IS THE GBIF APPROPRIATE FOR USE AS INPUT IN MODELS OF PREDICTING SPECIES DISTRIBUTIONS? STUDY FROM THE CZECH REPUBLIC

Zuzana Štípková[1], Spyros Tsiftsis[2], Pavel Kindlmann[1,3]

[1]*Global Change Research Institute, Czech Republic*
e-mail: zaza.zuza@seznam.cz
[2]*International Hellenic University, Greece*
e-mail: stsiftsis@for.ihu.gr
[3]*Charles University, Czech Republic*
e-mail: pavel.kindlmann@centrum.cz

Received: 02.11.2023. Revised: 09.02.2024. Accepted: 15.02.2024.

Questions concerning species diversity have attracted ecologists and biogeographers for over a century, mainly because the diversity of life on Earth is in rapid decline, which is expected to continue in the future. One of the most important current database on species distribution data is the Global Biodiversity Information Facility (GBIF), which contains more than 2 billion occurrences for all organisms, and this number is continuously increasing with the addition of new data and by combining with other applications. Such data also exist in several national databases, most of which are unfortunately often not freely available and not included in GBIF. We suspected that the national databases, mostly professionally maintained by governmental organisations, may be more comprehensive than GBIF, which is not centrally organised and therefore the national databases may give more accurate predictions than GBIF. To test our assumptions, we have compared: (i) the amount of data included in the Czech database called Nálezová databáze ochrany přírody (NDOP, Discovery database of nature protection) with the amount of data in GBIF after its restriction to the Czech Republic, and (ii) the overlap of the predictions of species distributions for the Czech Republic, based on these two databases. We have used the family Orchidaceae as a model group. We found that: (i) there is a significantly larger number of records per studied region (Czech Republic) in NDOP, compared with GBIF, and (ii) the predictions of Maxent based on orchid records in NDOP are overlapping to a great degree with the predictions based on data based on orchid records in GBIF. Bearing in mind these results, we suggest that if only one database is available for the region studied, we must use this one. If more databases are available for the region studied, we should use the database containing most locations (usually some of the local ones, like NDOP), because using more locations implies larger significance of predictions of species distributions.

Key words: databases, Global Biodiversity Information Facility, NDOP, orchid distribution, species distribution models

Introduction

Questions concerning species diversity have attracted ecologists and biogeographers for over a century, mainly because the diversity of life on Earth is in rapid decline (Spooner et al., 2018; Baker et al., 2019; Halley & Pimm, 2023), which is expected to continue in the future (Román-Palacios & Wiens, 2020). For a reliable analysis of the rules governing the trends in species diversity, good data are necessary. To get them, direct sampling in the field, but also data available in museums and herbaria, which contain samples collected over centuries of field exploration (Smith & Blagoderov, 2012) are used. Mass digitalisation of all these data via interactive digital databases is now leading to their massive public availability (Maldonado et al., 2015) and to analyses using new computational methods and bioinformatic tools (Soberón & Peterson, 2004; Newbold, 2010).

Currently, one of the most important databases on species distribution data is the Global Biodiversity Information Facility (GBIF) (e.g. Beck et al., 2014; Maldonado et al., 2015; Chadin et al., 2017; Guedes et al., 2018; Alhajeri & Fourcade, 2019; Moudrý & Devillers, 2020; De Araujo et al., 2022), which contains currently more than 2 billion occurrences for all organisms, and this number is continuously increasing with the addition of new data and by combining with other applications (e.g. iNaturalist.org). Similar kind of data also exists in some national databases, such as the Czech database called Nálezová databáze ochrany přírody (NDOP, Discovery database of nature protection; see https://portal.nature.cz/nd/), most of which are unfortunately often not freely available and not included in the GBIF.

Thanks to the availability of powerful computers and advanced software, the occurrence and distribution of threatened species is now determined by species distribution models (SDMs) in combination with GIS techniques, which use the above-mentioned databases of species occurrence records and environmental data on climate, land use, geological substrate and other parameters as inputs (e.g. Guisan & Thuiller, 2005; Elith & Leathwick, 2009; Jiang & Purvis, 2023). Based on these, numerous papers have been published on current and future potential distributions of many species, and their range shifts under various climate change scenarios (e.g. Kistner & Hatfield, 2018; Weterings & Vetter, 2018; Tsiftsis & Djordjević, 2020; Namkhan et al., 2022; Arotolu et al., 2023). Many of them have used GBIF as input (e.g. Salvà-Catarineu et al., 2021; Daba et al., 2023; Krapf, 2023; Mallen-Cooper et al., 2023).

We have used the family Orchidaceae as a model group. Orchidaceae have a great species richness with about 20 000–35 000 species (Dressler, 1993; Chase et al., 2003; Cribb et al., 2003; Christenhusz & Byng, 2016). They are heavily threatened by extinction, and dispose of many varieties of reproductive strategies (Steffelová et al., 2023) and have an extremely restricted distribution with relatively small populations (Švecová et al., 2023). These traits make orchids an ideal model group because they are (i) important in conservation biology (Pillon & Chase, 2007; Swarts & Dixon, 2009) and (ii) crucial for their distribution and conservation status (Zhang et al., 2015).

We suspected that on the local scale the national databases, mostly professionally maintained by governmental organisations, may be more comprehensive than GBIF, managed by the GBIF Secretariat including four groups, so it is not centrally organised and therefore the national databases may give more accurate predictions than GBIF. To our knowledge, no study was yet published comparing the outcomes of any SDM method by using data from GBIF with those using any other national database. To test our expectations, we have compared (i) the amount of data included in the Czech database NDOP with the amount of data in GBIF, when it is restricted to the Czech Republic, and (ii) the overlap of the predictions of species distributions for the Czech Republic based on these two databases.

Material and Methods

The Czech Republic was chosen as a model country because its orchid flora is very well studied (Štípková et al., 2021). It is covered mainly by highlands of moderate altitude and higher mountains occur at its borders, especially in the north and south. The climate of the Czech Republic is typically temperate with cold, cloudy winters and hot summers. However, there are some regional and local differences due to the relief that forms a complex topography in this area (Palacký University Olomouc, 2020). Because the Czech Republic is a relatively small country in terms of latitudinal range, temperature and precipitation are mostly affected by local heterogeneity and altitude (Štípková et al., 2020b).

Two databases were compared: (i) one of the most important current database on species distribution data, the Global Biodiversity Information Facility (GBIF), which is freely accessible on https://www.gbif.org/ and (ii) the database NDOP (https://portal.nature.cz/nd/) of the Nature Conservation Agency of the Czech Republic, which is unavailable to the public to preserve orchid localities in the country. We used 55 orchid taxa. Their classification and nomenclature follow Danihelka et al. (2012). All studied species are threatened and protected on the national level and included on the national Red List (Grulich & Chobot, 2017).

NDOP was chosen because we have enough experience with it. Previously, Štípková & Kindlmann (2015), Štípková et al. (2018, 2020a) worked on the revision of orchid records in 24 mapping squares (see the network of mapping squares used for these purposes on https://www.entospol.cz/sit-mapovych-ctvercu/) in South Bohemia based on NDOP. More than 82% of records included in these squares were confirmed in NDOP, when revised. It was therefore supposed that records included in NDOP would be similarly correct for the whole Czech Republic with a small number of errors. Thus, we considered the NDOP to be sufficiently reliable for the purpose of this study. Nature Conservation Agency is divided into many regional branches across the whole area of the Czech Republic and each branch manages a certain area of the country. All data from the regional branches are then centralised in one database that guarantees uniformity of the database records. Moreover, NDOP allows their users to easily provide feedback on specific records, whereas GBIF does not.

We used Maxent (Phillips et al., 2006; Phillips & Dudík, 2008; Elith et al., 2011) to predict the current potential distribution of orchid species in the Czech Republic. The maximum entropy algorithm in the Maxent application (Phillips et al., 2006; Phillips & Dudík, 2008; Elith et al., 2011) is used for modelling species distribution from presence-only species records (Elith et al., 2011). This approach is widely used for predicting current as well as future distributions of species from a set of occurrence records and environmental variables (Yi et al., 2016; Tsiftsis & Djordjević, 2020). A great advantage of this method is that it has a high predictive performance even for very small sample sizes (Hernandez et al., 2006; Elith & Leathwick, 2009; David et al., 2020).

Bioclimatic variables and map of geological substrates of the Czech Republic were used as environmental predictors in the SDMs. Initially, 21 environmental variables were selected as predictors. Nineteen of them were bioclimatic variables and the remaining two were altitude and geological substrate. The bioclimatic variables were obtained from the WorldClim database (Fick & Hijmans, 2017) in a 30-sec resolution (approximately 1 km^2). The map of geological substrate was obtained from the geological map of the Czech Republic based on the digital geological map 1:500 000 (Czech Geological Survey, 1998). Because the map of the geological substrate is in vector format, the layer was converted into a raster format at the same resolution and extent with the layers of the bioclimatic variables.

To account for multicollinearity between the 19 bioclimatic variables and avoid overfitting, Pearson correlation coefficients were calculated for all pairwise interactions. To eliminate highly correlated variables, only one (i.e. the one with the higher percent contribution and training gain) was selected among any pair of those with a correlation coefficient r in the range $|r| > 0.70$. Specifically, in modelling the potential distribution of the studied species, the non-highly intercorrelated bioclimatic variables were used BIO 01 (annual mean temperature), BIO 02 (mean diurnal temperature range), BIO 05 (maximal temperature of warmest month), BIO 09 (mean temperature of driest quarter), BIO 12 (annual precipitation), and BIO 15 (precipitation seasonality). In addition, the altitude and the geological substrate were also used. The geological map of the Czech Republic contains the only categorial variable used in the models, but we treated all geological categories as dummy variables.

For both databases (NDOP and GBIF), we removed duplicate records (records falling in the same 1 km^2 grid cell), and we ran Maxent models only for species having at least 12 records in both databases. For each orchid species and database used, ten models were run. At each run, species records were randomly divided into training and testing datasets using the ratio between 80% and 20%, and we used 10 000 background samples to characterise the environmental conditions of the area of interest. Based on the output of the ten replicates, we calculated the average prediction.

SDMs outputs are numerical predictions, which provide a measure of the habitat suitability in an area (for example, at a country level). To convert these maps into presence/absence (binary) maps, the Maximum Sensitivity plus Specificity (MaxSSS) threshold was applied for each orchid species and database. This threshold was selected, as it provides better results than other thresholds, independently of the data used either presence/absence or presence-only data (Liu et al., 2016).

A niche equivalency test was used that shows Schoener's D and Hellinger Distances I of niche overlap (Warren et al., 2008). These statistics use suitability scores and have been widely used previously (e.g. Nunes & Pearson, 2017; Martínez-Méndez et al., 2019). Both these variables (D and I) measure niche overlap using different calculations, and their values range from 0 (no overlap between the two distributions) to 1 (identical distributions). Only D statistic was used for comparisons of percentage niche overlap of orchid occurrence data using Maxent model, as it is widely used in pairwise comparisons (e.g. El-Gabbas & Dormann, 2018; Chevalier et al., 2022).

To examine, whether there are significant differences in the mean altitude of the distribution of each of the studied species, we extracted the altitude values of the grid cells where each orchid is potentially present after converting the habitat suitability values into presence/

absence data. Thus, we compared the altitudinal values of the species distributions between the predictions of the two different datasets used in Maxent by using the Mann-Whitney U test in R v. 4.1.2 (R Core Team, 2023).

Results

In total, 31 orchid taxa had more than 12 records in both databases after removing the duplicates (Table 1). The number of orchid records included in GBIF and NDOP differed to a great degree, when compared in the region of the whole Czech Republic (Fig. 1). Initially, GBIF database contained 4328 of orchid records, NDOP contained 105 810 orchid records. The number of grid cell records analysed here, i.e. those containing enough records, after the reduction for duplicates etc., ranged from 61 (*Neotinea tridentata* (Scop.) R. M. Bateman, Pridgeon & M. W. Chase) to 13 636 records (*Dactylorhiza majalis* (Rchb.) P. F. Hunt & Summerh.) in the NDOP database, and from 13 (*Gymnadenia densiflora* (Wahlenb.) A. Dietr.) to 384 (*Neottia ovata* (L.) R. Br.) records in the GBIF database (Table 1).

Table 1. Species records used in Maxent and *D* statistics showing the niche overlap between the predictions of the two databases considered of 31 orchid taxa of the Czech Republic using Maxent

Species	Number of species records		Maxent
	NDOP	GBIF	*D* statistics
Anacamptis morio (L.) R.M.Bateman, Pridgeon & M.W.Chase	927	115	0.790
Anacamptis pyramidalis (L.) Rich.	238	63	0.625
Cephalanthera damasonium (Mill.) Druce	3631	322	0.860
Cephalanthera longifolia (L.) Fritsch	1493	244	0.813
Cephalanthera rubra (L.) Rich.	542	48	0.698
Cypripedium calceolus L.	576	95	0.692
Dactylorhiza fuchsii (Druce) Soó	4912	143	0.754
Dactylorhiza incarnata (L.) Soó	397	51	0.667
Dactylorhiza maculata (L.) Soó	346	32	0.711
Dactylorhiza majalis (Rchb.) P.F.Hunt & Summerh.	13 636	233	0.867
Dactylorhiza sambucina (L.) Soó	1150	92	0.751
Epipactis atrorubens (Hoffm.) Besser	643	85	0.700
Epipactis helleborine (L.) Crantz	7109	259	0.866
Epipactis palustris (L.) Crantz	1363	91	0.775
Gymnadenia conopsea (L.) R.Br.	2254	76	0.765
Gymnadenia densiflora (Wahlenb.) A.Dietr.	306	13	0.549
Neotinea tridentata (Scop.) R.M.Bateman, Pridgeon & M.W.Chase	61	19	0.455
Neotinea ustulata (L.) R.M.Bateman, Pridgeon & M.W.Chase	1082	70	0.813
Neottia cordata (L.) Rich.	369	22	0.749
Neottia nidus-avis (L.) Rich.	4867	272	0.848
Neottia ovata (L.) Hartm.	5121	384	0.867
Ophrys apifera Huds.	99	31	0.533
Ophrys insectifera L.	121	30	0.501
Orchis mascula (L.) L.	3845	83	0.737
Orchis militaris L.	709	135	0.796
Orchis pallens L.	598	163	0.779
Orchis purpurea Huds.	478	349	0.765
Platanthera bifolia (L.) Rich.	6104	255	0.837
Platanthera chlorantha (Custer) Rchb.	2113	37	0.815
Spiranthes spiralis (L.) Chevall.	232	16	0.729
Traunsteinera globosa (L.) Rchb.	619	42	0.609

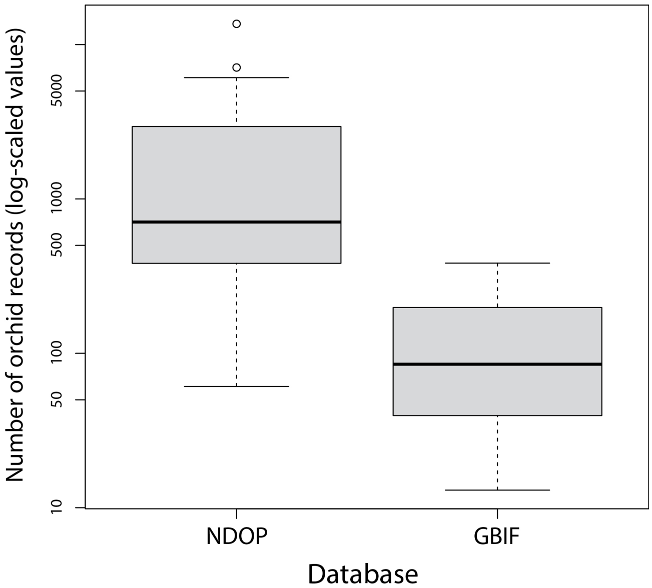

Fig. 1. Boxplot showing the number of orchid records (after removing duplicate records) in both databases (NDOP and GBIF) in the Czech Republic.

The values of the *D* statistics indicating the degree of niche overlap are presented in Table 1. The lowest niche overlap was observed in *Neotinea tridentata* (*D* value is 0.455), whereas the highest niche overlap was found in *Dactylorhiza majalis* and *Neottia ovata* (*D* value of both is 0.867). Most species showed a percentage overlap between 70% and 80%, but no species reached a percentage overlap between 90% and 100% (Fig. 2). Habitat suitability maps for each species based on data from the GBIF database and NDOP database are presented in Electronic Supplement 1. They show that GBIF often (but not always!) makes similar predictions to those made by NDOP.

The Mann-Whitney U test revealed significant altitudinal differences between data predictions from the NDOP and GBIF databases after Maxent had been applied (Table 2). Almost all data predictions of NDOP were significantly different from those of the GBIF database ($p < 0.001$). Only for *Spiranthes spiralis* (L.) Chevall. the *p*-value was lower ($p < 0.05$). The differences were not statistically significant only for two species, namely *Gymnadenia conopsea* (L.) R. Brown. and *Traunsteinera globosa* (L.) Rchb. The predictions of the Maxent model revealed statistically higher altitudinal distribution (in terms of the higher mean altitude) for 20 out of 31 studied species.

Table 2. Comparison of data presented in the NDOP and GBIF databases after Maxent predictions using Mann-Whitney U test in the Czech Republic

Species	Number of presence grid cells after Maxent predictions		Altitudinal statistics of the presence grid cells obtained through Maxent model using NDOP data				Altitudinal statistics of the presence grid cells obtained through Maxent model using GBIF data				Mann-Whitney U test between data of NDOP and GBIF
	NDOP	GBIF	Min	Max	Mean	SD	Min	Max	Mean	SD	
Anacamptis morio	15 867	30 444	183	866	447.01	124.68	187	699	397.17	88.95	**
Anacamptis pyramidalis	10 383	7460	162	841	423.56	128.64	162	656	376.06	86.39	**
Cephalanthera damasonium	37 059	46 600	86	671	356.33	94.41	131	598	323.52	84.06	**
Cephalanthera longifolia	23 313	26 284	200	841	420.34	98.67	187	1359	403.30	105.56	**
Cephalanthera rubra	21 286	20 198	97	825	415.48	87.41	245	745	423.81	74.78	**
Cypripedium calceolus	28 840	50 398	180	690	370.70	91.64	51	577	298.18	77.66	**
Dactylorhiza fuchsii	20 551	15 897	289	1524	721.20	186.47	157	1524	761.42	193.02	**
Dactylorhiza incarnata	32 281	27 520	51	1007	302.94	123.37	51	1524	277.10	115.65	**
Dactylorhiza maculata	23 650	10 966	223	1524	654.78	218.79	185	1524	806.85	212.42	**
Dactylorhiza majalis	47 016	36 189	382	1402	638.53	142.23	157	1461	637.57	190.93	**
Dactylorhiza sambucina	6179	9341	271	982	528.45	147.00	296	1407	646.23	195.12	**
Epipactis atrorubens	25 068	14 465	177	1248	492.41	188.78	235	1524	718.49	232.75	**
Epipactis helleborine	43 931	24 665	148	1461	517.68	205.98	235	1524	651.44	219.08	**
Epipactis palustris	24 220	25 905	159	1080	477.55	163.07	125	928	357.45	130.59	**
Gymnadenia conopsea	16 199	8385	183	1524	604.26	221.12	125	1524	622.69	244.74	0.168
Gymnadenia densiflora	14 816	43 087	125	1461	393.75	146.58	51	516	281.40	77.67	**
Neotinea tridentata	19 116	35 078	168	827	343.50	124.34	51	516	258.34	72.80	**
Neotinea ustulata	20 544	22 946	51	729	410.17	105.60	51	656	364.54	104.50	**
Neottia cordata	13 008	5154	288	1524	819.48	152.36	742	1524	953.35	131.51	**
Neottia nidus-avis	30 649	31 391	162	866	395.92	104.79	189	785	364.61	86.69	**
Neottia ovata	35 841	31 233	125	1325	446.99	184.64	125	1325	386.59	182.54	**
Ophrys apifera	13 701	21 397	162	671	343.17	105.90	134	545	274.90	86.71	**
Ophrys insectifera	18 257	5340	51	906	361.74	175.65	51	863	299.98	96.65	**
Orchis mascula	8705	8791	249	969	528.73	145.60	237	857	492.85	127.91	**
Orchis militaris	11 695	12 880	152	779	327.19	119.01	51	1524	304.25	124.16	**
Orchis pallens	9638	12 834	175	733	412.72	108.13	192	671	380.36	106.62	**
Orchis purpurea	28 813	22 468	86	623	333.34	84.03	86	559	285.07	68.93	**
Platanthera bifolia	44 770	29 432	189	1209	494.24	174.18	230	1461	553.56	217.43	**
Platanthera chlorantha	35 793	13 703	162	1282	581.17	210.81	51	1524	740.47	270.59	**
Spiranthes spiralis	32 768	92 074	122	1209	425.56	109.94	143	1133	420.58	143.28	*
Traunsteinera globosa	6289	3738	171	1461	537.32	185.68	171	952	521.89	142.36	0.182

Note: ** – $p < 0.001$, * – $p < 0.05$.

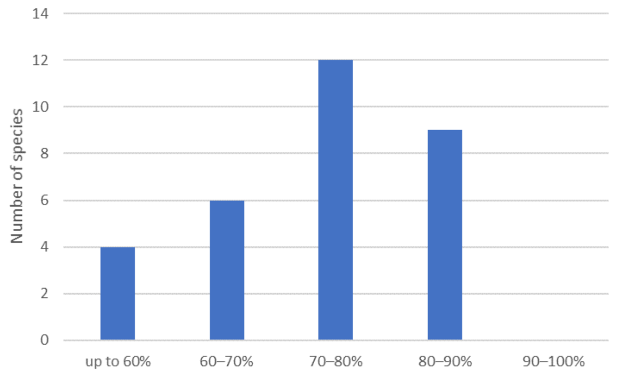

Fig. 2. Percentage overlap between data from NDOP and GBIF database using *D* statistic from Maxent application in the Czech Republic.

Fig. 3 shows the importance of the environmental variables when orchid records from NDOP and GBIF are used in Maxent. The evaluation of the importance of each environmental variable was based on the jackknife test using each predictor separately. The lengths of the bars correspond to the percentage contribution of each environmental predictor to the total training gain of each model. For example, in the line associated with *Anacamptis morio* (L.) R. M. Bateman, Pridgeon & M. W. Chase, when NDOP data are used, the longest bar (the dark green one) is the mean diurnal temperature range (BIO 02). This means that the most important environmental variable for *Anacamptis morio*, when NDOP data are used, is the mean diurnal temperature range (BIO 02). Another important output of Fig. 3 is that the importance of variables may vary to a great extent between various databases used in the Maxent model. Specifically, for *Gymnadenia densiflora*, the geological substrate was the most important variable when data from NDOP were used, whereas altitude was among the less important ones. On the contrary, when the GBIF data were used, the importance of altitude was high, whereas that of the geological substrate was not. Something similar was also observed in the case of *Spiranthes spiralis*: when NDOP data were used, variables had a rather equal importance in the model, whereas when GBIF data were used, precipitation seasonality (BIO 15) was by far the most important variable compared to the others.

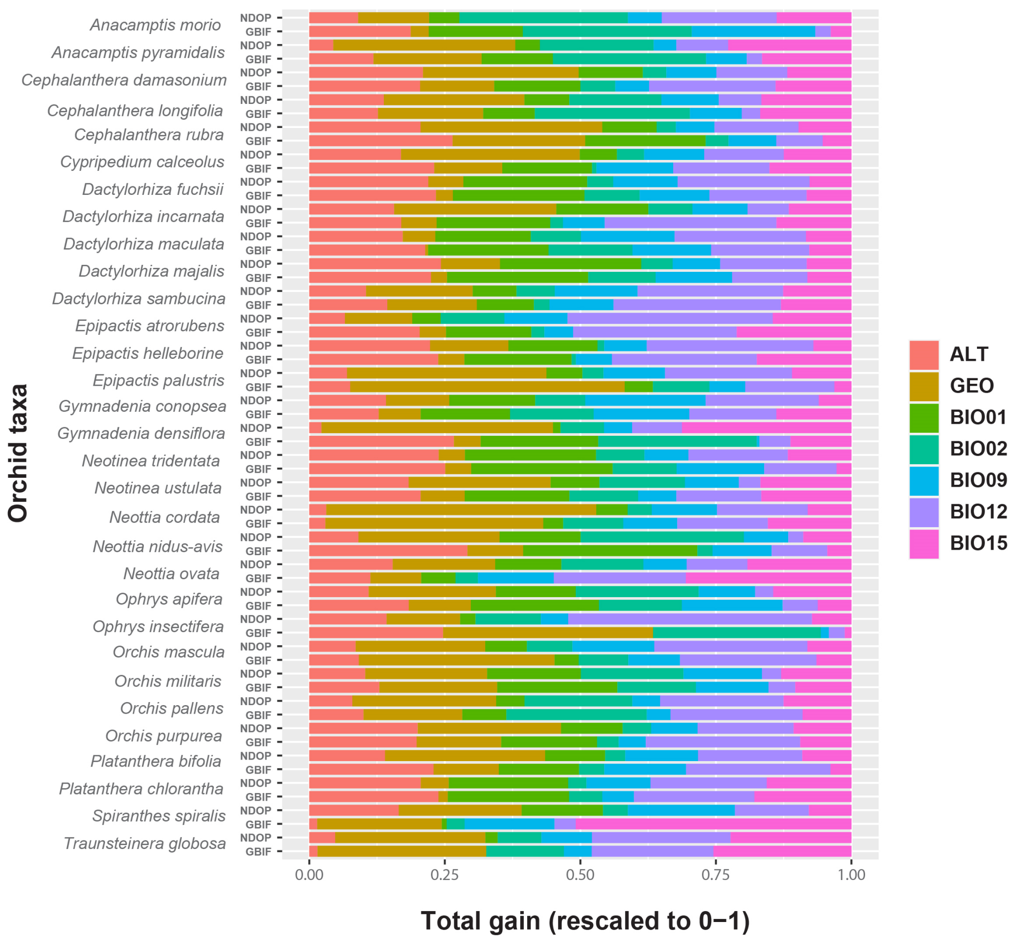

Fig. 3. The importance of the variables when orchid records from NDOP and GBIF are used in Maxent in the Czech Republic. The evaluation of the importance of each environmental variable was based on the jackknife test using each predictor separately. The lengths of the bars correspond to the percentage contribution of each environmental predictor to the total training gain of each model. Designation of the variables: ALT (altitude), GEO (geology), BIO 01 (annual mean temperature), BIO 02 (mean diurnal temperature range), BIO 09 (mean temperature of driest quarter), BIO 12 (annual precipitation) and BIO 15 (precipitation seasonality).

Differences in importance of the corresponding variables for the 31 orchid taxa when NDOP vs. GBIF data were used are documented in scatterplots in Fig. 4. The importance of altitude (ALT) and annual mean temperature (BIO 01) was higher (points above the diagonal in Fig. 4) when GBIF data were used, compared to the results of the NDOP data. On the contrary, when the NDOP data were used, the importance of the geological substrate for most orchid taxa was much stronger than when GBIF data were used (points below the diagonal in Fig. 4).

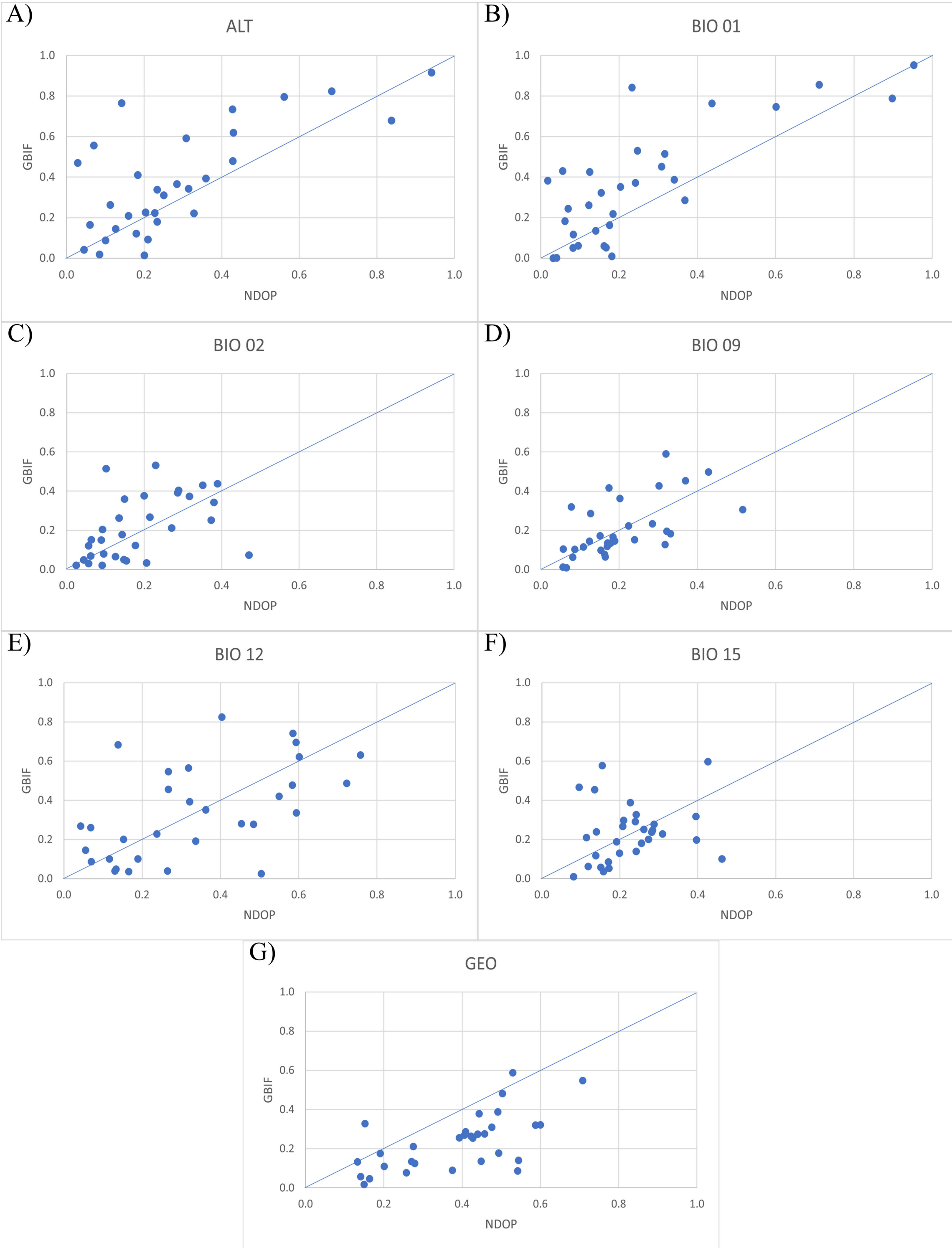

Fig. 4. Scatterplots showing the importance of each environmental variable based on the jack-knife test using each predictor separately in the case of the NDOP and GBIF database in the Czech Republic. Points above the main diagonal indicate a higher importance of the corresponding variable, when GBIF data are used and vice versa. Designation of the variables: A) ALT (altitude), B) BIO 01 (annual mean temperature), C) BIO 02 (mean diurnal temperature range), D) BIO 09 (mean temperature of driest quarter), E) BIO 12 (annual precipitation), F) BIO 15 (precipitation seasonality) and G) GEO (geology). Each dot represents an individual orchid species.

Discussion

The central topic of this paper is the comparison of accuracy of predictions based on public databases like GBIF against the governmentally controlled ones, like NDOP. We must admit that there are some practical advantages, when public databases, such as GBIF, are used for saving time and money and the uniformity of presented data that are ready to use for many analyses (Maldonado et al., 2015). However, how do the resulting predictions differ? What are the problems, when predictions based on records in the public databases like GBIF are compared with predictions of governmentally maintained ones, like NDOP?

First, our results show that there is a much larger number of orchid records in the governmentally maintained databases like NDOP than in public ones like GBIF, when like with like (i.e. records for the same region in both databases) is compared. In our study, the number of orchid records included in NDOP in the region specified at the beginning of the analysis (Czech Republic in this case) was much higher than that in GBIF (see Table 1). The reason for this is the long-term and systematic collection of data for NDOP from various parts of the Czech Republic. This renders a great advantage to the NDOP database for accuracy of predictions of species distribution in the region selected. The prevalence of records in the governmentally maintained databases, as opposed to the public ones, when like with like (the same region for both databases) is considered is not a solitary phenomenon of the Czech Republic. For example, the same occurs, when Greece is considered: GBIF for Greece has about 25 000 records (https://www.gbif.org/analytics/global), whereas the national database owned by Dr. Spyros Tsiftsis has more than 170 000 records (personal communication). So, the prevalence of records in the governmentally maintained databases, as opposed to the public ones, when like with like (the same region for both databases) is considered, seems to be a general phenomenon, if the governmentally maintained databases are good.

Second, in public databases like GBIF, the records are usually not as strictly controlled for correctness as governmentally maintained databases like NDOP. Questionable quality of unverified datasets, mistakes in the taxonomic identification of specimens or inaccurate georeferencing are common traits of public databases (Maldonado et al., 2015). Scientists and experts agree that a correct species name should be a minimum requirement for including the data in public databases, as well as an accurate georeferencing (Marcer et al., 2022), but this is not always the case. Mistakes in taxonomic identification can often be corrected by a taxonomist who has the possibility to access the specimen personally or at least see its image (Maldonado et al., 2015), and this is much more common in governmentally maintained databases like NDOP than in GBIF. A similar situation is with the errors in georeferencing (Graham et al., 2004).

Third, there is a common problem with records in public databases, like GBIF. Here, there are data spatially biased in most cases, which can greatly affect results of macroecological/biogeographical studies (Beck et al., 2014; Bowler et al., 2022; Boyd et al., 2022).

All these problematic inaccuracies can (and often will) affect results of studies dealing with biodiversity patterns, environmental niches and/or distribution predictions. Thus, information from public databases, like GBIF, must be used with caution due to important issues with data quality mentioned in the previous three paragraphs (Bowler et al., 2022; Boyd et al., 2022; Marcer et al., 2022). Just one example: it is well known that orchid distribution is strongly affected by the geological substrate (Djordjević & Tsiftsis, 2022). This is obvious when NDOP records, but not when the GBIF records are used (see Fig. 4G).

Surprisingly, despite of what was said in the four previous paragraphs, when two predictions were made: one based on records contained in NDOP and another one based on records contained in GBIF, then these two predictions were overlapping to a great degree in most cases (Table 1; Fig. 2), and there were often only rather small differences between them (Table 2; Fig. 4). Also, our results in Electronic Supplement 1 show that GBIF often (but not always!) makes similar predictions as NDOP. This suggests that GBIF may be used (with caution!) when no good local database is available.

No matter of what was said above here in the Discussion, there is one criterion that should be used, if the mentioned above does not suggest any preference for the use of public or governmentally based database: it is well known in statistics that the significance of the tests is positively correlated with the amount of data used in the test (Sokal & Rohlf, 2012). Therefore, the database containing more locations in the region considered should be preferred, because more locations imply a larger significance of predictions of species distribution.

Conclusions

Our analyses have shown that the predictions of species distributions based on data of orchid records from NDOP and GBIF databases are overlapping to a great degree. NDOP allows their users to easily provide feedback on specific records, whereas GBIF does not. Problematic inaccuracies might affect results of studies dealing with biodiversity patterns, environmental niches and/or distribution predictions, when based on public databases like GBIF, which therefore must be considered with caution. However, public databases have advantages in saving time and money in data collection and in uniformity of these data. With respect to significance of tests used, we suggest always using the database containing more locations (NDOP in our case), because more locations imply larger significance of predictions of species distributions.

Acknowledgements

We thank the Nature Conservation Agency of the Czech Republic for allowing us to use their dataset. We are greatly indebted to Tony Dixon for uncountable helpful hints how to improve the style of English in this paper. This study was supported by the Ministry of Education, Youth and Sports of CR within the CzeCOS program, grant number LM2023048 to Pavel Kindlmann, and by the PPLZ Program, grant number L200872201 to Zuzana Štípková.

Supporting Information

Habitat suitability maps of orchid species in the Czech Republic may be found in the **Supporting Information**.

References

Alhajeri B.H., Fourcade Y. 2019. High correlation between species-level environmental data estimates extracted from IUCN expert range maps and from GBIF occurrence data. *Journal of Biogeography* 46(7): 1329–1341. DOI: 10.1111/jbi.13619

Arotolu T.E., Wang H.N., Lv J.N., Shi K., Huang L.Y., Wang X.L. 2023. Modeling the current and future distribution of Brucellosis under climate change scenarios in Qinghai Lake basin, China. *Acta Veterinaria-Beograd* 73(3): 325–345. DOI: 10.2478/acve-2023-0025

Baker D.J., Clarke R.H., McGeoch M.A. 2019. The power to detect regional declines in common bird populations using continental monitoring data. *Ecological Applications* 29(5): e01918. DOI: 10.1002/eap.1918

Beck J., Böller M., Erhardt A., Schwanghart W. 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics* 19: 10–15. DOI: 10.1016/j.ecoinf.2013.11.002

Bowler D.E., Callaghan C.T., Bhandari N., Henle K., Barth B.M., Koppitz C., Klenke R., Winter M., Jansen F., Bruelheide H., Bonn A. 2022. Temporal Trends in the Spatial Bias of Species Occurrence Records. *Ecography* 2022(8): e06219. DOI: 10.1111/ecog.06219

Boyd R.J., Aizen M.A., Barahona-Segovia R.M., Flores-Prado L., Fontúrbel F.E., Francoy T.M., Lopez-Aliste M., Martinez L., Morales C.L., Ollerton J., Pescott O.L., Powney G.D., Saraiva A.M., Schmucki R., Zattara E.E., Carvell C. 2022. Inferring trends in pollinator distributions across the Neotropics from publicly available data remains challenging despite mobilization efforts. *Diversity and Distributions* 28(7): 1404–1415. DOI: 10.1111/ddi.13551

Chadin I., Dalke I., Zakhozhiy I., Malyshev R., Madi E., Kuzivanova O., Kirillov D., Elsakov V. 2017. Distribution of the invasive plant species *Heracleum sosnowskyi* Manden. in the Komi Republic (Russia). *Phytokeys* 77: 71–80. DOI: 10.3897/phytokeys.77.11186

Chase M.W., Cameron K.M., Barrett R.L., Freudebstein J.V. 2003. DNA data and Orchidaceae systematics: A new phylogenetic classification. In: K.W. Dixon, S.P. Kell, R.L. Barrett, P.J. Cribb (Eds.): *Orchid Conservation.* Kota Kinabalu: Natural History Publications (Borneo). P. 69–89.

Chevalier M., Zarzo-Arias A., Guélat J., Mateo R.G., Guisan A. 2022. Accounting for niche truncation to improve spatial and temporal predictions of species distributions. *Frontiers in Ecology and Evolution* 10: 944116. DOI: 10.3389/fevo.2022.944116

Christenhusz M.J.M., Byng J.W. 2016. The number of known plants species in the world and its annual increase. *Phytotaxa* 261(3): 201–217. DOI: 10.11646/phytotaxa.261.3.1

Cribb P.J., Kell S.P., Dixon K.W., Barrett R.L. 2003. Orchid conservation: A global perspective. In: K.W. Dixon, S.P. Kell, R.L. Barrett, P.J. Cribb (Eds.): *Orchid Conservation.* Kota Kinabalu: Natural History Publications (Borneo). P. 1–2.

Czech Geological Survey. 1998. *Geological map of the Czech Republic 1:500 000 (GEOCR500)*. Available from https://micka.geology.cz/en/record/basic/5f5b4530-a87c-4bf3-b45a-57d30a010852

Daba D., Kagnew B., Tefera B., Nemomissa S. 2023. Modelling the current and future distribution potential areas of *Peperomia abyssinica* Miq., and *Helichrysum citrispinum* Steud. ex A. Rich. in Ethiopia. *BMC Ecology and Evolution* 23(1): 71. DOI: 10.1186/s12862-023-02177-z

Danihelka J., Chrtek J.J., Kaplan Z. 2012. Checklist of vascular plants of the Czech Republic. *Preslia* 84: 647–811.

David O.A., Akomolafe G.F., Onwusiri K.C., Fabolude G.O. 2020. Predicting the distribution of the invasive species *Hyptis suaveolens* in Nigeria. *European Journal of Environmental Sciences* 10(2): 98–106. DOI: 10.14712/23361964.2020.11

De Araujo M.L., Quaresma A.C., Ramos F.N. 2022. GBIF information is not enough: national database improves the inventory completeness of Amazonian epiphytes. *Biodiversity and Conservation* 31(11): 2797–2815. DOI: 10.1007/s10531-022-02458-x

Djordjević V., Tsiftsis S. 2022. The role of ecological factors in distribution and abundance of terrestrial orchids. In: J.M. Mérillon, H. Kodja (Eds.): *Orchids Phytochemistry, Biology and Horticulture*. Cham: Springer Nature. P. 1–71. DOI: 10.1007/978-3-030-11257-8_4-1

Dressler R.L. 1993. *Phylogeny and Classification of the Orchid Family*. Cambridge: Cambridge University Press. 301 p.

El-Gabbas A., Dormann C.F. 2018. Wrong, but useful: regional species distribution models may not be improved by range-wide data under biased sampling. *Ecology and Evolution* 8(4): 2196–2206. DOI: 10.1002/ece3.3834

Elith J., Leathwick J. 2009. The contribution of species distribution modelling to conservation prioritization. In: A. Moilanen, A.K. Wilson, H.P. Possingham (Eds.): *Spatial conservation prioritization. Quantitative methods and computational tools*. New York: Oxford University Press Inc. P. 70–93.

Elith J., Phillips S.J., Hastie T., Dudík M., Chee Y.E., Yates C.J. 2011. A statistical explanation of MaxEnt for ecologist. *Diversity and Distributions* 17(1): 43–57. DOI: 10.1111/j.1472-4642.2010.00725.x

Fick S.E., Hijmans R.J. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12): 4302–4315. DOI: 10.1002/joc.5086

Graham C.H., Ferrier S., Huettman F., Moritz C., Peterson A.T. 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* 19(9): 497–503. DOI: 10.1016/j.tree.2004.07.006

Grulich V., Chobot K. 2017. Red list of threatened species of the Czech Republic vascular plants. *Příroda* 35: 1–178.

Guedes T.B., Sawaya R.J., Zizka A., Laffan S., Faurby S., Pyron R.A., Bérnils R.S., Jansen M., Passos P., Prudente A.L.C., Cisneros-Heredia D.F., Braz H.B., Nogueira C.D., Antonelli A. 2018. Patterns, biases and prospects in the distribution and diversity of Neotropical snakes. *Global Ecology and Biogeography* 27(1): 14–21. DOI: 10.1111/geb.12679

Guisan A., Thuiller W. 2005. Predicting species distribution: Offering more than simple habitat models. *Ecology Letters* 8(9): 993–1009. DOI: 10.1111/j.1461-0248.2005.00792.x

Halley J.M., Pimm S.L. 2023. The rate of species extinction in declining or fragmented ecological communities. *PloS ONE* 18(7): e0285945. DOI: 10.1371/journal.pone.0285945

Hernandez P.A., Graham C.H., Master L.L., Albert D.L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography* 29(5): 773–785. DOI: 10.1111/j.0906-7590.2006.04700.x

Jiang Y., Purvis A. 2023. How land use affects biodiversity: an analysis of the differences in the effects recorded on different continents. *European Journal of Environmental Sciences* 13(1): 15–22. DOI: 10.14712/23361964.2023.2

Kistner E.J., Hatfield J.L. 2018. Potential geographic distribution of Palmer Amaranth under current and future climates. *Agricultural and Environmental Letters* 3(1): 170044. DOI: 10.2134/ael2017.12.0044

Krapf P. 2023. Contribution of the public to the modelling of the distributions of species: Occurrence and current and potential distribution of the ant *Manica rubida* (Hymenoptera: Formicidae). *European Journal of Entomology* 120: 137–148. DOI: 10.14411/eje.2023.017

Liu C., Newell G., White M. 2016. On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution* 6(1): 337–348. DOI: 10.1002/ece3.1878

Maldonado C., Molina C.I., Zizka A., Persson C., Taylor C.M., Albán J., Chilquillo E., Rønsted N., Antonelli A. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases?. *Global Ecology and Biogeography* 24(8): 973–984. DOI: 10.1111/geb.12326

Mallen-Cooper M., Rodríguez-Caballero E., Eldridge D.J., Weber B., Büdel B., Höhne H., Cornwell W.K. 2023. Towards an understanding of future range shifts in lichens and mosses under climate change. *Journal of Biogeography* 50(2): 406–417. DOI: 10.1111/jbi.14542

Marcer A., Chapman A.D., Wieczorek J.R., Picó F.X., Uribe F., Waller J., Ariño A.H. 2022. Uncertainty matters: Ascertaining where specimens in natural history collections come from and its implications for predicting species distributions. *Ecography* 2022(9): e06025. DOI: 10.1111/ecog.06025

Martínez-Méndez N., Mejía O., Ortega J., Méndez-de la Cruz F. 2019. Climatic niche evolution in the viviparous *Sceloporus torquatus* group (Squamata: Phrynosomatidae). *PeerJ* 6: e6192. DOI: 10.7717/peerj.6192

Moudrý V., Devillers R. 2020. Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. *Ecological Informatics* 56: 101051. DOI: 10.1016/j.ecoinf.2020.101051

Namkhan M., Sukumal N., Savini T. 2022. Impact of climate change on Southeast Asian natural habitats, with focus on protected areas. *Global Ecology and Conservation* 39: e02293. DOI: 10.1016/j.gecco.2022.e02293

Newbold T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* 34(1): 3–22. DOI: 10.1177/0309133309355630

Nunes L.A., Pearson R.G. 2017. A null biogeographical test for assessing ecological niche evolution. *Journal of Biogeography* 44(6): 1331–1343. DOI: 10.1111/jbi.12910

Palacký University Olomouc. 2020. *Climatic Conditions of the Czech Republic*. Available from https://geography.upol.cz/soubory/lide/smolova/GCZ/GCZ_Klima.pdf

Phillips S.J., Dudík M. 2008. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography* 31(2): 161–175. DOI: 10.1111/j.0906-7590.2008.5203.x

Phillips S.J., Anderson R.P., Schapire R.E. 2006. Maximum entropy modeling of species geographic distribution. *Ecological Modelling* 190(3–4): 231–259. DOI: 10.1016/j.ecolmodel.2005.03.026

Pillon Y., Chase M. 2007. Taxonomic exaggeration and its effects on orchid conservation. *Conservation Biology* 21(1): 263–265. DOI: 10.1111/j.1523-1739.2006.00573.x

R Core Team. 2023. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from https://www.r-project.org/

Román-Palacios C., Wiens J.J. 2020. Recent responses to climate change reveal the drivers of species extinction and survival. *Proceedings of the National Academy of Sciences of the United States of America* 117(8): 4211–4217. DOI: 10.1073/pnas.1913007117

Salvà-Catarineu M., Romo A., Mazur M., Zielińska M., Minissale P., Dönmez A.A., Boratyńska K., Boratyński A. 2021. Past, present, and future geographic range of the relict Mediterranean and Macaronesian *Juniperus phoenicea* complex. *Ecology and Evolution* 11(10): 5075–5095. DOI: 10.1002/ece3.7395

Smith V.S., Blagoderov V. 2012. Bringing collections out of the dark. *ZooKeys* 209: 1–6. DOI: 10.3897/zookeys.209.3699

Soberón J., Peterson T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 359(1444): 689–698. DOI: 10.1098/rstb.2003.1439

Sokal R.R., Rohlf F.J. 2012. *Biometry: The Principles and Practice of Statistics in biological Research*, *4th ed.* New York: W.H. Freeman and Company. 915 p.

Spooner F.E.B., Pearson R.G., Freeman R. 2018. Rapid warming is associated with population decline among terrestrial birds and mammals globally. *Global Change Biology* 24(10): 4521–4531. DOI: 10.1111/gcb.14361

Steffelová M., Traxmandlová I., Štípková Z., Kindlmann P. 2023. Pollination strategies of deceptive orchids – a review. *European Journal of Environmental Sciences* 13(2): 110–116. DOI: 10.14712/23361964.2023.12

Štípková Z., Kindlmann P. 2015. Extent and reasons for meadows in South Bohemia becoming unsuitable for orchids. *European Journal of Environmental Sciences* 5(2): 142–147. DOI: 10.14712/23361964.2015.87

Štípková Z., Kosánová K., Romportl D., Kindlmann P. 2018. Determinants of orchid occurrence: a Czech example. In: B. Şen, O. Grillo (Eds.): *Selected Studies in Biodiversity.* London: InTechOpen. P. 133–155. DOI: 10.5772/intechopen.74851

Štípková Z., Romportl D., Kindlmann P. 2020a. Which environmental factors drive distribution of orchids? A case study from South Bohemia, Czech Republic. In: J.M. Mérillon, H. Kodja (Eds.): *Orchids Phytochemistry, Biology and Horticulture*. Cham: Springer Nature. P. 1–33. DOI: 10.1007/978-3-030-38392-3_27

Štípková Z., Tsiftsis S., Kindlmann P. 2020b. Pollination mechanisms are driving orchid distribution in space. *Scientific Reports* 10(1): 850. DOI: 10.1038/s41598-020-57871-5

Švecová M., Štípková Z., Traxmandlová I., Kindlmann P. 2023. Difficulties in determining distribution of population sizes within different orchid metapopulations. *European Journal of Environmental Sciences* 13(2): 96–109. DOI: 10.14712/23361964.2023.11

Štípková Z., Tsiftsis S., Kindlmann P. 2021. Distribution of orchids with different rooting systems in the Czech Republic. *Plants* 10(4): 632. DOI: 10.3390/plants10040632

Swarts N.D., Dixon K.W. 2009. Terrestrial orchid conservation in the age of extinction. *Annals of Botany* 104(3): 543–556. DOI: 10.1093/aob/mcp025

Tsiftsis S., Djordjević V. 2020. Modelling sexually deceptive orchid species distributions under future climates: the importance of plant-pollinator interactions. *Scientific Reports* 10(1): 10623. DOI: 10.1038/s41598-020-67491-8

Warren D.L., Glor R.E., Turelli M. 2008. Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution* 62(11): 2868–2883. DOI: 10.1111/j.1558-5646.2008.00482.x

Weterings R., Vetter K.C. 2018. Invasive house geckos (*Hemidactylus* spp.): their current, potential and future distribution. *Current Zoology* 64(5): 559–573. DOI: 10.1093/cz/zox052

Yi Y.J., Cheng X., Yang Z.F., Zhang S.H. 2016. Maxent modeling for predicting the potential distribution of endangered medicinal plant (*H. riparia* Lour) in Yunnan, China. *Ecological Engineering* 92: 260–269. DOI: 10.1016/j.ecoleng.2016.04.010

Zhang Z., Yan Y., Tian Y., Li J., He J.S., Tang Z. 2015. Distribution and conservation of orchid species richness in China. *Biological Conservation* 181: 64–72. DOI: 10.1016/j.biocon.2014.10.026

ПРИМЕНИМЫ ЛИ СВЕДЕНИЯ БАЗЫ GBIF В КАЧЕСТВЕ ИСХОДНЫХ ДАННЫХ ДЛЯ МОДЕЛИРОВАНИЯ ПРОСТРАНСТВЕННОГО РАСПРЕДЕЛЕНИЯ ВИДОВ? ИССЛЕДОВАНИЕ ИЗ ЧЕШСКОЙ РЕСПУБЛИКИ

З. Штипкова[1], С. Цифцис[2], П. Киндлманн[1,3]

[1]*Исследовательский институт глобальных изменений, Чехия*
e-mail: zaza.zuza@seznam.cz
[2]*Международный греческий университет, Греция*
e-mail: stsiftsis@for.ihu.gr
[3]*Карлов университет, Чехия*
e-mail: pavel.kindlmann@centrum.cz

Вопросы, касающиеся изучения видового разнообразия, привлекают внимание экологов и биогеографов уже более столетия, главным образом потому, что разнообразие жизни на Земле быстро сокращается, что, как ожидается, продолжится и в будущем. На настоящий момент одной из наиболее крупных баз данных о распространении видов является Global Biodiversity Information Facility (GBIF), которая содержит более 2 миллиардов находок всех организмов, и это число постоянно увеличивается с добавлением новых данных и в сочетании с другими приложениями. Такие данные также содержатся в национальных базах данных, большинство из которых, к сожалению, часто не находятся в свободном доступе и не ассоциированы с GBIF. Мы предположили, что национальные базы данных, в основном профессионально поддерживаемые правительственными организациями, могут быть более полными, чем GBIF, который не имеет централизованной организации, и что поэтому национальные базы данных могут давать более точные прогнозы распределения видов, чем GBIF. Чтобы проверить наши гипотезы, мы сравнили: (1) объем данных, включенных в базу данных Чешской Республики «Nálezová databáze ochrany přírody» (NDOP, [База данных местонахождений для охраны природы]), с объемом данных в GBIF в пределах территории Чешской Республики, и (2) перекрытие прогностических карт пространственного распределения видов в Чешской Республике на основании этих двух баз данных. В качестве модельной группы растений мы использовали семейство Orchidaceae. Мы обнаружили, что: (i) существует значительно большее количество записей для территории исследования (Чешская Республика) в базе NDOP по сравнению с базой GBIF, и (ii) прогнозы пространственного распределения видов с использованием Maxent, основанные на информации о местонахождениях орхидей в базе NDOP, в значительной степени перекрываются с таковыми, основанными на данных о местонахождениях видов в базе GBIF. Учитывая эти результаты, мы полагаем, что, если для исследуемой территории доступна только одна база данных, необходимо использовать именно ее. Если же для территории исследования доступно больше баз данных, мы должны использовать ту из них, которая включает большее количество местонахождений видов (обычно это одна из баз данных местного значения, как NDOP), поскольку использование большего количества местонахождений подразумевает более высокую значимость моделирования пространственного распределения видов.

Ключевые слова: Global Biodiversity Information Facility, NDOP, базы данных, модели распределения видов, распространение орхидей