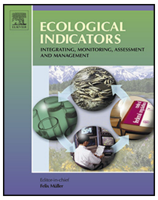
Ecological Indicators 74 (2017) 526–529

Contents lists available at ScienceDirect

Ecological Indicators

journal homepage: www.elsevier.com/locate/ecolind



Short communication

Which breeding bird categories should we use in models of species distribution?

Vítězslav Moudrý*, Jan Komárek, Petra Šímová

Department of Applied Geoinformatics and Spatial Planning, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamycka 129, 165 21 Prague 6, Czechia

ARTICLE INFO

Article history:
Received 28 February 2016
Received in revised form 29 October 2016
Accepted 2 November 2016
Available online 11 November 2016

Keywords:
Ambiguity
Atlas
Calibration
Species distribution model
Validation
Wetland birds

ABSTRACT

Studies often use breeding bird atlases to assess species' habitat requirements or to estimate species' potential distribution under environmental changes. In breeding bird atlases, one of the attributes recorded for each grid square is evidence of breeding. The attribute represent probability of breeding (confirmed, probable, possible) categorized according to breeding behaviour. However, the majority of studies often make arbitrary decisions on which category to use. This may have severe consequences for results. This study evaluated whether models' discrimination ability change by inclusion of ambiguous breeding categories (probable, possible). We compared models' predictions for distribution of nine wetland birds derived from Atlas of the breeding distribution of birds in the Czech Republic. For each species, we developed generalized linear models using combinations of the breeding categories as input to model calibration and validation. Our results show that the discrimination ability (AUC) decreased in most cases when all breeding categories were uncritically used in calibration and validation process. On the other hand, however, inclusion of probable and possible breeding categories to model calibration did not affect models' abilities to predict confirmed presences and absences. This implies that inclusion of ambiguous breeding categories has more serious impact on models' performance when added to validation than to calibration data set. We advocate for more rigorous use of different breeding categories and emphasize that widely used atlases from citizen science programmes offer more than simple occurrence data. Additional attributes (e.g. breeding category, sampling effort) should be used to select high quality data to validate the models.

1. Introduction

The description of the relationship between species and their environment has long been a focus for ecology and biogeography. Reliable descriptions of species' habitat requirements and distribution are fundamental for their conservation and as such should be explicitly accompanied by uncertainty estimates (Rocchini et al., 2011). In general, uncertainty arises from necessary simplification of the complex, continuous nature of the real world into discrete representation in spatial databases. Species distribution models (SDM) use species occurrence data and environmental data in order to produce a set of rules that identify and scale the environmental space where species were observed. These three elements (species occurrence, environmental data and method) are the main sources of uncertainty in SDM.

While the influence of different methods on SDMs' performance has been widely studied, the importance of different datasets has been recognized only more recently (e.g. Syphard and Franklin, 2009; Mateo-Tomás and Olea, 2015). The inclusion of uncertainty in distributional data has been emphasized by a few reviews (Moudrý and Šímová, 2012; Rocchini et al., 2011) and a few studies have examined the influence of survey methods on the output of SDMs (Bino et al., 2014; Monk et al., 2012; Tulowiecki, 2014). In a recent study, Duputié et al. (2014) stated that ecologists are in an awkward position due to the lack of the accurate species distribution data which is necessary to calibrate and validate the sophisticated SDMs. They further suggested that citizen science programmes, among other sources, could be a way to acquire such data.

Undoubtedly, the most widely researched and understood animal group is birds, due to the traditional popularity of recreational bird-watching and coordinated participatory science programmes over large areas which generally result in publication of a book

---

* Corresponding author at: Department of Applied Geoinformatics and Spatial Planning, Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamycka 129, 165 21 Prague 6, Czechia.
E-mail address: moudry@fzp.czu.cz (V. Moudrý).

http://dx.doi.org/10.1016/j.ecolind.2016.11.006

or electronic atlas. Broad-scale bird monitoring projects are the longest-running and largest citizen science programmes (see Gibbons et al., 2007 for review) and play an important role in ecology (Robertson et al., 2010). Breeding bird atlases exist for regions (e.g. Atlas of Breeding Birds of Wallonia; Jacob et al., 2010), countries (e.g. Atlas of Breeding Bird Distribution in the Czech Republic; Šťastný et al., 2006), and continents (e.g. EBCC Atlas of European Breeding Birds; Hagemeijer and Blair, 1997).

In recent study on cetaceans, Rayment et al. (2015) highlighted the need to account for reproductive status to refine species distribution models. Surprisingly, studies using breeding bird atlases have often neglected this important attribute—breeding categories. In breeding bird atlases, one of the attributes recorded for each grid square is evidence of breeding. Bird species breeding status is usually recorded as: 0—Non-breeding, A—Possible breeding (e.g. singing male present in breeding season), B—Probable breeding (e.g. pair observed in suitable nesting habitat in breeding season, agitated behaviour or anxiety calls from adults, bird observed building a nest), or C—Confirmed breeding (e.g. used nest or eggshells found, recently fledged young, nest containing eggs). However, it is common practice for authors arbitrary to select the breeding categories to use in their study without providing any justification. For example, Virkkala et al. (2014), Šímová et al. (2015) and Russell et al. (2015) deemed all three categories as presence, while Beale et al. (2008) and Moudrý and Šímová (2013) deemed only the probable and confirmed categories as presence. Importantly, both approaches are problematic due to model calibration and validation on ambiguous data (possible and probable categories can represent either presence or absence). In theory, however, SDMs should only be calibrated and validated using true presences and true absences. Evidently, the question of which breeding category to use in a study as presence is a trade-off between data quality and quantity (e.g. Nichols et al., 2007).

Our objective in this study is to examine following questions: (1) Are there differences in predictive performance among models when ambiguous breeding categories (possible and probable breeding) are used uncritically in model calibration and validation process? (2) Does inclusion of ambiguous breeding categories to model calibration influence ability to predict confirmed presences and absences?

2. Materials and methods

2.1. Study area and species distribution data

The study area encompassed the Czech Republic, a central European country. More specifically, it comprised a delineated area of almost 79,000 km^2 divided into 678 grid squares of 10′ east longitude × 6′ north latitude (hereafter referred to as mapping squares) to which biological and environmental data are referred.

For this study, data on bird species were obtained from the Third Atlas of Breeding Bird Distribution in the Czech Republic (Šťastný et al., 2006). The fieldwork for the atlas was conducted between 2001 and 2003, and field observations of the bird species occurring in each mapping square were recorded using 17 numerical breeding codes. Breeding occurrence of each bird species within a given mapping square was given in one of three categories: possible breeding, probable breeding, and confirmed breeding (Hagemeijer and Blair, 1997).

This study focused on nine migratory bird species (Table 1) that choose a similar nesting environment of standing water, and in particular a littoral zone of ponds, swamps, and other wetlands or habitats surrounding wetlands. The main criterion for species selection was to include species with different numbers of occupied squares within each category. At the same time, we endeavoured to include species from different ecological groups.

Table 1
Species under study.

Latin name	English name	Arrival period
Tachybaptus ruficollis	Little Grebe	March–April
Podiceps nigricollis	Black-necked Grebe	March–April
Ixobrychus minutus	Little Bittern	April–May
Anas strepera	Gadwall	March–April
Anas crecca	Common Teal	March–April
Anas clypeata	Northern Shoveler	March–April
Netta rufina	Red-crested Pochard	February–March
Bucephala clangula	Common Goldeneye	February–March
Rallus aquaticus	Water Rail	March–April

2.2. Environmental data

Appropriate selection of explanatory variables is critical to avoid under-specified models or over-fitting (Williams et al., 2012). For broad-scale species distribution modelling, in Europe, habitat variables are often derived from the national Corine Land Cover databases (Feranec et al., 2010). Considering the importance of water habitats for the studied species, we defined two variables representing area of water bodies (Corine 5.1.2) and wetlands (Corine 4.1.1. and 4.1.2.) within mapping squares. Other land cover variables were: area of agricultural areas (Corine 2) and area of the forest and semi-natural areas (Corine 3). We also included variable representing the area of artificial surfaces (Corine 1).

Current climatic data were downloaded from the widely used WorldClim database (Hijmans et al., 2005). Following earlier studies (Moudrý and Šímová, 2013; Virkkala et al., 2013), we used mean temperature and mean precipitation according to the arrival dates of each bird species (Table 1). Variables were downloaded at a resolution of 5′ ($\sim$ 10 km^2) for current conditions (1950–2000) and then averaged inside each mapping square to match the same grid format as species distribution data. All geodata were processed using ArcGIS 10.3 (ESRI, CA, USA).

2.3. Scenarios and statistical analysis

Based on commonly adopted schemes in the literature (Virkkala et al., 2014; Šímová et al., 2015; Russell et al., 2015) we assumed that species in all three categories can be considered as breeding presence. To answer the first question, we developed three scenarios differing according to breeding categories used in both model calibration and validation processes. For the first scenario (S1) we selected only confirmed breeding category and absences and randomly divided them into k independent partitions. We used $k-1$ of the partitions to calibrate the model, and evaluated it on the left-out partition (we used $k=5$). For the second scenario (S2) we randomly divided probable breeding category into k independent partitions and added each partition to its equivalent of the first scenario. Thus, we still had k number of partitions, but with higher number of presences. For the third scenario (S3) we repeated the latter process with possible breeding category (Table 2). As already mentioned, however, such approach (S2, S3) is problematic (although often adopted) because developed models are evaluated using ambiguous breeding categories (probable and possible breeding), which may lead to spurious conclusions about model performance (see Foody, 2011). Thus, we further examined whether inclusion of probable and possible occurrences in addition to confirmed occurrences lead to models that better discriminate between confirmed presences and absences. To assess the effect of probable and possible occurrences on model calibration we only added $k-1$ partitions, first the former (S4) and then the latter (S5), into $k-1$ calibration