Ecological Informatics 19 (2014) 10–15

Contents lists available at ScienceDirect

Ecological Informatics

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

Spatial bias in the GBIF database and its effect on modeling species' geographic distributions

Jan Beck [a,*], Marianne Böller [a,b], Andreas Erhardt [a], Wolfgang Schwanghart [b,1]

[a] University of Basel, Department of Environmental Science (Biogeography & Conservation Biology), St. Johanns-Vorstadt 10, CH-4056 Basel, Switzerland
[b] University of Basel, Department of Environmental Science (Physical Geography), Klingelbergstrasse 27, CH-4056 Basel, Switzerland

ARTICLE INFO

Article history:
Received 29 August 2013
Received in revised form 7 November 2013
Accepted 8 November 2013
Available online 12 November 2013

Keywords:
AUC
CSCF
Ecological niche modeling
Lepidoptera
Maxent
Small tortoiseshell

ABSTRACT

Species distribution modeling, in combination with databases of specimen distribution records, is advocated as a solution to the problem of distributional data limitation in biogeography and ecology. The global biodiversity information facility (GBIF), a portal that collates digitized collection and survey data, is the largest online provider of distribution records. However, all distributional databases are spatially biassed due to uneven effort of sampling, data storage and mobilization. Such bias is particularly pronounced in GBIF, where nation-wide differences in funding and data sharing lead to huge differences in contribution to GBIF.

We use a common Eurasian butterfly (*Aglais urticae*) as an exemplar taxon to provide evidence that range model quality is decreasing due to the spatial clustering of distributional records in GBIF. Furthermore, we show that such loss of model quality would go unnoticed with standard methods of model quality evaluation. Using evaluations of model predictions of the Swiss distribution of the species, we compare distribution models of full data with data where a subsampling procedure removes spatial bias at the cost of record numbers, but not of spatial extent of records. We show that data with less spatial bias produce better predictive models even though they are based on less input data. Our subsampling routine may therefore be a suitable method to reduce the impact of spatial bias to species distribution models.

Our results warn of automatized applications of species distribution models to distributional databases (as has been advocated and implemented), as internal model evaluation did not show the decline of model quality with increased spatial bias (but rather the opposite) while expert evaluation clearly did.

1. Introduction

Research in biogeography, ecology and biodiversity depends on data on species distributions and environmental conditions to uncover the mechanisms shaping the spatial distribution of life on Earth. However, while a surge of high-quality, satellite-derived remote sensing data on environmental conditions contributed to new insights over the last two decades, a shortage of data on species distributions is increasingly seen as a limiting factor in many fields of basic and applied ecology (Beck et al., 2012; Jetz et al., 2012). The mobilization of available, yet hard to access data (for example from natural history collections) in the form on online databases is seen as key advance to address this problem. The global biodiversity information facility (GBIF; www.gbif.org) is aiming at mobilizing biodiversity data from museums, surveys, and other data sources by collating locally digitized and stored data in an online data search portal.

GBIF is not the only initiative of its kind, but it is by far the largest and is therefore seen as a major step in closing the aforementioned data gap (Beck et al., 2012; Jetz et al., 2012). However, this pivotal position also implies continuing scrutiny of its methods and the data provided. Data quality issues and the lack of transparency of data quality have been noted by many and were publicly criticised (Graham et al., 2008; Soberón et al., 2002; Yesson et al., 2007). Introducing a peer-review system for data publications has been suggested (Chavan and Penev, 2011; Costello et al., 2013) and has begun to be applied as possible solution (Narwade et al., 2011).

Spatial bias in species distribution data is a general phenomenon with the potential of strongly distorting our view on large-scale biodiversity patterns (Ballesteros-Mejia et al., 2013; Boakes et al., 2010; Yang et al., 2013; and references therein). A multitude of factors, such as where surveys were carried out and at what spatial scale, what data or specimens were collected, and which of these data were stored and finally mobilized, can cause such biases. Data provided by GBIF are no exception to these problems. The national structure of funding museum data digitalization and policies of sharing data with GBIF may weigh particularly high as factors leading to spatial bias in the data made available.

Recently, Beck et al. (2013) have shown for the European member of a Lepidopteran family that GBIF data, despite being much more

[*] Corresponding author. Tel.: +41 61 2670810.
E-mail address: jan.beck@unibas.ch (J. Beck).
[1] Present address: University of Potsdam, Institute of Earth and Environmental Science, Karl-Liebknecht-Str. 24–25, D-14476 Potsdam-Golm, Germany.

1574-9541/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecoinf.2013.11.002

numerous, are less informative with respect to species ranges and climatic niches than targeted data compilations from museums, collections and literature. Presumably, clustering of GBIF data in certain countries causes a shift in perceived species occurrence and commonness (in geographic and environmental space). For example, many more records of a species are available from well-financed, data-sharing countries such as Sweden or the UK, even if the real density of occurrences of the species may be higher elsewhere (e.g., species-rich Balkan countries).

Ecological niche modeling or species distribution modeling (SDM; Elith and Leathwick, 2009) is a quantitative way of estimating species geographic ranges from occurrence records and the environmental conditions found there. Despite manifold criticism on some aspects of its implementation (among them, uninformed use and input data quality; Beale and Lennon, 2012; Joppa et al., 2013) it is an important tool to provide geographic range estimates for many poorly known species. SDM is increasingly applied to data provided by GBIF (Costello et al., 2013; Guralnick and Hill, 2009). However, it must be assumed that ecological niche models are very sensitive to the distortion of observed environmental conditions in specimen records caused by spatial bias (Dudík and Phillips, 2005; Lintz et al., 2013; Phillips et al., 2009).

In the present study, we use a common Eurasian butterfly, the Small Tortoiseshell (*Aglais urticae*) as an exemplar taxon to investigate how SDM quality is affected by the spatial clustering of records in GBIF data. By comparing SDM predictions based on subsampled data with known occurrence data for Switzerland, we (1) test whether spatial clustering leads to a decrease in model prediction accuracy, (2) investigate whether our subsampling procedure can be used to improve model predictions under such data conditions, and (3) check whether standard model quality metrics correctly indicate potential problems due to spatial clustering.

2. Methods

2.1. Study taxon and input data

A. urticae is a nymphalid butterfly of the northern-temperate zone of Eurasia. Larvae feed on widespread nettle (*Urtica*). The highly mobile adults are commonly seen flying or sun-basking. With its conspicuous colouration the species can hardly be misidentified within the European butterfly fauna.

As GBIF data for invertebrates are generally sparse outside industrialized countries, we restricted the extent of input data and of SDMs to Europe and the Mediterranean (see Fig. 1 for extent). We downloaded distributional records for the species from GBIF (www.gbif.org, accessed January 2010), which led to 32,355 records after excluding data with missing or clearly false locality coordinates. Notably, because public institutions in Switzerland had not contributed data to GBIF at the time of accessing GBIF, there are very few records of the species from this country. Switzerland is situated in the centre of the modeled region and features, due to its mountainous topography, a wide variety of climatic zones and landscapes.

2.2. Distribution models and internal evaluation

To create SDMs, we used the most widely utilized method and software, Maxent (v. 3.3.2; Phillips and Dudík, 2008; Phillips et al., 2006; see also Joppa et al., 2013). Maxent was found to be a very good SDM method in critical comparisons of presence-only data modeling approaches (Elith et al., 2006; but see Fitzpatrick et al., 2013). We used mostly standard software settings (i.e., using logistic output and no bias-file for background data sampling, see the Discussion section) but we sampled 100,000 background points to balance the large number of input records. As environmental predictor variables we utilized altitude and 12 climatic variables (from www.worldclim.org, accessed Feb. 2009), as well as vegetation cover data from MODIS remote sensing (http://glcf.umiacs.umd.edu/data/vcf; accessed Feb. 2009; see Appendix for details). Data were used in a resolution of 2.5 arc min (ca. 5 × 5 km).

Following standard practices of 'internal' SDM quality evaluation, input distribution records were separated randomly into a "training data" set (75% of data) used for model fitting and remaining "test data" used for calculating the integral of the receiver-operating characteristic ("area under the curve", AUC). AUC is a measure of predictive model quality independent of applying a threshold for converting continuous modeled "suitability" into predictions of presence or absence (Brown and Davis, 2006; Pearce and Ferrier, 2000). Because AUC calculation for presence-only data, as provided by Maxent, replaces missing commission error data with predicted area size, we denote this as AUC_{Maxent} to distinguish it from true AUC (Brown and Davis, 2006; see below). All models and AUC_{Maxent} values presented are averages from 5 replicate runs with different random separations of data into "test" and "training". AUC_{Maxent} are thus based on the entire modeling extent, i.e. Europe and the Mediterranean.

2.3. Manipulating data density

To test whether spatial bias in distribution data directly affects model quality, we devised a subsampling routine that removes records from high-density regions. After converting data to an equal-area projection (Lambert Conformal Conic) we divided the modeling region into 100 × 100 km cells. From GBIF data, a predefined maximum number of records per cell were randomly chosen (if available). Allowing only few, to the extreme only one, records per cell produced data sets with lower spatial clustering (and less records) than full data, whereas the spatial extent (hence, sampling of large-scale environmental variation) remained quite unchanged. Using more records per cell led to more clustered data sets. We generated subsampled data sets by iteratively and randomly extracting up to 20 records per cell, if available in a cell. We tested the spatial distribution of the subsampled data against a Poisson distribution. A MATLAB script of the subsampling procedure is provided in the Appendix. We ran SDMs for full GBIF data as well as for the 20 subsamples with manipulated data density (see Fig. 1). An alternative subsampling procedure, based on kernel densities of records, did not remove spatial clustering sufficiently to affect model qualities (details in the Appendix).

2.4. Independent tests of model accuracy

Averages of five replicate MaxEnt runs were viewed as one model prediction and assessed further. We used Swiss predictions of Europe-wide models for an independent evaluation of model quality. While GBIF data contained very few distribution records of the species for Switzerland, detailed records of the species are stored in the databank of the *Centre Suisse de Cartographie de la Faune* (CSCF — Centre Suisse de Cartographie de la Faune, 2011; accessed June 2011, 5 × 5 km grid cells). Because Switzerland is a densely populated country with a high level of faunistic data collection, and because *A. urticae* is so obvious due to its behaviour and colouration, we assumed that CSCF grid cells without records indicate true absences of the species. This allowed computing true omission and commission errors of SDMs for Switzerland, hence true AUC_{Swiss} values (including tests for significantly better than random predictions, i.e., $AUC_{Swiss} > 0.5$; software: SPSS 18.0).

As a second, independent metric of SDM prediction quality for Switzerland, A. Erhardt, an expert for the ecology, behaviour and distribution of the Swiss Lepidoptera visually interpreted and evaluated predictive maps (Maxent logistic output, identical colour scheme for all maps), applying the Swiss high school grading system (1–6, in steps of 0.5; best grade is 6). Notably, beforehand he was neither informed about the differences between models nor about the objective of the study to assure his independence of assessment.

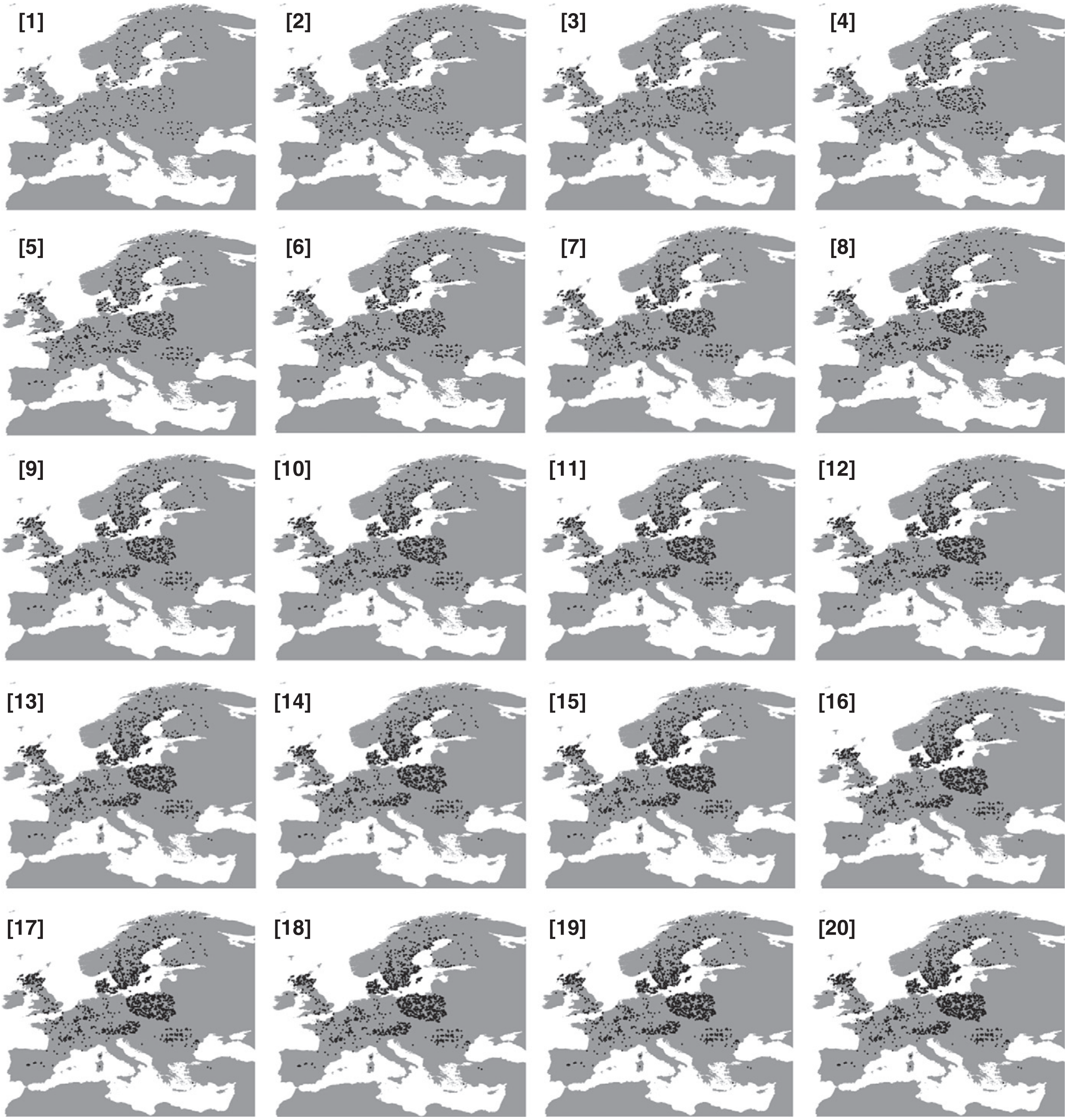

Fig. 1. Subsampled GBIF data with defined upper limits of records allowed per 100 × 100 km cell. Numbers in square brackets give maximum numbers of records per cell. Low numbers lead to reduced spatial clustering of data (see the Methods section for details, Appendix for map of full data set).

3. Results

Decreasing numbers of records per cell during subsampling diminished the spatial clustering of records (Figs. 1 & 2). Tests against a Poisson distribution revealed that subsamples with 7 or more records per cell were significantly clustered (Fig. 2).

Fig. 3 shows SDM predictions for Switzerland. Reducing spatial clustering by subsampling clearly weakens the east–west gradient in model predictions (also in Europe-wide data, Appendix), a gradient that is not observable in CSCF distribution data. Data with fewer points per 100 × 100 km cell are graded much better by the expert than those with many (Table 1; linear regression of records vs. expert grades: $N = 20$, $r = -0.926$, $p < 0.0001$). However, no relationship was found between record number per cell and AUC_{Swiss} ($r = 0.106$, $p = 0.657$).

Using less points (i.e., removing spatial bias) leads to a decrease of model quality as measured internally by AUC_{Maxent} (linear regression of records vs. AUC_{Maxent}: $r = 0.765$, $p < 0.0001$). Consequently, expert grades and AUC_{Maxent} come to opposing conclusions of model quality ($r = -0.916$, $p < 0.0001$), whereas the relationship between AUC_{Maxent} and AUC_{Swiss} is quite ambiguous ($r = 0.444$, $p = 0.050$). Notably, AUC_{Swiss} are generally indicating very poor models (although most are

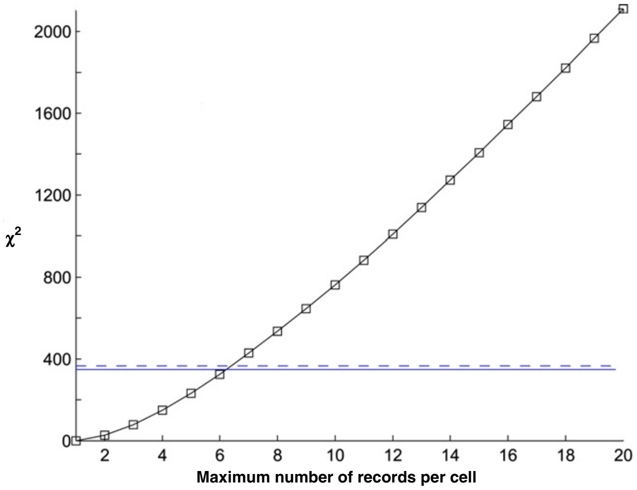

Fig. 2. Effect of subsampling (x-axis: number of records per 100 × 100 km cell) on spatial clumping (y-axis: χ^2-value based on a comparison of observed across-cells variance of records and expectation according to a Poisson distribution). Horizontal lines indicate critical thresholds for significant deviation from the expectation (solid: $p < 0.05$; broken: $p < 0.01$).

still significantly better than a random prediction) while $\text{AUC}_{\text{Maxent}}$ rate models much better (Table 1).

4. Discussion

Our results show that spatial bias in specimen distribution records can reduce the quality of predictive distribution models, as judged by an expert on the study taxon's distribution. In parallel to our study, Kramer-Schadt et al. (2013) have recently devised similar subsampling regimes (one record per 10 km radius, and a bias-file approach) which confirmed our assessment of model quality effects in an entirely different system (a tropical carnivore). Spatial bias (i.e., record density being affected by other factors than true density of the species in a region) is prevalent in probably all distributional data sets (e.g., Ballesteros-Mejia et al., 2013; Boakes et al., 2010; Yang et al., 2013) and such bias has been identified as a concern for SDM before (Dudík and Phillips, 2005; Phillips et al., 2009). However, such bias is probably particularly prominent in the GBIF database as, in addition to effects of data collection and storage, the national structure of funding and policy heavily affects data mobilization.

We carried out this study on an exemplar species in a particular region, but similar and even stronger biases exist for other taxa (e.g., Beck et al., 2013; Kramer-Schadt et al., 2013) or when comparing European to non-European distribution records. *A. urticae*, for example, occurs from Europe throughout northern Asia to the Pacific coast and Japan (e.g., Tutsov, 2000). Global extent-GBIF data (not shown), in contrast, indicates a huge, non-existent distribution gap across north-central Asia.

AUC as a measure of predictive quality of SDMs has been criticised previously for being affected by modeling extent, for being not consistent with other evaluation criteria, and (if applied to presence-only data) for not representing “true” AUC (Barve et al., 2011; Jiménez-Valverde, 2011; Lobo et al., 2008; Peterson et al., 2008; L. Ballesteros-Mejia, I.J. Kitching & J. Beck, unpubl.). Our data (Table 1) show that $\text{AUC}_{\text{Maxent}}$ and the butterfly expert came to opposing conclusions regarding the effect of clustered input data on model quality. We interpret this as an indication that $\text{AUC}_{\text{Maxent}}$ itself is misled by data bias, and hence unable to reveal real quality changes.

However, for $\text{AUC}_{\text{Swiss}}$, calculated from the records of the CSCF national databank, we did not find significant effects of spatial clustering on SDM quality. Rather, all SDMs were judged as very poor ($\text{AUC} < 0.57$). The inconsistency of conclusions based on expert grades and $\text{AUC}_{\text{Swiss}}$ is surprising as both, expert and CSCF data, suggested that the east–west gradient of habitat suitability, indicated by SDMs on highly clustered data, is an artefact (Fig. 3). Probably the expert assessment weighted such large-scale pattern (see also European predictions, Appendix) heavier than $\text{AUC}_{\text{Swiss}}$, which is based on a cell-by-cell error quantification. Hence, the two criteria may measure different types of error. A weak correspondence of AUC and expert grades when comparing SDMs for 64 moth species was also found by L. Ballesteros-Mejia, I.J. Kitching & J. Beck (unpubl.).

Dudík and Phillips (2005) and Phillips et al. (2009), with particular reference to the Maxent method (and software), have suggested using a bias file that targets the probability of background environmental samples to regions that have actually been well-sampled. Such a bias file can be created from a priori knowledge on sampling intensity, from densities of records for species that are likely to be sampled and mobilized together with the target species, or from features such as population density or traffic infrastructure that may allow predicting specimen sampling (Ballesteros-Mejia et al., 2013). However, for single-species studies such information is often not available. Our subsampling routine offers an alternative option of removing spatial bias in data availability, at least for cases where record numbers are very large and therefore a loss of sample size does not gravely affect model quality (cf. Kramer-Schadt et al., 2013).

Our procedures have some caveats that need to be kept in mind. First, we did not evaluate the quality of models within any of the densely sampled regions (e.g., southern Scandinavia). Subsampling may not improve model quality if modeling is restricted to such regions. Second, we did not replicate subsampling (and model evaluations based on different runs of subsamples). Thus, we cannot quantify the variation in model quality caused by the random choice of records within cells. However, the strong correlations of expert grades with subsample size (number of points per cell) indicate that random effects must be relatively weak.

5. Conclusions

Strong spatial bias in specimen distribution data, as prominently found in the GBIF database, has potential to gravely distort species distribution modeling. Internal model evaluation by AUC-values based on presence-only data, as typically applied to SDMs, may not only fail to indicate this loss in quality but can actually suggest opposing trends. These two aspects together are a major impediment to apply automated SDMs to GBIF and other large distributional databases, as suggested by Flemons et al. (2007) or Guralnick and Hill (2009). Subsampling distribution data, as applied in this study, can be used to remove spatial bias and hereby produce better SDMs, at least if original input data are so abundant that reduced sample size cannot be expected to reduce prediction quality.

Acknowledgements

We thank CSCF for allowing us to use their distributional databank for *A. urticae*. L. Ballesteros-Mejia helped with SDM implementations, S. Steiner contributed to our work with (unpubl.) investigations of the altitudinal distribution of *A. urticae*. A. Wilting and an anonymous reviewer made useful suggestions for presentation. The study is part of a project supported by the Swiss National Science Foundation (SNF) (grant 31003A_119879; SNF had no influence on study design).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ecoinf.2013.11.002.

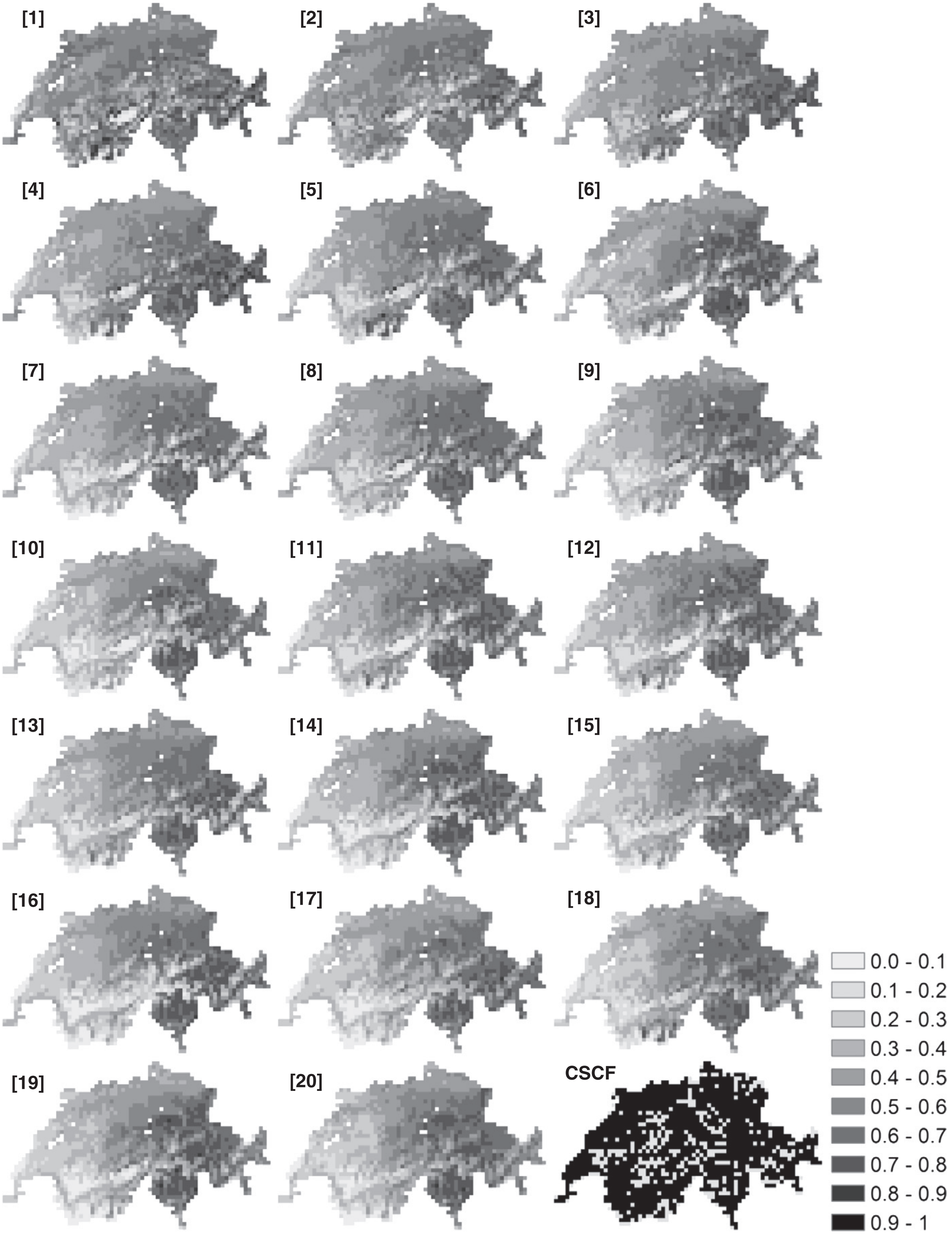

Fig. 3. Maxent model predictions (logistic output) for Switzerland based on subsampled data. Numbers in square brackets give maximum numbers of records per cell for input data. The last map (low-right corner, labelled CSCF) shows known presences (black) and assumed absences (light grey) from the CSCF database. See Appendix for Europe-wide models.

Table 1
Model evaluation of 20 subsampled datasets, with increasing sample size (maximum records per 100 × 100 km grid cell) and spatial clustering. Asterisks (*) indicate AUC_{Swiss} significantly better than for a random prediction ($p < 0.05$). See the Methods section for acronyms. Expert grades range from 1 (bad) to 6 (very good) in increments of 0.5.

Records	AUC_{Swiss}	AUC_{Maxent}	Expert grade
1	0.532	0.759	5.0
2	*0.545	0.786	4.0
3	0.526	0.796	3.5
4	0.529	0.806	3.0
5	*0.549	0.816	3.0
6	*0.557	0.816	3.0
7	*0.565	0.826	2.5
8	*0.537	0.814	2.5
9	*0.547	0.818	2.5
10	*0.535	0.813	2.0
11	*0.554	0.821	2.5
12	*0.554	0.824	2.0
13	*0.549	0.827	2.0
14	*0.547	0.825	2.0
15	*0.543	0.823	2.0
16	*0.551	0.826	1.5
17	0.526	0.824	1.5
18	*0.543	0.826	1.5
19	*0.539	0.823	1.5
20	*0.546	0.833	1.0
Mean	0.544	0.815	2.45

References

Ballesteros-Mejia, L., Kitching, I.J., Jetz, W., Nagel, P., Beck, J., 2013. Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. Glob. Ecol. Biogeogr. 22, 586–595.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J., Villalobos, F., 2011. The crucial role of the accessible area in ecological niche modelling and species distribution modelling. Ecol. Model. 222, 1810–1819.

Beale, C.M., Lennon, J.J., 2012. Incorporating uncertainty in predictive species distribution modeling. Phil. Trans. R. Soc. B 367, 247–258.

Beck, J., Ballesteros-Mejia, L., Buchmann, C.M., Dengler, J., Fritz, S., Gruber, B., Hof, C., Jansen, F., Knapp, S., Kreft, H., Schneider, A.-K., Winter, M., Dormann, C.F., 2012. What's on the horizon of macroecology? Ecography 35, 673–683.

Beck, J., Ballesteros-Mejia, L., Nagel, P., Kitching, I.J., 2013. Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Divers. Distrib. 19, 1043–1050.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biol. 8, e1000385.

Brown, C.D., Davis, H.T., 2006. Receiver operating characteristics curves and related decision measures: a tutorial. Chemom. Intell. Lab. Syst. 80, 24–38.

Chavan, V., Penev, L., 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinforma. 12, S2. http://dx.doi.org/10.1186/1471-2105-12-S15-S2.

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.-Q., Bourne, P.E., 2013. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol. Evol. 28, 454–461.

CSCF im kurzem. In: CSCF — Centre Suisse de Cartographie de la Faune (Ed.), (http://www.cscf.ch/cscf/page-20464_de_CH.html [accessed 29.06.2011]).

Dudík, M., Phillips, S.J., 2005. Correcting sample selection bias in maximum entropy density estimation. Advances in Neural Information Processing Systems. 18, pp. 323–330.

Elith, J., Leathwick, J., 2009. Species distribution models: ecological explanation and prediction across space and time. Ann. Rev. Ecol. Evol. Syst. 40, 677–697.

Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, Jm, Peterson, A., Phillips, S., Richardson, K., Scachetti-Pereira, R., Shapire, R., Soberón, J., Williams, S., Wisz, M., Zimmermann, N., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Fitzpatrick, M.C., Gotelli, N.J., Ellison, A.M., 2013. MaxEnt versus MaxLike: empirical comparisons with ant species distributions. Ecosphere 4, art55.

Flemons, P., Guralnick, R., Krieger, J., Ranipeta, A., Neufeld, D., 2007. A web-based GIS tool for exploring the world's biodiversity: the global biodiversity information facility mapping and analysis portal application (GBIF-MAPA). Ecol. Inform. 2, 49–60.

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend, P.A., Loiselle, B.A., 2008. The influence of spatial errors in species occurrence data used in distribution models. J. Appl. Ecol. 45, 239–247.

Guralnick, R., Hill, A., 2009. Biodiversity informatics: automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25, 421–428.

Jetz, W., McPherson, J.M., Guralnick, R.P., 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends Ecol. Evol. 23, 151–159.

Jiménez-Valverde, A., 2011. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Glob. Ecol. Biogeogr. 21, 498–507.

Joppa, L.N., McInerny, G., Harper, R., Salido, L., Takeda, K., O'Hara, K., Gavaghan, D., Emmott, S., 2013. Troubling trends in scientific software use. Science 340, 814–815.

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19, 1366–1379.

Lintz, H.E., Gray, A.N., McCune, B., 2013. Effect of inventory method on niche models: random versus systematic error. Ecol. Inform. 18, 20–34.

Lobo, J., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151.

Narwade, S., Kalra, M., Jagdish, R., Varier, D., Satpute, S., Khan, N., Talukdar, G., Mathur, V., Vasudevan, K., Pundir, D.S., Chavan, V., Sood, R., 2011. Literature based species occurrence data of birds of northeast India. ZooKeys 150, 407–417.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133, 225–245.

Peterson, A.T., Papeş, M., Soberón, J., 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol. Model. 213, 63–72.

Phillips, S.J., Dudík, M., 2008. Modelling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Phillips, S.J., Anderson, R., Schapire, R., 2006. Maximum entropy modelling of species geographic distributions. Ecol. Model. 190, 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197.

Soberón, J., Arriaga, L., Lara, L., 2002. Issues of quality control in large, mixed-origin entomological databases. In: Saarenmaa, H., Nielsen, E.S. (Eds.), Towards a global biological information infrastructure. European Environment Agency, Copenhagen, pp. 1–72.

Tutsov, V.K., 2000. Guide to the butterflies of Russia and adjacent territories. vol. 2. Pensoft, Sofia & Moscow (580 pp.).

Yang, W., Ma, K., Kreft, H., 2013. Geographical sampling bias in a large distributional database and its effects on species richness–environment models. J. Biogeogr. http://dx.doi.org/10.1111/jbi.12108 (early view).

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M., Gray, W.A., White, R.J., Jones, A.C., Bisby, F.A., Culham, A., 2007. How global is the global biodiversity information facility? PLoS One 2, e1124.