.396	1.357	1.708	0.954
Mosaic cropland (>50%) /natural vegetation (tree, shrub, herbaceous cover)(<50%)	-	1.183	1.234	1.253	1.363	-	0.750	1.475	1.283	1.232	-	0.773
Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%)/cropland (<50%)	1.281	-	1.121	1.163	1.326	-	0.609	1.249	1.222	1.202	-	0.709
Tree cover, broadleaved, evergreen, closed to open (>15%)	12.230	0.495	0.354	0.505	0.734	0.391	0.036	0.686	1.354	-	-	0.949
Tree cover, broadleaved, deciduous, closed to open (>15%)	2.016	1.477	1.564	-	1.496	-	0.294	-	0.813	1.451	-	0.834
Tree cover, needleleaved, evergreen, closed to open (>15%)	3.695	0.250	-	-	-	2.011	0.643	0.704	0.431	-	-	0.886
Tree cover, needleleaved, deciduous, closed to open (>15%)	-	-	-	-	1.960	-	-	-	-	-	-	0.868
Tree cover, mixed leaf type (broadleaved and needleleaved)	5.990	0.520	0.705	-	1.510	-	0.463	-	-	-	-	0.898
Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	1.217	-	1.294	-	1.318	-	0.429	1.152	-	1.285	-	0.724
Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	0.788	1.543	1.232	1.675	-	-	-	-	-	-	-	0.874
Shrubland	0.535	1.758	2.133	1.309	-	1.678	1.603	-	0.702	-	0.626	0.873
Grassland	0.468	1.469	1.612	2.132	1.348	1.928	1.280	1.281	0.700	1.240	1.453	0.806
Lichens and mosses	-	-	-	19.148	-	-	-	-	-	-	-	0.972
Sparse vegetation (tree, shrub,herbaceous cover)(<15%)	0.111	1.547	2.073	2.205	-	1.770	3.919	-	0.449	-	-	0.937
Tree cover, flooded, fresh or brakish water	1.824	-	0.622	0.660	2.793	-	-	-	-	-	2.109	0.886
Tree cover, flooded, saline water	-	-	-	-	2.021	-	0.292	-	1.374	-	1.733	0.888
Shrub or herbaceous cover, flooded, fresh/ saline/brakish water	0.649	1.430	1.191	-	2.585	0.636	-	1.242	-	-	1.990	0.802
Urban area	-	-	1.352	-	1.611	-	0.676	1.755	1.409	3.582	1.718	0.768
Bare areas	0.174	-	2.147	1.457	-	2.717	5.375	-	0.558	-	-	0.877
Water bodies	0.651	-	-	-	2.167	-	0.487	1.221	-	1.598	1.939	0.749

Positive association (odds ratio = 1.121 - 1.393) **Positive association (odds ratio = 1.396 - 1.704)** **Positive association (odds ratio = 1.708 - 19.148)**
Negative association **Non significant association**

Figure 3. Odds ratio values describing the association between ESA-CCI land-cover classes and IUCN habitat classes. Odds ratio values significantly < 1 indicate a negative association, values significantly > 1 indicate a positive association. The positive associations are divided into tertiles (shown in shades of green), indicating three possible options for setting a threshold to convert continuous variables into a binary association/non-association variable for creating AOH maps. AUC indicates the values of Area Under the Curve from a ROC, a measure of accuracy of a classification model.

Savanna, *Shrubland* and *Grassland* habitat classes were associated with “shrubs”, “herbaceous vegetation” and “cultivated and managed vegetation agriculture” in CGLS-LC100 land cover, and “cropland”, “herbaceous cover”, “shrubland”, “grassland”, “sparse vegetation”, “mosaic cropland” and “mosaic herbaceous cover” in ESA-CCI. However, the power of association varied between these different combinations. The *Savanna* habitat class was also associated with some forest classes while *Shrubland* and *Grassland* habitats were also associated with bare areas.

We divided artificial terrestrial habitats into three different classes: *Artificial arable and pasture lands*, *Artificial degraded forest and plantations*, and *Artificial urban and rural gardens*. These habitats had the least certain relationships because the odds ratio values were the closest to 1 (CGLS-LC100 average positive odds ratio = 1.367, 1.333 and 1.577 respectively; ESA-CCI average positive odds ratio = 1.468, 1.370 and 1.579 respectively). Some unexpected land-cover classes were associated with these habitat classes, e.g. *Arable and pasture lands* and *degraded forest and plantations* were associated with “urban areas”. However, these unexpected associations disappeared when increasing the threshold.

Wetland and *Artificial aquatic* habitats had intermediate odds ratio values (CGLS-LC100 average positive odds ratio = 1.7; ESA-CCI average positive odds ratio = 1.8). In terms of land-cover associations, they were associated (in some cases strongly) with land-cover classes related to water, but also to some land-cover classes that have no relation with wetlands or aquatic environments (e.g. some type of forest or cultivated areas).

The AUC of models for CGLS-LC100 ranged from 0.644 to 0.940. The land-cover classes with the lowest AUC were the “open and closed unknown forest” (AUC = 0.644 and 0.736) classes, followed by “water bodies” (AUC = 0.745) and “urban areas” (AUC = 0.763). Those with the highest AUC values were the other forest classes (AUC range 0.854 – 0.940) and “bare and sparse vegetation” (AUC = 0.924). For ESA-CCI, the AUC ranged from 0.709 to 0.972. The land-cover classes with the lowest AUC were mosaic land-cover classes (AUC range 0.709 - 0.874), followed by “water bodies” (AUC = 0.750) and “urban areas” (AUC = 0.768). The land-covers with the highest AUC values were “lichens and mosses” (AUC = 0.972), “cropland irrigated or post-flooding” (AUC = 0.954), “sparse vegetation” (AUC = 0.937) and tree cover land classes (AUC range 0.834 – 0.949).

The results of the models can also be mapped spatially (Fig. 4) using one of the three thresholds of associations between habitat and land-cover classes. In such maps, habitats are overlaid because the same land-cover class may represent more than one habitat class and/or because both habitats occur in the same geographical areas. The overlap among habitats increases as the threshold of association is reduced.

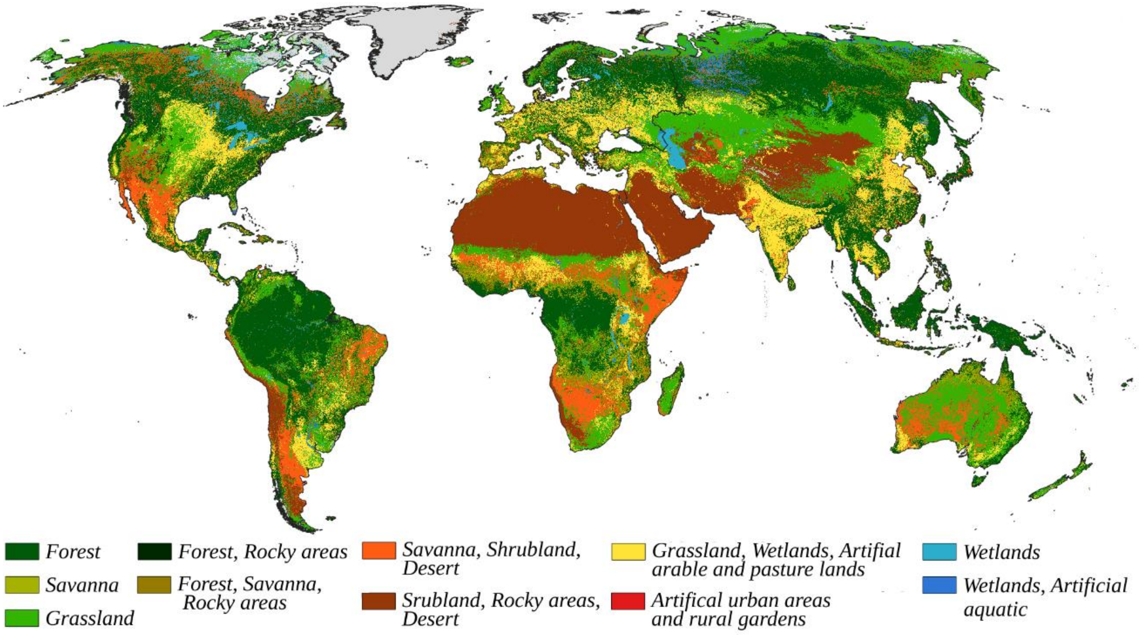

Forest
Forest, Rocky areas
Savanna, Shrubland, Desert
Grassland, Wetlands, Artifial arable and pasture lands
Wetlands
Savanna
Forest, Savanna, Rocky areas
Srubland, Rocky areas, Desert
Artifical urban areas and rural gardens
Wetlands, Artificial aquatic
Grassland

Figure 4 Map of habitat classes (Level 1) from the IUCN Habitat classification scheme based on the highest threshold for CGLS-LC100 data-derived translation (Fig 2) .

To compare the performance of the data-driven table and the expert-derived table, we used 211,304 point localities for 489 species of mammal and 461,277 point localities for 2,112 species of bird. We compared the point prevalence (the proportion of georeferenced points falling in the land-cover classes assigned by each translation table according to the habitat of each species) between our data-driven method and the expert based assessment of Santini et al., and found that point prevalence in Santini et al., (2019) was similar to the point prevalence we found from our model when using the middle and high odd-ratio thresholds (Table 1). The ratio between point prevalence and model prevalence (the proportion of the range remaining after apparently unsuitable land cover classes are excluded) between the two methods was also very similar, and higher than 1, indicating that the habitat associations were better than random for both approaches.

Table 1. Mean point prevalence and model prevalence for birds and mammals using the three tertile thresholds for ESA CCI land cover derived from data-driven assessment (see Figure 4) and the expert knowledge-based assessment of Santini et al., (2019).

	Lower terti17le threshold	*Middle tertile threshold*	*Upper tertile threshold*	*Santini et al., (2019)*
		BIRDS		
Point prevalence	0.94	0.81	0.66	0.74
Model prevalence	0.91	0.76	0.59	0.68
		MAMMALS		
Point prevalence	0.93	0.82	0.67	0.73
Model prevalence	0.90	0.80	0.62	0.70

DISCUSSION

By modeling the relationship between IUCN habitat classes and the CGLS-LC100 and ESA-CCI land-cover classes, we generated two translation tables, quantifying the strength of association between habitat and land cover classes. The strength of association is represented by the odd ratio values, which indicate the extent to which a species being coded to a particular habitat class increases or decreases the odds of that species being found in a particular land-cover class. The relationship between IUCN habitat classes and land cover classes is expressed as a continuous variable.

Among habitat classes, *Forest*, *Desert* and *Rocky areas* have the strongest associations with land-cover classes, perhaps owing to the higher accuracy of the relevant land-cover classes. For both CGLS-LC100 and ESA-CCI, the highest classification accuracy classes are “forest”, “tree cover areas” and “bare soil”. Using a different approach based on a decision tree, Jung et al., (2020) found that *Forest* has the highest validation accuracy, although they obtained lower validation accuracy for *Rocky areas* and *Desert* habitat classes.

On the other hand, *Wetlands* and *Artificial* habitats are more difficult to represent using land-cover maps. Wetland-related land-cover classes have the lowest classification accuracy in both land-cover maps. From a remote sensing perspective, wetlands are difficult to map because they are highly dynamic, with rapid phenological changes through the year (Gallant, 2015; Lumbierres et al., 2017). Remote sensing products at a global scale cannot distinguish small ponds or temporary water bodies (Pekel et al., 2016; Klein et al., 2017). Therefore, wetland land-cover classes have more omission errors, and this has a direct impact on the results of our model.

Artificial land-cover classes are also difficult to map, as they tend to be more heterogeneous (Álvarez-Martínez et al., 2018), producing misclassifications among land-cover classes. Land-cover maps have difficulty separating artificial land-cover classes from natural ecosystems, e.g., plantation from forest, grassland from cropland, or lake from reservoir (Álvarez-Martínez et al., 2018). Overall, species richness and average abundance are often lower in artificial environments than in their natural equivalent, even if there is variation across different biogeographical contexts (Barlow et al., 2007; Newbold et al., 2015) and this introduces commission errors. Moreover, we found that artificial land-cover are associated with some natural habitat classes. This is likely a consequence of greater accessibility of these habitats, and hence disproportionate prevalence in citizen science data (Meyer et al., 2015). Because a high proportion of citizen science point location data are recorded in artificial land-cover classes, there is an increased probability that species primarily associated with natural habitats are reported there, so a data-driven method may associate some natural habitats with artificial land-cover classes.

There are several differences between the two land-cover layers used to produce the translation tables that could determine the use of the table. CGLS-LC100 has a resolution of 100 m while ESA-CCI has a coarser resolution of 300 m, also CGLS-LC100 has an overall classification accuracy of 80.2% compared with 71.1% for ESA-CCI. Moreover, CGLS-LC100 avoids using mosaics classes and in general, mapping less complex habitats is easier than more heterogeneous habitats (Corbane et al., 2015; Álvarez-Martínez et al., 2018). However, ESA-CCI has the advantage of being available as a longer time series, 1992-2020 for ESA-CCI vs 2015-2019 for CGLS-LC100, which allows studying habitat changes. For both land cover maps we excluded some land-cover classes because of the lack of point localities; we recommend adding these land cover classes manually when using the translation tables, according to the user's needs.

The coding of habitats to each species on the IUCN Red List could introduce some noise to the modeling process. Coding is based on qualitative assessment by experts, and is therefore susceptible to individual biases (Brooks et al., 2019; Santini et al., 2019). The current version of the Habitat Classification Scheme on the IUCN website is described as a draft version. We, therefore, recommend that IUCN updates and improves this document and anticipate this would influence our odds ratio estimates.

Both the data-driven table and the expert knowledge translation table represented land cover distribution inside the range better than random. However, our data-driven approach presents several advantages compared to an expert knowledge approach. We present the relationship between IUCN habitat and land cover classes as a continuous variable, allowing greater flexibility in using the results. To use the model to produce AOH maps, the user is able to decide a threshold of association to transform the results into a binary table according to the required balance between omission and commission errors. Moreover, a data-driven approach allows us to quantify the uncertainty associated with the habitat to land cover association and could help to evaluate potential uncertainties in the AOH maps. This approach can be used to develop a translation table between any set of habitat codes and any set of land cover variables at a global or regional scale. As better data (species point and land-cover maps) become available, the translation table can be improved, assuring objectivity, standardization, and repeatability.

ACKNOWLEDGEMENTS

This research is part of the Inspire4Nature Innovative Training Network, funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 766417. MDM acknowledges support from the MIUR Rita Levi MontalcinLevii programme.

AUTHOR CONTRIBUTIONS

ML, CR, PFD, and SHMB conceived the study. ML and PFD develop the analysis. ML led the writing of the manuscript. All authors contributed to drafts and gave final approval for publication.

REFERENCES

Álvarez-Martínez JM, Jiménez-Alfaro B, Barquín J, Ondiviela B, Recio M, Silió-Calzada A, Juanes JA. 2018. Modelling the area of occupancy of habitat types with remote sensing. Methods in Ecology and

Evolution **9**:580–593.

Barlow J et al., 2007. Quantifying the biodiversity value of tropical primary, secondary, and plantation forests. Proceedings of the National Academy of Sciences of the United States of America **104**:18555–18560.

Beresford AE, Buchanan GM, Donald PF, Butchart SHM, Fishpool LDC, Rondinini C. 2011. Poor overlap between the distribution of Protected Areas and globally threatened birds in Africa. Animal Conservation **14**:99–107.

Bird TJ et al., 2014. Statistical solutions for error and bias in global citizen science datasets. Biological Conservation **173**:144–154. Elsevier Ltd. Available from http://dx.doi.org/10.1016/j.biocon.2013.07.037.

Boitani L, Maiorano L, Baisero D, Falcucci A, Visconti P, Rondinini C. 2011. What spatial data do we need to develop global mammal conservation strategies? Philosophical Transactions of the Royal Society B: Biological Sciences **366**:2623–2632.

Bradley BA, Olsson AD, Wang O, Dickson BG, Pelech L, Sesnie SE, Zachmann LJ. 2012. Species detection vs. habitat suitability: Are we biasing habitat suitability models with remotely sensed data? Ecological Modelling **244**:57–64. Elsevier B.V. Available from http://dx.doi.org/10.1016/j.ecolmodel.2012.06.019.

Bradter U, Mair L, Jönsson M, Knape J, Singer A, Snäll T. 2018. Can opportunistically collected Citizen Science data fill a data gap for habitat suitability models of less common species? Methods in Ecology and Evolution **9**:1667–1678.

Brooks TM et al., 2019. Measuring Terrestrial Area of Habitat (AOH) and Its Utility for the IUCN Red List. Trends in Ecology & Evolution **xx**:1–10. The Authors. Available from https://linkinghub.elsevier.com/retrieve/pii/S0169534719301892.

Buchanan GM, Butchart SHM, Dutson G, Pilgrim JD, Steininger MK, Bishop KD, Mayaux P. 2008. Using remote sensing to inform conservation status assessment: Estimates of recent deforestation rates on New Britain and the impacts upon endemic birds. Biological Conservation **141**:56–66.

Buchhorn M, Smets B, Bertels L, Lesiv M, Tsendbazar N-E, Herold M, Fritz SA. 2019. Copernicus Global Land Service: Land Cover 100m: epoch 2015: Globe. Dataset of the global component of the Copernicus Land Monitoring Service le.

Copernicus Global Land Operations “Vegetation and Energy.” 2018a. PRODUCT USER MANUAL, MODERATE DYNAMIC LAND COVER 100M VERSION 2.

Copernicus Global Land Operations “Vegetation and Energy.” 2018b. VALIDATION REPORT MODERATE DYNAMIC LAND COVER COLLECTION 100M VERSION 2.0:1–44.

Corbane C, Lang S, Pipkins K, Alleaume S, Deshayes M, García Millán VE, Strasser T, Vanden Borre J, Toon S, Michael F. 2015. Remote sensing for mapping natural habitats and their conservation status - New opportunities and challenges. International Journal of Applied Earth Observation and Geoinformation **37**:7–16. Elsevier B.V. Available from http://dx.doi.org/10.1016/j.jag.2014.11.005.

Crawford BA, Olds MJ, Maerz JC, Moore CT. 2020. Estimating population persistence for at-risk species using citizen science data. Biological Conservation **243**:108489. Elsevier. Available from https://doi.org/10.1016/j.biocon.2020.108489.

Di Marco M, Watson JEM, Possingham HP, Venter O. 2017. Limitations and trade-offs in the use of species distribution maps for protected area planning. Journal of Applied Ecology **54**:402–411.

Díaz S et al., 2019. Pervasive human-driven decline of life on Earth points to the need for transformative change. Science **366**.

Ducatez S, Sayol F, Sol D, Lefebvre L. 2018. Are urban vertebrates city specialists, artificial habitat exploiters, or environmental generalists? Integrative and Comparative Biology **58**:929–938.

eBird Basic Dataset. Version: EBD_January-2019. Cornell Lab of Ornithology, Ithaca, New York. 2019.

ESA. 2017. Land Cover CCI Product User Guide Version 2. Tech. Rep. Available at: maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf, Paria, France. Available from http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf.

Ficetola GF, Rondinini C, Bonardi A, Baisero D, Padoa-Schioppa E. 2015. Habitat availability for amphibians and extinction threat: A global analysis. Diversity and Distributions **21**:302–311.

Ficetola GF, Rondinini C, Bonardi A, Katariya V, Padoa-Schioppa E, Angulo A. 2014. An evaluation of the robustness of global amphibian range maps. Journal of Biogeography **41**:211–221.

Gallant AL. 2015. The challenges of remote monitoring of wetlands. Remote Sensing **7**:10938–10950.

GBIF.org (04 March 2020) GBIF Occurrence Download. 2020. Available from https://doi.org/10.15468/dl.tvtiqq.

GBIF.org (14 January 2019) GBIF Occurrence Download. 2019. Available from https://doi.org/10.15468/dl.tk87g2.

GBIF.org (23 December 2020) GBIF Occurrence Download. 2020. Available from https://doi.org/10.15468/dl.swey54.

GBIF.org (24 February 2020) GBIF Occurrence Download. 2020. Available from https://doi.org/10.15468/dl.5vqa7s.

GBIF.org (26 January 2019) GBIF Occurrence Download. 2019. Available from https://doi.org/10.15468/dl.8bfl5p.

Gueta T, Carmel Y. 2016. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. Ecological Informatics **34**:139–145. Elsevier B.V. Available from http://dx.doi.org/10.1016/j.ecoinf.2016.06.001.

IUCN. 2012. Habitats Classification Scheme (Version 3.1).

IUCN. 2013. Documentation standards and consistency checks for IUCN Red List assessments and species accounts. Version 2. Available from http://www.iucnredlist.org/documents/RL_Standards_Consistency.pdf.

IUCN. 2020. The IUCN Red List of Threatened Species. Version 2020-2.

Johnson CJ, Gillingham MP. 2004. Mapping uncertainty: Sensitivity of wildlife habitat ratings to expert opinion. Journal of Applied Ecology **41**:1032–1041.

Jung M, Dahal PR, Butchart SHM, Donald PF, De Lamo X, Lesiv M, Kapos V, Rondinini C, Visconti P. 2020. A global map of terrestrial habitat types. Scientific Data **7**:1–8. Springer US. Available from http://dx.doi.org/10.1038/s41597-020-00599-8.

King G, Zeng L. 2001. Logistic Regression in Rare. Political Analysis **9**:137–163.

Klein I, Gessner U, Dietz AJ, Kuenzer C. 2017. Global WaterPack – A 250 m resolution dataset revealing the daily dynamics of global inland water bodies. Remote Sensing of Environment **198**:345–362. Elsevier Inc. Available from http://dx.doi.org/10.1016/j.rse.2017.06.045.

Lumbierres M, Méndez PF, Bustamante J, Soriguer R, Santamaría L. 2017. Modeling biomass production in seasonal wetlands using MODIS NDVI land surface phenology. Remote Sensing **9**:1–18.

Meyer C. 2012. Limitations in global information on species occurrences. Frontiers of Biogeography **4**:217–220.

Meyer C, Jetz W, Guralnick RP, Fritz SA, Kreft H. 2016. Range geometry and socio-economics dominate species-level biases in occurrence information:1181–1193.

Meyer C, Kreft H, Guralnick RP, Jetz W. 2015. Global priorities for an effective information basis of biodiversity distributions. Nature Communications **6**:1–8. Nature Publishing Group. Available from http://dx.doi.org/10.1038/ncomms9221.

Montesino Pouzols F, Toivonen T, Di Minin E, Kukkala AS, Kullberg P, Kuusterä J, Lehtomäki J, Tenkanen H, Verburg PH, Moilanen A. 2014. Global protected area expansion is compromised by projected land-use and parochialism. Nature **516**:383–386. Nature Publishing Group. Available from http://www.nature.com/articles/nature14032 (accessed October 1, 2019).

Newbold T et al., 2015. Global effects of land use on local terrestrial biodiversity. Nature **520**:45–50.

Pekel JF, Cottam A, Gorelick N, Belward AS. 2016. High-resolution mapping of global surface water and its long-term changes. Nature **540**:418–422. Nature Publishing Group. Available from http://dx.doi.org/10.1038/nature20584.

Pimm SL, Jenkins CN, Abell R, Brooks TM, Gittleman JL, Joppa LN, Raven PH, Roberts CM, Sexton JO. 2014. The biodiversity of species and their rates of extinction, distribution, and protection. Science

344:1246752–1246752. Available from http://www.sciencemag.org/cgi/doi/10.1126/science.1246752 (accessed November 20, 2019).

Pozzolo AD, Caelen O, Johnson RA, Bontempi G, Dal Pozzolo A, Caelen O, Bontempi G, Johnson RA. 2015. Calibrating Probability with Undersampling for Unbalanced Classification. 2015 IEEE Symposium Series on Computational Intelligence:159–166. IEEE. Available from https://www.researchgate.net/publication/283349138 (accessed May 28, 2019).

Ranganathan P, Pramesh C, Aggarwal R. 2017. Common pitfalls in statistical analysis: Logistic regression. Perspectives in Clinical Research **8**:148–151. Medknow Publications.

Rondinini C et al., 2011. Global habitat suitability models of terrestrial mammals. Philosophical Transactions of the Royal Society B: Biological Sciences **366**:2633–2641.

Rondinini C, Boitani L. 2012. Mind the map: trips and pitfalls in making and reading maps of carnivore distribution. Pages 31–46 Carnivore Ecology and Conservation A Handbook of Techniques.

Rondinini C, Stuart S, Boitani L. 2005. Habitat suitability models and the shortfall in conservation planning for African vertebrates. Conservation Biology **19**:1488–1497.

Rondinini C, Wilson KA, Boitani L, Grantham H, Possingham HP. 2006. Tradeoffs of different types of species occurrence data for use in systematic conservation planning. Ecology Letters **9**:1136–1145.

Santini L, Butchart SHM, Rondinini C, Benítez-López A, Hilbers JP, Schipper AM, Cengic M, Tobias JA, Huijbregts MAJ. 2019. Applying habitat and population-density models to land-cover time series to inform IUCN Red List assessments. Conservation Biology **00**:1–10.

Seoane J, Bustamante J, Díaz-Delgado R. 2005. Effect of expert opinion on the predictive ability of environmental models of bird distribution. Conservation Biology **19**:512–522.

Tomaselli V et al., 2013. Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: A Mediterranean assessment. Landscape Ecology **28**:905–930.