The Effects of False Positive Errors in Species Occurrence Data on the Performance of Species Distribution Models: Case Study on the Wild Duck (*Anas platyrhynchos*)

Vítězslav Moudrý, Petra Šímová
Czech University of Life Sciences

Abstract

In ecological research, particularly in species distribution modelling (SDM), the uncertainty resulting from data deficiencies is of recent interest. Beside other things, it includes positional accuracy of species occurrence data. In our study, we investigated the influence of the positional error, caused by false positive detection of species, in species occurrence data on the performance of the models. We used occurrences of Wild Duck (*Anas platyrhynchos*) mapped in fine scale resolution (300 x 300 meters) and habitat variables derived from the Base map of the Czech Republic. Generalized additive models (GAM) with a stepwise selection procedure were used to select relevant habitat variables. Model performance was evaluated using area under the receiver operating characteristics curve (AUC), sensitivity and specificity. Incorporated positional error led to a reduction in model prediction accuracy, although not enough to reject the models.

Keywords: spatial uncertainty, positional accuracy, SDM (species distribution modelling)

Introduction

With consideration to the broad use of GIS tools in many research topics, spatial uncertainty of geodata is an important issue (Heuvelink 1998; Shi et al. 2002). Although only factors as, for example, geodata availability and price are often considered for selecting inputs for particular research, many recent studies have focused on influence of uncertainty resulting from data deficiencies on results of analyses (e.g. Barry and Elith 2006). In ecological research, particularly in species distribution modelling (SDM), the problem of uncertainty includes, beside other things, positional accuracy of species occurrence

data (see Moudrý and Šímová 2012 for review). Species occurrence data are increasingly available thanks to data sharing activities (e.g. www.gbif.org), however its positional accuracy can be influenced by a variety of factors, including field mapping method, inaccuracy in the measurement of location or georeferencing error (e.g. Graham et al. 2008).

For the mapping of bird assemblages on the fine scale (extent of tens of square kilometers), the ornithological point sampling method (Bibby et al. 2000) is often used. Sampling points create a network of 300 × 300 meters and all birds which are heard or seen in the vicinity of the point are localized to the point. Although this method is commonly used in ornithological research (Ralph et al. 1995) and gives ecologically reasonable results (e.g. López and Moro 1997; Davis 2004), it is a data source prone to both false negative and false positive errors. False negative detections occur because it is generally impossible to detect every individual within a sampled area, whereas false positive detections occur when species that are absent are erroneously detected (e.g. commonly heard species). For example, in a study by Miller et al. (2012) 8.1% of recorded occurrences were due to false positive error. It is a question, whether the false positive errors can influence the performance of species distribution model.

The effect of positional error in species occurrence data on the performance of species distribution models is of recent interest (Graham et al. 2008; Osborne and Leitao 2009). However, to our best knowledge, there is no study dealing with this topic at fine scale resolution. For our case study, we selected the Wild Duck (*Anas platyrhynchos*), because it is easily detectable and thus the data are less prone to false positive and false negative errors. We investigated the influence of positional error, caused by false positive detection of species, in species occurrence data on the performance of SDM at 300 × 300 meters resolution.

Materials and Methods

Data sources

The study area is located in the western part of Žďárské vrchy Protected Landscape Area. The data on the Wild Duck occurrences were obtained from the field mapping, which were conducted during the high breading season in May – June 1999 to 2003. The occurrence of bird species was mapped in 300 × 300 m network on 1141 sampling points. The habitat variables were obtained from the Base map of the Czech Republic at scale 1:10 000 (ZABAGED), obtained from Czech Office for Surveying, Mapping, and Cadastre. At each mapping point (square 300 × 300 m) we calculated the area of following habitats: forest (*for*), arable land (*ara*), grasslands (*gra*), watercourses and water bodies (*wat*), villages (*vill*), and wetlands (*wet*).

Introduction of positional error

To investigate the effect positional error, caused by false positive detection of species, on the performance of the model, we generated 5 datasets with introduced positional error. We assumed that false positive error occurs when a species is wrongly assigned to the mapping point instead of neighbouring point. For the first dataset D1, we randomly selected 20 percent of recorded presences of the Wild Duck and moved them randomly to one of the eight mapped neighbouring points (i.e. one pixel at the resolution analysed). To assess more extreme errors, we moved randomly 40 percent for D2, 60 percent for D3, 80 percent for D4, 100 percent for D5 of presences of the Wild Duck. If moving the occurrences resulted in transferring them outside the mapped points (i.e. study area) we recalculated the shift until the point remained on a mapped point. All geodata processing was performed using geographic information system ArcGIS 9.3 (ESRI, CA, USA).

Statistical analysis

To assess which habitat variables are important for breeding distribution of the Wild Duck, we used generalized additive models (GAM, Hastie and Tibshirani 1990) with an automated bidirectional stepwise selection method based on Akaike's Information Criterion (AIC, Akaike 1974). The modelling was performed with a binomial error distribution and a logit link function. All data (1141 sampling points) were used for model training and testing.

Model performance was evaluated using area under the receiver operating characteristics curve (AUC). AUC is a threshold independent measure of the ability of a model to discriminate between sites where species is present and those where it is absent (Fielding and Bell 1997) and ranges from 0.5 for models with no discrimination ability to 1 for models with perfect discrimination. Additionally, we calculated sensitivity (ability to predict presences) and specificity (ability to predict absences) as the most simple and straightforward measure of model performance. The threshold was chosen to maximize the sum of sensitivity and specificity (Jiménez-Valverde and Lobo 2006). The analysis was carried out using the 'gam 1.06.2' (Hastie 2011) and 'PresenceAbsence 1.1.5' packages (Freeman 2007) in free statistical software R version 2.12.2 (R Development Core Team 2010).

Results

Area of water, forest and villages were selected as important habitat variables in the model for the Wild Duck. The model explained 36.5% of the variation in distribution of the Wild Duck (total deviance change 162.9 out of 446.8). The model of the Wild Duck distribution received relatively high AUC value 0.92 indicating "excellent" accuracy of the model.

In all cases the datasets with introduced positional error have lower AUC scores than AUC score obtained with the original dataset (Tab. 1). However, the drop in AUC was relatively low for D1 and the model was still regarded as

"excellent" when judged by AUC. Even when each recorded presence had been shifted (D5), the performance of the model would be regarded as "good".

Table 1: AUC values, Sensitivity and Specificity for the model with original data and models with incorporated error in the species occurrences (D1— D5).

	Original data	**D1**	**D2**	**D3**	**D4**	**D5**
AUC	0.92±0.03	0.90±0.02	0.85±0.03	0.79±0.03	0.78±0.03	0.79±0.03
Sensitivity	0.90±0.07	0.80±0.05	0.73±0.06	0.72±0.06	0.72±0.06	0.67±0.07
Specificity	0.81±0.02	0.80±0.01	0.81±0.01	0.72±0.01	0.66±0.01	0.79±0.01

Discussion and Conclusions

The habitat variables important for occurrence of the Wild Duck are in concurrence with our expectations. Water is important nesting habitat of the Wild Duck and forests are the most important component of the landscape of the mapped area. Also, there are small ponds in villages, where the Wild Duck occur and which are not recorded in water habitat data for their relatively small area.

In our study, we investigated the influence of positional error, caused by false positive detection of species, by shift in species occurrences on one of the eight mapped neighbouring points. Our finding showed that positional error incorporated into the occurrence of the Wild Duck led to a reduction in model prediction accuracy. However, it has a small effect on the performance of models judged by AUC, which is in concurrence with Graham et al. (2008) and Osborne and Leitao (2009). According to Graham et al. (2008), useful predictions can be made even when species occurrence data include some positional error. However, it depends on the range of spatial autocorelation in the habitat variables, which reduces the impact of positional error on the predictions (Naimi et al. 2011). It can explain why the shift in 60% and more of species occurrences did not lead to further decrease in model performance. Moreover, the reliability of AUC as a comparative measure of accuracy between model results has been recently criticised (Lobo et al. 2008). In consequence, models developed with species occurrence data containing false positive errors would be probably interpreted as relevant, when judged by AUC. The incorporated false positive error in species occurrence data had the greatest influence on model sensitivity. In our opinion, sensitivity and specificity should be additionally reported to better indicate the erroneous models.

We are aware that our models are built and tested on the same data, which can lead to overly optimistic estimates of model accuracy and that evaluation of the models using independent data is preferred (Newbold et al. 2010). However, it was not our intention to assess relevant habitat variables and to provide reliable model of distribution of Wild Duck, but to evaluate the performance of a model built with data that contains false positive errors.

Acknowledgement

This study was funded by the Czech University of Life Sciences, Prague (Grant No. 20124215).

Reference

MOUDRÝ, V. and P. ŠÍMOVÁ, 2012: Influence of positional accuracy, sample size and scale on modelling species distributions: a review. *International Journal of Geographical Information Science*. **26**, 2083–2095. ISSN 1365-8816.

AKAIKE, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. **19**, 716–723. ISSN 0018-9286.

BARRY, S. and J. ELITH, 2006. Error and uncertainty in habitat models. *Journal of Applied Ecology*. **43**, 413–423. ISSN 1365-2664.

BIBBY, C. J. et al., 2000. *Bird Census Techniques*. Second edition. London: Academic Press. ISBN 978-0120958313.

DAVIS, S. K., 2004. Area sensitivity in grassland passerines: effects of patch size, patch shape, and vegetation structure on bird abundance and occurrence in southern Saskatchewan. *Auk*. **121**, 1130–1145. ISSN 0004-8038.

FIELDING, A.H. and J. F. BELL, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*. **24**, 38–49. ISSN 0376-8929.

FREEMAN, E., 2007. PresenceAbsence: An R Package for Presence-Absence Model Evaluation. CRAN. R package version 1.1.5.

GRAHAM, C. H. et al., 2008. The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*. **45**, 239–247. ISSN 1365-2664.

HASTIE, T., 2011. gam: Generalized Additive Models. CRAN. R package version 1.06.2.

HASTIE, T. and R. TIBSHIRANI, 1990. *Generalized additive models*. London: Chapman and Hall. ISBN 978-0412343902.

HEUVELINK, G., 1998. *Error propagation in environmental modelling with GIS*. London: Taylor and Francis. ISBN 978-0748407446.

JIMÉNEZ-VALVERDE, A. and J. M. LOBO, 2006. The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions*. **1**2, 521–524. ISSN 1472-4642.

LOBO, J. M. et al., 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*. **17**, 145–151. ISSN 1466-8238.

LÓPEZ, G. and M. J. MORO, 1997. Birds of Aleppo pine plantations in south-east Spain in relation to vegetation composition and structure. *Journal of Applied Ecology.* **34**, 1257–1272. ISSN 1365-2664.

MILLER, D. A. W. et al., 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications.* **22**, 1665–1674. ISSN 1051-0761.

NAIMI, B. et al., 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography.* **38**, 1497–1509. ISSN 0305-0270.

NEWBOLD, T. et al., 2010. Testing the accuracy of species distribution models using species records from a new field survey. *Oikos.* **119**, 1326–1334. ISSN 1600-0706.

OSBORNE, P. E. and P. J. LEITÃO, 2009. Effects of species and habitat positional errors on the performance and interpretation of species distribution models. *Diversity and Distributions.* **15**, 671–681. ISSN 1472-4642.

SHI, W., P. F. FISHER and M. F. GOODCHILD, 2002. *Spatial Data Quality.* London: Taylor and Francis. ISBN 978-0415258357.

R Development Core Team, 2010. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

RALPH, C. J., S. DROEGE and J. R. SAUER, 1995. Managing and monitoring birds using point counts: Standards and applications. In: Ralph, C. J., J. R. Sauer and S. Droege (ed.). *Monitoring Bird Populations by Point Counts.* U.S. Department of Agriculture, Forest Service General Technical Report PSW-GTR-149. 161–168. ISBN 978-0899046518.

Vliv nesprávného určení výskytu druhu na modelování distribuce druhů: Případová studie s kachnou divokou (*Anas Platyrhynchos*)

V poslední době je v popředí zájmu ekologického výzkumu, zejména pak v modelování distribuce druhů, neurčitost vyplývající z nedostatků datových vstupů. Toto téma zahrnuje mimo jiné polohovou přesnost určení výskytu druhu. Předkládaná studie zkoumá vliv polohové chyby způsobené nesprávným určením výskytu druhu na výsledky modelování distribuce druhů. Pro analýzu byla využita data o výskytu kachny divoké (*Anas platyrhynchos*) mapované v rozlišení 300 × 300 metrů a proměnné prostředí získané ze Základní mapy ČR v měřítku 1 : 10 000. Environmentální proměnné, nejlépe vysvětlující výskyt kachny divoké, byly zjištěny krokovou regresí zobecněného aditivního modelu. Přesnost modelu byla vyhodnocena dle hodnot AUC, sensitivity a specificity. Uměle zahrnutá polohová chyba vedla ve všech případech ke snížení přesnosti modelu, i když ne takové, aby mohl být model zamítnut.

Klíčová slova: prostorová neurčitost, polohová přesnost, SDM (modelování distribuce druhů)

Kontaktní adresa:
Ing. Vítězslav Moudrý, Katedra aplikované geoinformatiky a územního plánování, Fakulta životního prostředí, Česká zemědělská univerzita, Kamýcká 129 165 21 Praha 6, e-mail: *moudry@fzp.czu.cz*

MOUDRÝ, V. and P. ŠÍMOVÁ, The Effects of False Positive Errors in Species Occurrence Data on the Performance of Species Distribution Models: Case Study on the Wild Duck (*Anas Platyrhynchos*). *Littera Scripta.* 2012, **5**(2), 243–249. ISSN 1802-503X.
