ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Fakulta životního prostředí

Katedra prostorových věd

Jakým způsobem ovlivňuje sampling bias ekologickou interpretabilitu modelů druhové distribuce?

Bakalářská práce

Vedoucí práce: Ing. Lukáš Gábor, Ph.D.

Autor práce: Tomáš Gregor

Praha 2022

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Fakulta životního prostředí

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Tomáš Gregor

Územní technická a správní služba v životním prostředí

Název práce

Jakým způsobem ovlivňuje sampling bias ekologickou interpretabilitu modelů druhové distribuce?

Název anglicky

How does sampling bias affect the ecological interpretability of species distribution models?

Cíle práce

Cílem této bakalářské práce je s využitím virtuálního druhu otestovat, jaký vliv má sampling bias ve výskytových datech na ekologickou interpretabilitu modelů druhové distribuce.

Metodika

Druhové distribuční modely představují široce používaný nástroj v biogeografii, makroekologii a ochraně přírody, který se s postupným rozvojem stal důležitým prostředkem, využívaným například při určení lokalit potencionálně ohrožených invazními druhy nebo studování vlivu klimatických změn na biodiverzitu. S postupujícím rozvojem začalo být zřejmé, že jedním ze zásadních limitujících faktorů modelování druhové distribuce jsou vstupní data. Nicméně ověření vlivu kvality prostorových dat na modely, je s využitím reálných dat obtížné. Avšak pomocí virtuálního druhu lze snadno určit, jaký vliv mají například rozdílné metody sběru dat nebo jejich kvalita na výsledný model. Úkolem autora je zpracovat literární rešerši na téma druhových distribučních modelů, generování virtuálního druhu a vlivu sampling bias na modely druhové distribuce. V praktické části pak pomocí virtuálního druhu otestovat, do jaké míry ovlivňuje sampling bias ve výskytových datech ekologickou interpretabilitu modelů druhové distribuce.

Doporučený rozsah práce

30 stran

Klíčová slova

Formulace klíčových slov je úkolem autora

Doporučené zdroje informací

AustinM. P.2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Model.157: 101–118.

Gábor, L., Moudrý, V., Barták, V., & Lecours, V. (2020). How do species and data characteristics affect species distribution models and when to use environmental filtering?. International Journal of Geographical Information Science, 34(8), 1567-1584.

Leitão, P. J., Moreira, F., & Osborne, P. E. (2011). Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. International Journal of Geographical Information Science, 25(3), 439-454.

Meynard CH. N. & Kaplan M. D., The effect of a gradual response to the environment on species distribution modelling performance. Ecography. 2012, 35(6), 499-509.

Miller J., Species Distribution Modelling. Geography Compass. 2010, 4(6), 490-509.

Zurell D. et al., The virtual ecologist approach: simulating data and observers. Oikos. 2010, 119(4), 622-635.

Předběžný termín obhajoby

2020/21 LS – FŽP

Vedoucí práce

Ing. Lukáš Gábor, Ph.D.

Garantující pracoviště

Katedra prostorových věd

Elektronicky schváleno dne 21. 3. 2022

doc. Ing. Petra Šímová, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 22. 3. 2022

prof. RNDr. Vladimír Bejček, CSc.

Děkan

V Praze dne 23. 03. 2022

Prohlášení

Prohlašuji, že jsem bakalářskou práci na téma *Jakým způsobem ovlivňuje sampling bias ekologickou interpretabilitu modelů druhové distribuce?* vypracoval samostatně, pod vedením Ing. Lukáše Gábora, Ph.D., a že jsem uvedl všechny literární prameny, ze kterých jsem čerpal. Prohlašuji, že tištěná verze se shoduje s verzí odevzdanou přes Univerzitní informační systém.

V Praze dne 31.03.2022 .....

Poděkování

Největší poděkování patří Ing. Lukášovi Gáborovi, Ph.D., vedoucímu bakalářské práce, za nesmírnou trpělivost, odborné vedení a cenné rady, bez kterých by nebylo možné tuto práci prohlásit za hotovou. Velký dík patří také celé mé rodině za důvěru, trpělivost a podporu během celého studia a v neposlední řadě spolužákům a kamarádům za příjemně strávené chvíle během studia.

Abstrakt

Druhové distribuční modely (z anglického originálu Species Distribution Models, dále jen SDM) se staly důležitým nástrojem pro studování vztahu mezi druhem a prostředím ve kterém se vyskytuje. S rozvojem SDM začalo být zřejmé, že hlavním limitujícím faktorem v modelování druhové distribuce je přesnost druhových dat. Nejsnáze dostupná, pouze prezenční data, jsou často sbíraná ve snadno dostupných lokalitách, v okolí silnic nebo městské zástavby. Tento fenomén se nazývá sampling bias a předchozí studie ukázaly jeho negativní vliv na přesnost na SDM. Nicméně, jejich vliv na ekologickou interpretaci modelu zůstal neprozkoumán. Cílem této práce proto bylo prozkoumat vliv sampling bias na ekologickou interpretabilitu modelů. Pro účely bakalářské práce byl použit virtuální druh, pro který byly na území Španělska vygenerovány dvě datové sady. První datová sada byla vygenerována rovnoměrně po celém vhodném území. Druha datová sada byla pak vygenerována jen v blízkosti velkých silničních tahů tak, aby simulovala sampling bias v reálných datech. Pomoci modelovací techniky Maxent byly vytvořeny modely druhové distribuce. Výkon modelů byl vyhodnocen pomoci tří validačních metrik (Sorensen index, Míra nadhodnocené predikce a Míra podhodnocené predikce). Kromě toho byla též porovnána schopnost modelu odhadnout důležitost jednotlivých environmentálních proměnných a výsledné křivky, ukazující reakci druhu na změny v prostředí. Výsledky ukázaly, že ačkoliv u modelů ovlivněných sampling bias klesá jejich přesnost, ekologická interpretabilita je nicméně stále možná. Modely ovlivněny sampling bias, byly schopny detekovat důležitost environmentálních proměnných stejně jako reakci druhu na změnu prostředí, které byly srovnatelné s modely, které byly vytvořeny pomocí rovnoměrně „nasbíraných“ dat. Ačkoliv je nutné mé závěry ověřit s použitím vice environmentálních proměnných či v jiném měřítku, jsou moje závěry důležité pro budoucí aplikace SDM a to především v ochraně přírody.

Klíčová slova: modely druhové distribuce, virtuální druh, sampling bias, ekologické modelování

Abstract

Species Distribution Models (SDMs) have become an important tool for studying the relationship between species and their environment. With the development of SDMs, it has become obvious that the main limiting factor in species distribution modelling is the accuracy of the species data. The most readily available, presence-only data are often collected in easily accessible locations, around roads or urban developments. This phenomenon is called sampling bias, and previous studies have shown its negative effect on accuracy on SDM. However, their effect on the ecological interpretability of the model remained unexplored. Therefore, the aim of this thesis was to investigate the effect of sampling bias on the ecological interpretability of the model. For the purpose of this thesis, a virtual species was used for which two datasets were generated in Spain. The first dataset was generated uniformly over the suitable territory. The second dataset was then generated only in the vicinity of major road alignments to simulate sampling bias in the real data. Species distribution models were generated using the MaxEnt modelling technique. Model performance was evaluated using three validation metrics (Sorensen index, Over-prediction rate and Under-prediction rate). In addition, I also compared the model's ability to estimate the importance of each environmental variable and the resulting curves showing the species' response to environmental changes. The results showed that although the model's accuracy decreases for sampling bias-influenced models, ecological interpretability is still possible. Models influenced by sampling bias were able to detect the importance of environmental variables as well as the species' response to environmental change, which were comparable to models that were created using uniformly "collected" data. Although my conclusions need to be validated using more environmental variables or at a different scale, my findings are important for future applications of SDM, particularly in conservation.

Keywords: Species distribution models, virtual species, sampling bias, ecological modelling

Obsah

SEZNAM POUŽITÝCH ZKRATEK

AUC	Area under the ROC curve (Plocha pod ROC křivkou)
BAM	Biotic, Abiotic, Movement factors (Biotické, Abiotické a Pohybové faktory)
BRT	Boosted regression trees (Posílený regresní strom)
ENFA	Ecological niche factor analysis (Analýza faktorů ekologické niky)
FN	Fundamental niche (Fundamentální nika)
GAM	Generalized additive model (Generalizovaný aditivní model)
GARP	Generatic Algorithm for Rule Set Production (Generický algoritmus pro tvorbu pravidel)
GBIF	Global Biodiversity Information Facility (Globální informační zařízení pro biologickou rozmanitost)
GIS	Geographic Information System (Geografický informační systém)
GLM	Generalized linear model (Generalizovaný lineární model)
GNSS	Global Navigation Satellite System (Globální družicový polohový systém)
GPS	Global Positioning System (Globální polohový systém)
MAXENT	Maximum entropy model (Model maximální entropie)
OBIF	The Ocean-Bottom Instrumentation Facility (Zařízení pro měření na dně oceánu)
OPR	Overprediction rate (Míra nadhodnocené predikce)
RN	Realized niche (Realizovaná nika)
ROC	Receiver Operating Characteristic Curve (Operační charakteristická křivka)
SDM	Species distribution model (Model druhové distribuce)
SI	Sorensen index (Sorensenův index)
TSS	True skill statistic (Statistika skutečných schopností modelu)
UPR	Underprediction rate (Míra podhodnocené predikce)

1. Úvod

Druhová data jsou stále intenzivněji, vědci i širokou veřejností, zaznamenávána do veřejných globálních databází, jako jsou GBIF (www.gbif.org), eBird (www.ebird.org) nebo iNaturalist (www.inaturalist.org). Zatímco počet záznamů v těchto databázích neustále roste, je čím dál tím jasnější, že přesnost těchto dat je velmi diskutabilní (Moudrý & Devillers 2020). V praxi jsou obvykle nejčastěji dostupné záznamy o druzích získané pomocí nesystematických sběrů dat (například díky pozorování lidi, kteří jdou na výlet a cestou zaznamenají náhodný výskyt několika ptačích druhů). Tento typ dat označujeme jako pouze prezenční data (Guillera-Arroita et al. 2015) a často trpí tzv. sampling bias, tedy nadměrným sběrem v některých oblastech. Sampling bias je často vázán na data sbíraná ve snadno přístupných lokalitách, ve chráněných oblastech, nebo v oblastech s vysokou hustotou zalidnění (Boakes et al. 2010, Geldmann et al. 2016). Předchozí studie ukázaly, že je při modelování druhové distribuce nutné brát sampling bias v potaz, protože negativně ovlivňuje přesnost modelu (Leitão et al. 2011, Boria et al. 2014). Pro kompenzování tohoto negativního vlivu byly navrženy různé metody, jako je například manipulace s background daty při použití modelovací techniky Maxent (Phillips et al. 2009) nebo environmentální a geografické (prostorové) filtrování (Tessarolo et al. 2014, Varela et al. 2014). Jejich použití nicméně vede ke snížení celkového počtu dat, která mohou být použita pro tvorbu modelu a navíc, jak bylo ukázáno, jejich aplikace je velmi diskutabilní (Gábor et al. 2020).

SDM se také často používají ke zjištění důležitosti environmentálních proměnných na druhovou distribuci (Smith & Santos 2020) nebo ke zjištění, jakým způsobem reaguje druh na změny v prostředí (Dvorský et al. 2017). Proto je překvapivé, že až do současné doby všechny předchozí studie zaměřené na sampling bias zkoumaly pouze vliv na přesnost modelů, zatímco ekologická interpretabilita zůstávala stranou.

Cílem této bakalářské práce je prozkoumat, jakým způsobem ovlivňuje sampling bias ekologickou interpretabilitu modelů. Konkrétně jsem se v této práci zaměřil na vliv sampling bias na schopnost modelů detekovat důležitost environmentálních proměnných a schopnost modelů správně detekovat vliv měnícího se prostředí na druhouvou distribuci.

2. Literární rešerše

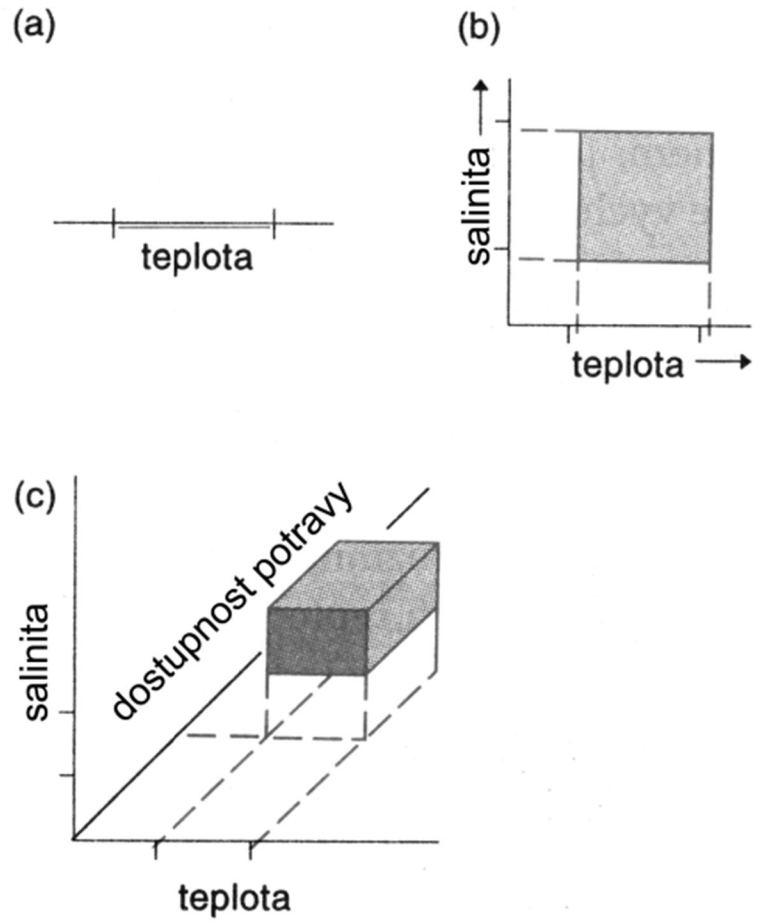
2.1 Ekologická nika

Pod pojmem ekologická nika chápeme komplexní začlenění druhu v prostředí. Ekologická nika zahrnuje zapojení druhu v potravních sítích (potravní nároky), požadavky na další zdroje (světlo, voda, minerální látky), jeho prostorové nároky (umístění hnízda, místa výskytu, odpočinku, úkryty), časové rozložení aktivity (denní a sezónní rytmy), požadavky na místa a období rozmnožování a další životní projevy (Studijní texty předmětu Z0025 ekologie a životní prostředí; elektronická učebnice 2013).

Věda zná výraz nika od roku 1917, kdy jej poprvé ve svém článku o Kalifornském drozdu použil americký biolog Joseph Grinnell. Jeho koncept niky byl založen na distribuci, která byla dána fyzickými a klimatickými bariérami, s malým důrazem na zásoby potravy nebo interakce s jinými druhy (Gaffney 1975). Moderní definice ekologické niky pochází od Evelyna Hutchinsona, který formálně definoval metodu, jíž lze popsat způsob života organismu pomocí vymezení požadavků, tolerancí a jejich vzájemných interakcí (Hutchinson 1957, Jarošík 1987).

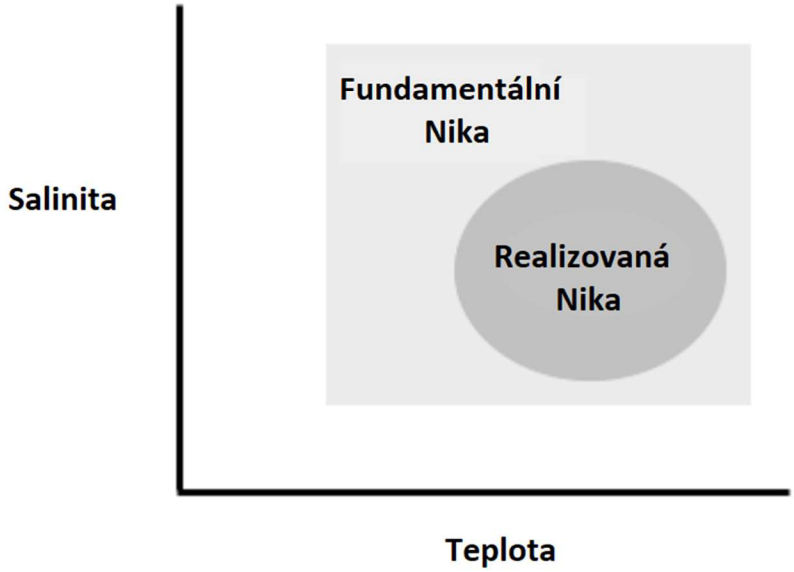
Hutchinsonova definice (1957) se nejlépe ilustruje na příkladech. Podle Hutchinsona organismy jakéhokoli druhu mohou přežívat, růst, rozmnožovat se a zachovávat životaschopnou populaci jen v určitém rozmezí faktorů ovlivňujících jejich distribuci (např. teplota, salinita, srážky apod.). To se dá demonstrovat například na teplotě, která může ekologickou niku ovlivňovat. Teplotní rozsah je ekologickou nikou daného druhu v jednom rozměru (Obr. 1.a). Jistě, na organismus ale nebude působit jen teplota. Daný organismus může kupříkladu zároveň tolerovat pouze určitý rozsah salinity. Uvažujeme-li o obou faktorech současně, stává se ekologická nika dvourozměrnou a může být zobrazena jako plocha (Obr. 1.b). Vezmeme-li v úvahu další podmínky a zdroje (což bychom nepochybně měli), pak je dalším krokem trojrozměrná nika, kterou můžeme znázornit jako objem (Obr. 1.c). Je nasnadě, že začlenění více než tří rozměrů již nelze graficky znázornit. V tomto

procesu se dá abstraktně pokračovat a ekologickou niku si představit jako n-rozměrný nadprostor. Takové pojetí je dnes jedním ze základních kamenů ekologického myšlení (Jarošík 1987).



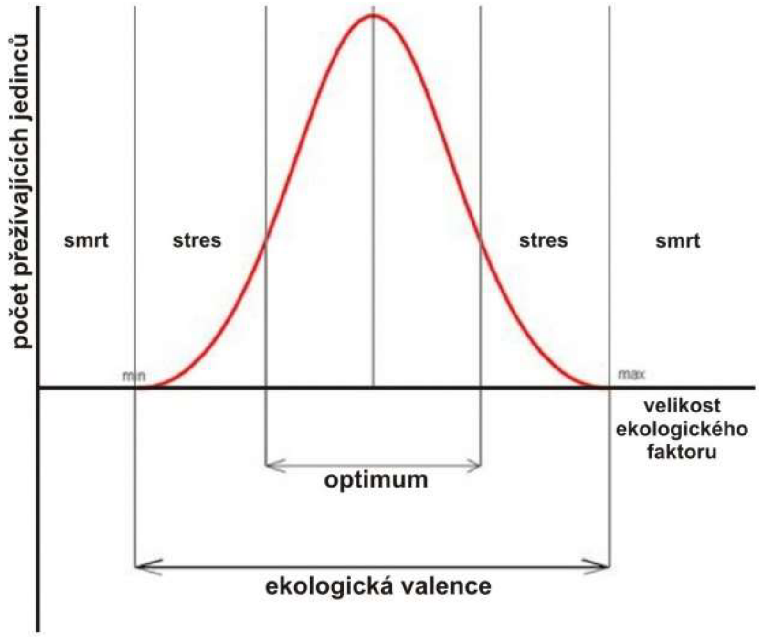
Obr. 1 Teoretická definice niky podle Hutchinsona (1957). Ekologická nika a) v jednom (teplota), b) dvou (teplota a salinita) a c) třech (teplota, salinita a dostupnost potravy) rozměrech. Upraveno podle Begona et al. (1997)

Ekologickou niku rozlišujeme na fundamentální (základní) a realizovanou (Obr. 2). Fundamentální nika je výsledkem evoluční historie druhu a představuje geneticky daný potenciál jeho funkčního zapojení. Realizovaná nika je vždy užší a je výsledkem konkrétní situace v obývaném prostředí. K jejímu omezení dochází jak vlivem abiotických podmínek (teplota, salinita), tak nejrůznějšími vztahy k ostatním druhům (potravní nabídka, konkurence) (Studijní texty předmětu Z0025 ekologie a životní prostředí; elektronická učebnice 2013).



Obr. 2 Grafické znázornění rozpětí fundamentální a realizované niky působením abiotických proměnných (salinity a teploty)
(Studijní texty předmětu Z0025 ekologie a životní prostředí; elektronická učebnice 2013).

Tolerované rozmezí působení kteréhokoli ekologického faktoru nazýváme ekologická valence (Obr. 3). Organismy vyžadují ke své zdárné existenci určitou teplotu, vlhkost, potravu, stanoviště apod. Organismus nejlépe prospívá, tj. dosahuje nejvyšší zdatnosti, v oblasti optima. Zdatnost lze chápat jako schopnost mít nejvíce potomků a tak nejvíce ovlivnit genofond potomstva populace, což je důležité k přežití budoucích populací druhu.



Obr. 3 Ekologická valence znázorněna pomocí Gaussovy křivky
(Studijní texty předmětu Z0025 ekologie a životní prostředí; elektronická učebnice 2013).

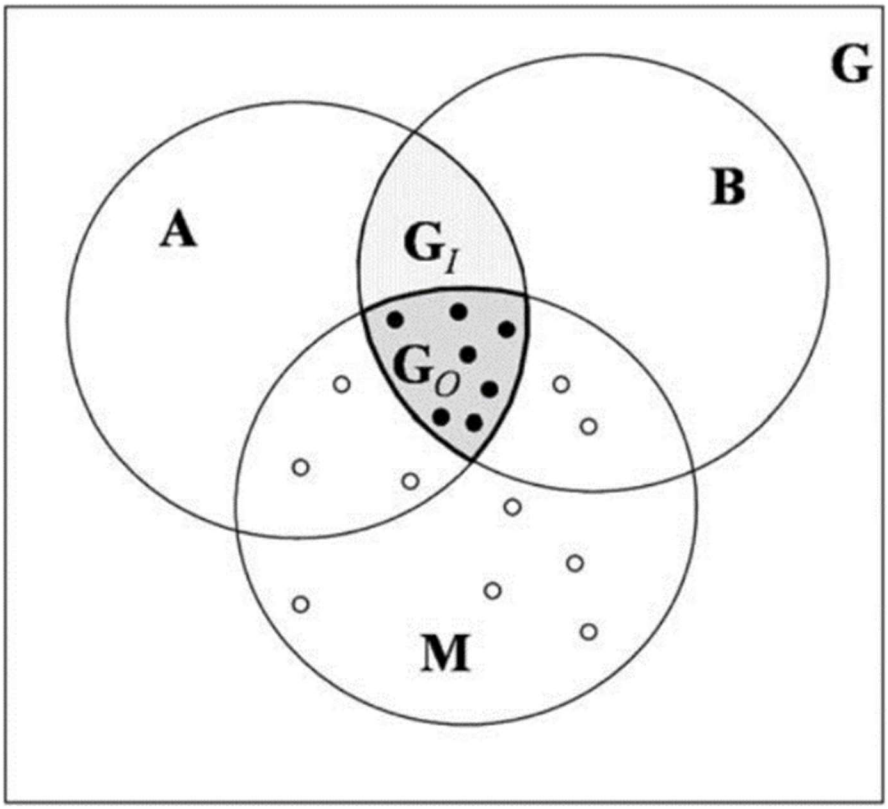
Na obě strany od optima se životní projevy nebo některá z životně důležitých funkcí zpomalují až do situace, kdy reprodukční schopnost již nedokáže kompenzovat úmrtnost. Překročení mezí ekologické valence tedy obvykle nevede k smrti jedince, ale populace v daném prostředí nemůže trvale existovat, rozmnožování je ztížené nebo k němu nedochází, v některém stádiu vývoje nastává vysoká mortalita. Díky modelům druhové distribuce můžeme získat výsledky, jež nám usnadňují chápání rozšíření druhů na územích (Studijní texty předmětu Z0025 ekologie a životní prostředí; elektronická učebnice 2013).

2.2 BAM diagram

Hutchinson (1987) si představoval fundamentální niku jako širokou, klimaticky určenou abiotickou niku, která je do určité míry redukována biotickými interakcemi nezbytnými pro přetrvávání druhu v regionu, aby se vytvořila „realizovaná“ nika, která zdůrazňuje roli negativních interakcí při snižování ekologického potenciálu druhů. Hutchinson (1978) však zanedbával účinky geografické heterogenity. Soberón a Peterson (2005) předložili rámec pro pochopení geografického rozšíření druhů v podobě BAM (Biotic, Abiotic a Movement, v češtině biotický, abiotický, pohybový) diagramu. Dodatečný rámec M od Soberón & Peterson (2005) tak zahrnuje i úvahy o pohybu a přístupu druhu jako další omezení distribučního potenciálu druhů. Tato kombinace aspektů biotických interakcí, abiotických podmínek a pohybu (odtud „BAM“) nastiňuje hlavní faktory ovlivňující distribuční potenciál druhů.

BAM diagram (Obr. 4) je abstraktní reprezentace geografického prostoru (Soberón & Peterson 2005 , Soberón 2007, Soberón & Nakamura 2009). Diagram je rozdělen do oblastí, kde tři hlavní faktory omezují distribuci druhu (Peterson 2006 ,Hirzel & Le Lay 2008): abiotické (A), biotické (B) a historické nebo rozptylové (M) faktory. Mimo prostor vymezený třemi faktory je stanoviště pro daný druh nevhodné. Nevhodná je však i mimo oblast společnou všem třem faktorům. Tato oblast představuje oblasti v prostoru, kde se druh skutečně vyskytuje. Oblast společná pro A a B proto představuje oblasti v prostoru,

kde se vyskytuje realizovaná nika (RN); a A představuje oblasti v prostoru, kde se vyskytuje základní nika (FN). Vhodná oblast definovaná rozptylovými faktory odpovídá především populacím (Pulliam 2000).



Obr. 4 BAM (biotický, abiotický, pohybový) diagram. Oblast G představuje celou uvažovanou zeměpisnou oblast. Oblast A je oblast, ve které jsou pro druh příznivé abiotické podmínky. Oblast B je oblast, ve které jsou pro daný druh vhodné biotické podmínky. Oblast M je oblast, kterou je druh schopen dosáhnout v daném časovém období. Gi představuje oblast, kde se daný druh může vyskytovat díky příznivým podmínkám oblasti A a B, ale jelikož oblast může být potencionálně napadena predátorem, tak se do ní dosud druh nedostal. Konečně Go představuje skutečnou oblast rozšíření druhu. Černé body představují zdrojové populace. Prázdné body představují populace, jež strádají jeden z faktorů pro úspěšné rozmnožení (Soberón 2010).

2.2.1 B faktor

Faktor B pracuje se vztahy a s interakcemi mezi organismy v ekosystému. Většina modelů druhové distribuce (z anglického originálu Species Distribution Models, dále jen SDM) pracuje s předpokladem, že biotické interakce nemají vliv na výskyt druhů (Huntley et al. 1995, Bakkenes et al. 2002), případně s domněním malého rozsahu ovlivnění (Pearson & Dawson 2003, Dormann et al. 2007, Heikkinen et al. 2007). Tato tvrzení však

vyvrátil Araújo et al. (2014) podklady, které demonstrují, jak důležité je zahrnutí biotických interakcí do SDM (Araújo & Luoto 2007).

Toto vzájemné působení organismů lze rozdělit na vnitrodruhové a mezidruhové. Vnitrodruhové působení znamená, že probíhá kompetice mezi organismy stejného druhu např. o zdroje nebo o partnera/ku k následnému rozmnožení. Ovšem k interakci nemusí docházet jen mezi stejnými druhy, ale také mezi druhy odlišnými. V tomto případě se jedná o mezidruhové vztahy. Vztahy mohou nabývat tří typů - pozitivní, negativní a neutrální. Pozitivní jsou komenzálismus (jeden těží výhody a druhého to nijak neovlivňuje) a symbióza (kdy je vztah pro oba prospěšný). Mezi negativní řadíme parazitismus (kde máme parazita, který se pase na hostiteli) a predaci (kdy je na jedné straně predátor a na druhé kořist). Při neutrálním vztahu se organismy vzájemně neovlivňují (Soberón & Peterson 2005).

2.2.2 A faktor

Vhodné abiotické podmínky umožnují přetrvání druhu v areálu jeho rozšíření (Sobéron & Peterson 2005). Pulliam (2000) zdůrazňuje důležitost abiotických podmínek pro výskyt druhu, a proto vyvinul simulaci založenou na kvantitativním popisu abiotických podmínek. Abiotický faktor je neživá část ekosystému, která utváří jeho prostředí. V pozemském ekosystému mohou příklady zahrnovat teplotu, světlo, vliv klimatu a edafické podmínky. V mořském ekosystému by abiotické faktory zahrnovaly slanost vody a mořské proudy. Abiotické a biotické faktory spolupracují na vytvoření jedinečného ekosystému.

2.2.3 M faktor

Faktor M (Movement) je určen určitou kombinací současné schopnosti rozptylu (např. kapacita pro pohyb mezi natálními a reprodukčními místy) a historickými posuny kontinentů, které otevřely přístup do oblastí relativně vzdálenějších od současných distribučních oblastí. (Soberón & Peterson 2005). Důležitost posledně jmenovaného fenoménu je doložena četností izolovaných populací, jejíž rozsah byl během dřívějších epoch mnohem širší (např. Smith et al. 2000). Migrace byla ovšem často zastavena „tvrdými“ bariérami (tj.

bariérami, které se neposouvají s měnícími se klimatickými podmínkami), jako jsou například břehy velkých vodních toků, horská pásma, hluboká údolí. Tyto rysy, alespoň obecně, zůstaly do značné míry neměnné, i když došlo k velkým klimatickým posunům, například během posledních milionů let, s událostmi zalednění v pleistocénu (Peterson 2009).

2.3 Druhové distribuční modely

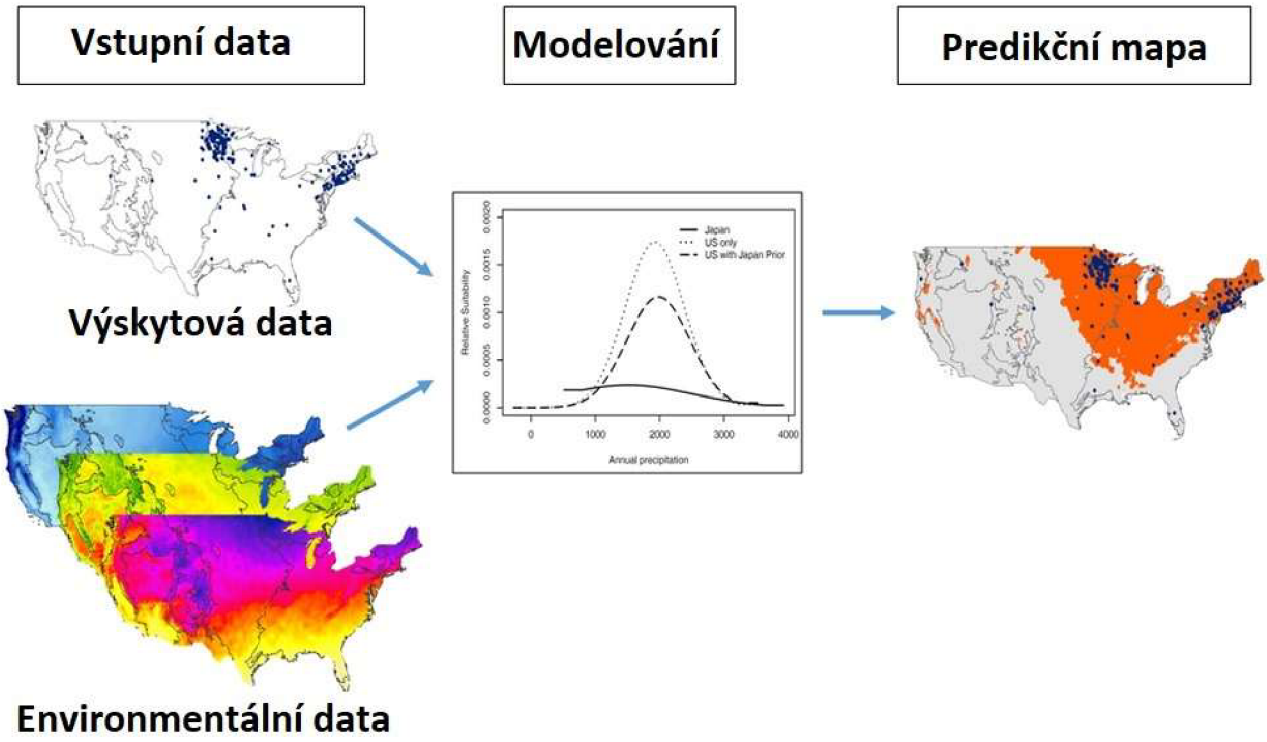
Modely druhové distribuce mají svůj původ na konci 70. let minulého století, kdy ještě ovšem byla omezena výpočetní technika pro velkoobjemové výpočty a také byla značně omezena znalost o využití GIS v ekologii. První práce v této oblasti se soustředila především na vývoj metod k efektivnímu modelování druhové distribuce (Austin 1987, Austin et al. 1990). Metodika a její rámec byly shrnuty před více než 10 lety a tyto syntézy jsou stále široce používány jako referenční orientační body v současné literatuře modelování distribuce. (Franklin 1995, Guisan 2000, Zimmermann et al. 2010).

SDM jsou statistické nástroje, které kombinují pozorování výskytu s environmentálními proměnnými (Obr. 5) (Jane Elith & John R. Leathwick 2009), aby bylo možné buď popsat vztahy mezi nimi anebo vymezit chráněná území (Gaubert Papes & Peterson 2006), nebo vytvořit mapy druhové bohatosti (Loiselle et al. 2003; Rissler et al. 2006), a k predikování geografického rozšíření invazních druhů nebo studiu vlivu klimatických změn na distribuci druhu (Peterson 2003).

Údaje o výskytu druhů jsou čím dál tím více dostupné, a to především díky globálním databázím jako GBIF (Global Biodiversity Information Facility) (Graham et al. 2004 ; Suarez & Tsutsui 2004). Zatímco údaje o výskytu druhů poskytují základ pro mnoho výzkumů, jejich použití může být problematické, protože obsahují pouze prezenční záznamy, množství záznamů je často malé (především u ohrožených druhů) a navíc tyto data mohou obsahovat geografické zkreslení a nízkou polohovou přesnost. Ačkoliv jsou přesná data pro SDM velmi

důležitá, existuje jen málo studií, které se zabývají vlivem různé kvality druhových dat na modely druhové distribuce (Hijmans et al. 2000 ; Elith et al. (2006).

Robustnost modelu může být ovlivněna výběrem relevantních environmentálních prediktorů modelovacího algoritmu, nebo měřítkem, ve kterém je distribuce predikována (rozlišení environmentálních prediktorů, rozloha studované oblasti) (Jane Elith & John R. Leathwick 2009).



Obr. 5 Proces tvorby modelu druhové distribuce.
(https://earthdata.nasa.gov/learn/pathfinders/biodiversity/species-distribution)

2.4 Modelovací algoritmy

Výběr správného algoritmu není vzhledem k množství existujících algoritmů, které se liší typem vstupních dat ((prezenčními (PO – only presence), prezence-absenčními daty (PA – presence-absence) nebo s daty prezence a pozadí (PB – presence-background)) nebo použitím statistických metod (glm, gam, brt, machine learning) modelů a možností

aplikace mnoha různými způsoby jednoduchým procesem. V současné době podporuje SDM více jak 15 modelovacích algoritmů (Tab. 1) (Naimi & Araujo 2016).

Modelovací algoritmy
Modely presence-absenční
GLM (Obecný linearní model)
GAM (Obecný aditivní model)
MARS (Multivariační adaptivní regresní spline)
ANN(Umělé neuronové sítě)
CART (Klasifikační a regresní strom)
BRT (Posílený regresní strom)
Modely prezenční
ENFA (Environemtální analýza faktorů)
Bioclim
Domain
Mahalanobis
Maxlike
RF (Náhodné lesy)
Model prezence a pozadí
Maxent (Model maximální entropie)

Tab. 1 Přehled modelovacích algoritmů (Naimi & Araujo 2016).

Když jsou pro vývoj modelů k dispozici výskytová prezence-absenční data druhu ve studované oblasti, v takovém případě lze použít obecné statistické metody (přehled rozmanitosti aktuálně používaných technik (viz Corsi et al. 2000, Guisan & Zimmerman 2000, Elith 2002, Scott et al. 2002). Pro modelování s datovými soubory prezence-absence se běžně používají obecné statistické metody, jako jsou zobecněné lineární modely (GLM) a zobecněné aditivní modely (GAM) (Phillips et al. 2006).

Většinou jsou ale přístupná pouze prezenční data. Pro taková data se dá použít BIOCLIM (Busby & Nix 1986), který předpovídá vhodné podmínky v „bioklimatickém obalu“, sestávajícím z přímočaré oblasti v environmentálním prostoru představující rozsah (nebo určité procento z nich) pozorovaných hodnot prezence v každé dimenzi prostředí. Podobně DOMAIN (Carpenter et al. 1993) používá metriku podobnosti, kde je předpovězený index vhodnosti dán výpočtem minimální vzdálenosti v prostoru prostředí k jakémukoliv záznamu

prezence (Phillips et al. 2006). GARP (Stockwell & Noble 1992 , Stockwell & Peters 1999) používá rámec umělé inteligence nazývaný genetické algoritmy. Vytváří sadu pozitivních a negativních pravidel, která společně dávají binární předpověď. Ecological niche factor analysis (ENFA, Hirzel et al. 2002) používá lokality prezence spolu s údaji o životním prostředí pro celou studovanou oblast, aniž by vyžadoval, aby byl vzorek pozadí považován za absence. Je podobný analýze hlavních komponent, který zahrnuje lineární transformaci environmentálního prostoru na ortogonální faktory „marginality" a „specializace". Vhodnost prostředí je pak modelována jako vzdálenost Manhattanu v transformovaném prostoru (Phillips et al. 2006).

S daty prezence a pozadí pracuje model Maxent (maximum entropy model, v češtině model maximální entropie). Maxent, předpovídá výskyt druhů tak, že najde rozdělení, které je nejvíce rozprostřené nebo nejblíže rovnoměrně rozložené v celé studované oblasti, přičemž bere v úvahu limity proměnných prostředí známých lokalit. Maxent používá pouze prezenční údaje a algoritmus porovnává místa, kde byl druh nalezen, se všemi prostředími, která jsou ve studovaném regionu k dispozici. Tato dostupná prostředí definuje výběrem velkého počtu background bodů v celé studované oblasti, které se označují jako body pozadí (body pozadí zahrnují lokality, kde je znám výskyt druhu) (https://support.bccvl.org.au/support/solutions/articles/6000083216-maxent). V poslední době je široce používán, protože si vede lépe ve srovnání s ostatními modely distribuce druhů (Elith et al. 2006).

2.5 Validace modelů

Validace neboli hodnocení je stěžejní součástí druhových distribučních modelů. Toto hodnocení nám umožňuje kvantifikovat výkonnost modelů z hlediska toho, jak dobře předpovědi odpovídají pozorováním, což je základní a objektivní součástí každé teoretické, aplikované nebo metodické studie.

Metriky, které jsou pro hodnocení SDM často používané, se zakládají na použití takzvané chybové matice, tj. matice (Tab. 2) porovnávající předpovězené a pozorované prezence a absence (Alloche et al. 2006).

		Data získaná pozorováním	
		Prezence	**Absence**
Simulovaná data	**Prezence**	**a** skutečně pozitivní	**b** falešně pozitivní
	Absence	**c** falešně negativní	**d** skutečně negtivní

Tab. 2 Chybová matice složená z presence a absence dat (Kienast *et al.* 2012)

V literatuře o SDM nejvíce rezonují tři klasifikační metriky, tj. plocha pod křivkou operační charakteristiky (AUC, Area under the ROC curve), statistika skutečných dovedností (TSS, True skill statistic) a dříve používaná Cohenova KAPPA. AUC zavedli v ekologii Fielding & Bell (1997), ale od té doby byla opakovaně kritizována (Lobo, Jiménez-Valverde & Real 2008; Lobo, Jiménez-Valverde & Hortal, 2010; Jiménez-Valverde, 2012) kvůli své závislosti na prevalenci (tj. podíl zaznamenaných lokalit, kde se daný druh vyskytuje). A navíc je založena na použití prezenčně-absenčních dat, které se však používají velice málo. Cohenův Kappa byl také opakovaně kritizován ze stejného důvodu (McPherson, Jetz & Rogers 2004; Allouche, Tsoar & Kadmon 2006; Lobo et al. 2010). Na druhé straně TSS (Peirce 1884) se od svého zavedení dařilo poměrně dobře Allouchem et al. (2006), a to především proto, že byla prokázána jako nezávislá na prevalenci. V poslední době byla tato metrika zpochybněna, protože byla původně navržená pro prezenčně-absenční data (Leroy B, Delsol R, Hugueny B et al. 2018).

Přehled validačních metrik		
Metriky	Popis	Výpočet
Sensitivity	poměr správně předvídané pozorované prezence	Sn=a/(a+c)
Specificity	poměr správně předvídané pozorované absence	Sp=d(b+d)
True skill statistic	statistika skutečné dovednosti	TSS=Sn+Sp-1
Jaccard's similarity index	Jaccardův index podobnosti	Jaccard=a(c+2a+d)
Sorensen's similarity index	Sorensenův index podobnosti	Sorensen=2a(c+2a+d)
Proxy of F-measure	zástupce F-měření	Fpb= 2*Jaccard
Overprediction rate	míra nadhodnocení predikce	Opr=d/(a+d)
Underprediction rate	míra podhodnocení predikce	Upr=c/(a+c)=1-Sn
Kappa	procento shody	Kappa=[(a+d)±(((a +c)(a+b)+(b+d)(c+d))/n)] [n ± (((a+c)(a +b)+(b+d)(c +d))/n)
PPP	procento skutečné predikované absence	PPP=a/(a+b)
NPP	procento skutečné predikované prezence	NPP=d/(c+d)

Tab. 3 Přehled validačních metrik *(Leroy B, Delsol R, Hugueny B et al. 2018).*

Protože výše uvedené metriky byly čím dál častěji kritizovány, navrhl Leroy (2018) používat Sorensenový index a Jaccardový index, které při vyhodnocení modelu nepoužívali falešně negativní data (Tab. 2), a proto jsou vhodnější pro prezence-only modely.

Kromě výše uvedených metrik se dají k vyhodnocení modelů použít také metriky zkoumající míru podhodnocení a nadhodnocení predikované vhodnosti habitatu OPR „míra nadhodnocené predikce“ a UPR „míra podhodnocené predikce“. OPR měří do jaké míry model nadhodnocuje vhodnosti habitatu (Barbosa, Real, Muñoz, & Brown 2013). UPR měří do jaké míry model podhodnocuje vhodnost habitatu (Fielding & Bell 1997). Při hodnocení modelu by se vždy mělo používat více metrik, což umožní získat úplný přehled o přesnosti modelu a díky tomu přesněji interpretovat výsledné výstupy (Leroy B, Delsol R, Hugueny B et al. 2018).

Obecně u většiny metrik je rozsah 0 až 1 a čím vyšší je číslo, tím je lepší přesnost modelu, přičemž hodnoty pod 0,5 zejména u AUC a Schoner indexu ukazují náhodnou predikci, která se reálně nedá použít (Leroy B, Delsol R, Hugueny B et al. 2018).

2.6 Vstupní data

Pro modelování druhové distribuce jsou potřeba dva typy vstupních dat, tj. záznamy o výskytu druhu získané například pomocí GPS a environmentální proměnné.

2.6.1 Environmentální data

Environmentální data jsou složena z vrstev vzniklých využitím dat získaných dálkovým průzkumem Země a aplikací nástrojů GIS, jež musí podle Millera (2010) představovat vhodnou kombinaci přímých a nepřímých proměnných. Přímé proměnné jsou např. světlo, teplota, vzduch, voda, půdní živiny apod., a nepřímé proměnné jsou např. půdní struktura, nadmořská výška, sklon svahu, vítr. Jako opětovně používané prediktory při tvorbě SDMs jsou napříč odbornou veřejností aplikovány bioklimatické a topografické proměnné, jejichž kombinace dokáže vystihnout fyziologickou závislost druhů na vodu, teplotu a ve větších měřítkách i druhovou variaci, závislost na vlhkosti prostředí nebo tok energií v rámci ekosystému (Kienast et al. 2012).

Bioklimatické proměnné představují roční trendy (např. průměrná roční teplota, roční srážky), sezónnost (např. roční rozsah teplot a srážek) a extrémní nebo limitující faktory prostředí (např. teplota nejchladnějšího a nejteplejšího měsíce a srážky mokrých a suchých prostor). Měsíční údaje obsahují informace o klimatu pro minimální, střední a maximální teploty, srážky, sluneční záření, rychlost větru, tlak vodní páry a úhrn srážek. Často používaným zdrojem těchto proměnných může být např. Worldclim (https://www.worldclim.org), kde je možné stáhnout celkem 12 proměnných (aktuální a budoucí podmínky počasí a klimatu) pro celý svět ve čtyřech prostorových rozlišeních, mezi 30 sekundami (~1 km^2) až 10 minutami (~340 km^2) (https://www.worldclim.org/-data/bioclim.html).

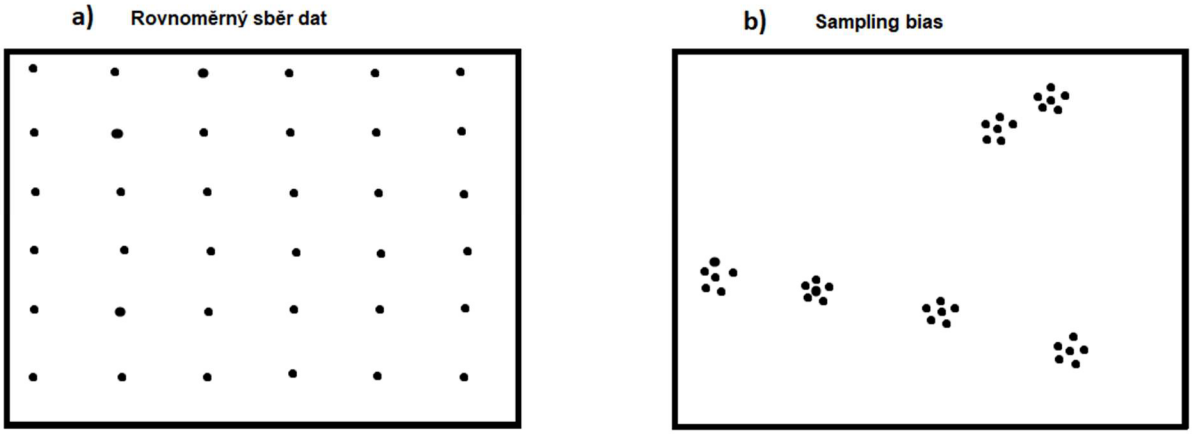
Topografická data poskytují informace o svažitosti terénu, proudění vody, množství slunečního záření nebo nadmořské výšce v dostupných souborech topografických dat (Macdonald, H. 2004).

2.6.2 Druhová data

Výskytová data jsou získávána z odborných terénních průzkumů, ale také díky pozorování laiků napříč celým světem. Cenným zdrojem dat jsou také muzejní databáze, přírodovědné sbírky, jež poskytují rozsáhlé informace více než 2,5 miliardy vzorků po celém světě v muzeích, herbářích a dalších institucích. Tyto údaje jsou stále častěji zpřístupňovány prostřednictvím internetových portálů ((GBIF (Global Biodiversity Information Facility), iNaturalist, OBIF (The Ocean-Bottom Instrumentation Facility)). Klíčovou výzvou při používání těchto dat je nejistota ohledně přesnosti pozorování, která je spojena i s moderními systémy GNSS (Global Navigation Satellite System). Navíc velká část dat, která byla shromážděna před popularizací technologie GPS, byla zaznamenána jako textové popisy místa, kde byl nález zaznamenán, a proto může být získaní přesných souřadnic obtížné, často téměř nemožné (poloha může být i několik kilometrů nepřesná). Tento problém, takzvaná polohová nejistota, se stává důležitým, když jsou data použita k modelování druhové distribuce. Souřadnice se používají k extrakci společně umístěných proměnných prostředí, a proto se polohová nejistota přenese na nepřesné charakterizace vztahu mezi druhem a prostředím (Naimi et al. 2014, Gabor et al. 2020).

Výskytová data jsou také ovlivněna tzv. sampling bias. Sampling bias je způsoben sběrem dat ve snadno dostupných lokalitách, v blízkosti silnic, městských sídel a řek, a proto získaná data nemusí reprezentovat skutečný rozsah podmínek prostředí, ve kterých se druh vyskytuje (Obr. 6). Sampling bias je často považován za jeden z hlavních faktorů, které mají negativní dopad na SDM (např. Araujo & Guisan 2006, Leitão et al. 2011, Duputié et al. 2014, Guillera-Arroita et al. 2015, Ranc et al. 2016).

Ačkoliv se vlivem sampling bias na SDM v poslední dekádě zabývalo mnoho studií např. filtrování dat v geografickém nebo environmentálním prostoru (Varela et al. 2014), v současné době neexistuje žádný nástroj, který by efektivně umožnil obejít negativní vliv sampling bias na přesnost modelu. Proto modely druhové distribuce, u kterých se použijí takto zkreslená data, ukazují spíše úsilí při sběru dat než skutečnou distribuci druhu. Proto je pro budoucí studie klíčové pochopit, jak sampling bias ovlivňuje všechny aspekty SDM (Syfert et al. 2013).

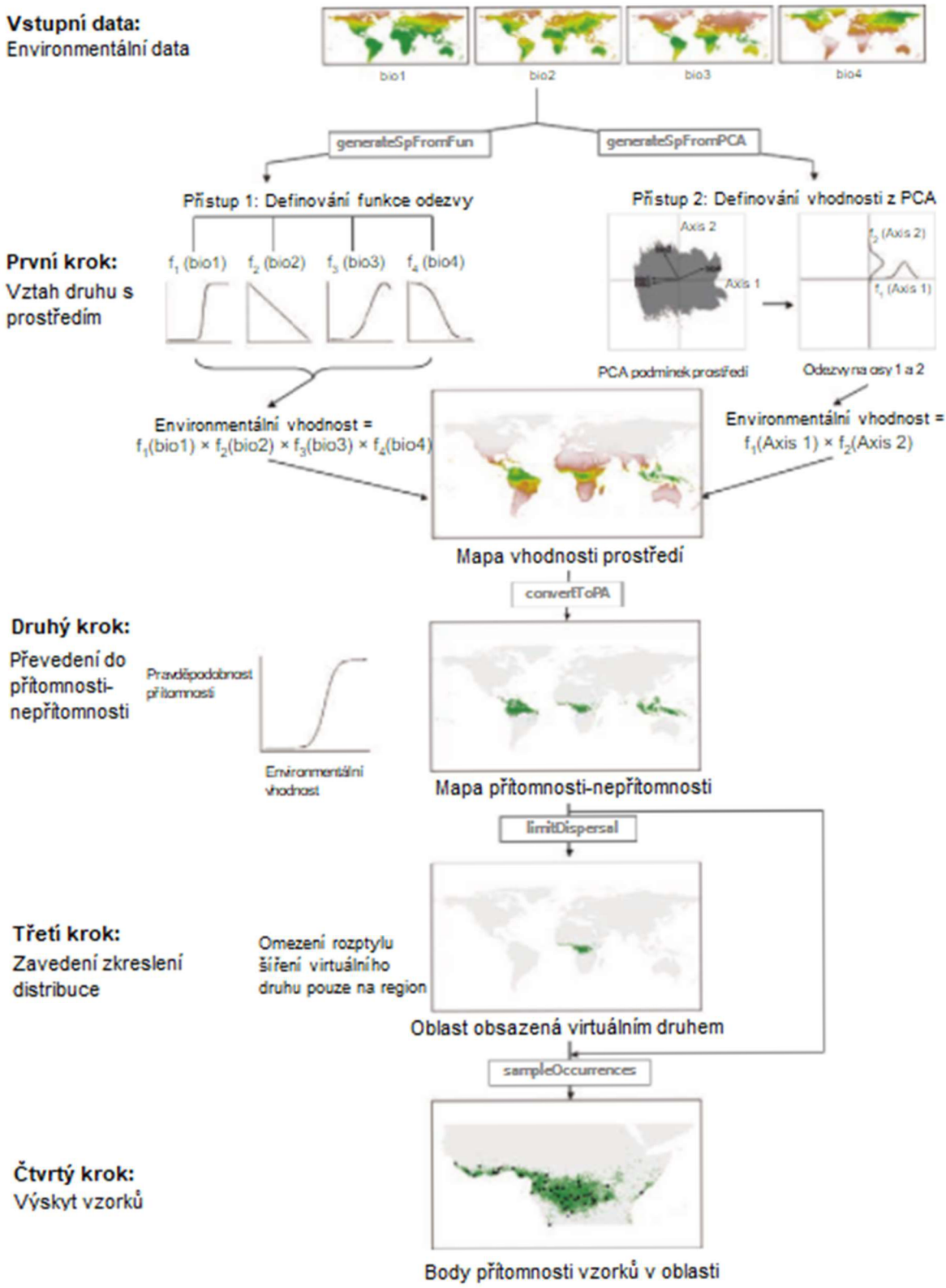


Obr. 6 Rozdíl mezi způsoby měření dat ze stejné oblasti, kde (a) jsou rovnoměrně sesbíraná data, tedy máme informace z celé studované oblasti a (b) nerovnoměrně sesbíraná data z dostupných lokalit (komunikací a sídel), kde dochází k sampling bias.

2.7 Virtuální druh

Testování SDM je s reálnými daty vždy velkou výzvou, protože nejistota reálných dat je většinou neznámá a navíc není možné přesně kvantifikovat charakteristiky zkoumaného druhu. Proto se pro metodologické studie řešící vliv různé kvality prostorových dat začal používat virtuální druh (Hirzel et al. 2001; Meynard & Quinn, 2007 ; Elith & Graham, 2009).

Virtuální druh představuje efektní způsob, jak mít plnou kontrolu, jak nad kvalitou a kvantitou výskytových dat, tak nad ekologickými vlastnostmi druhu. Generování virtuálního druhu se skládá ze čtyř na sebe navazujících kroků (Obr. 7):



Obr. 7 Tvorba virtuálního druhu (Leroy *et al.* 2015)

1) definování reakce virtuálního druhu na environmentální proměnné,

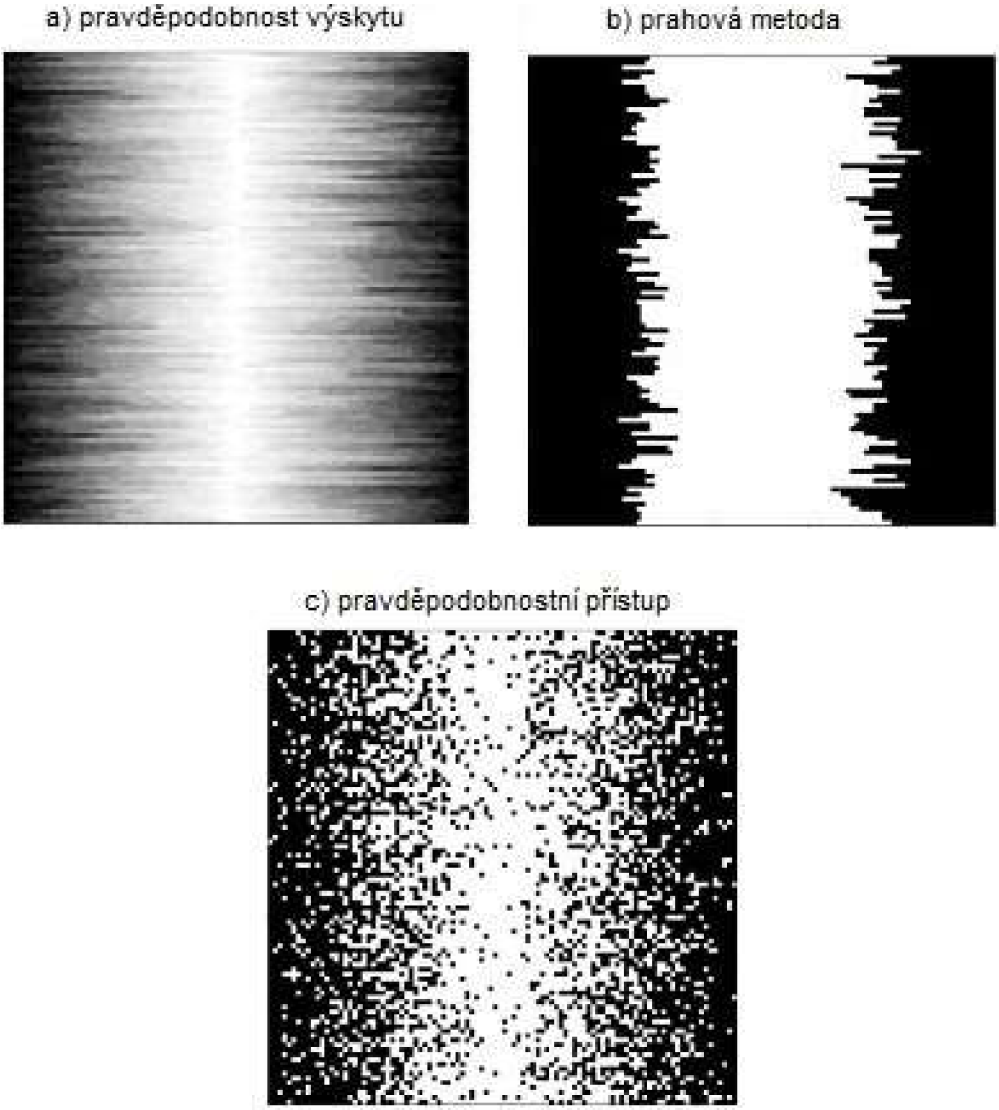
2) kombinace jednotlivých reakcí a generování pravděpodobnostního rasteru,

3) omezení šíření virtuálního druhu pouze na studovanou oblast,

4) převod pravděpodobnostního rasteru na prezence-absenční raster a samplování výskytových dat. (Leroy et al. 2015).

2.7.1 Rozdíl mezi prahovou metodou a pravděpodobnostním přístupem

Při generování virtuálního druhu je důležité, zda-li se při převodu vhodnosti habitatu na prezence-absenční rastr použije prahová nebo pravděpodobnostní metoda. V mnoha simulačních studiích se používá fixní prahová hodnota k převodu simulovaných pravděpodobností výskytu virtuálních druhů (Obr. 8.a) na prezenčně-absenční mapu (např. Hirzel et al. 2001 ; Real et al. 2006 ; Jiménez-Valverde & Lobo, 2007 ; Albert & Thuiller 2008 ; Jiménez-Valverde et al. 2009 ; Santika & Hutchinson, 2009 ; Peterson 2011 ; Bombi & D'Amen 2012). Tato metoda (Obr. 8.b) nicméně začíná být čím dál více kritizována, protože v reálném prostředí má prahovou reakci na změnu v prostředí jen minimum druhů (Elith & Graham 2009, Santika 2011, Meynard & Kaplan 2012, Moudrý 2015). Proto se častěji používá pravděpodobnostní přístup, což je náhodný proces spojený s pravděpodobnostmi výskytu, který může postupně reagovat na proměnné prostředí, a proto se hodnota může pohybovat v rozmezí od 0 do 1. Při použití tohoto přístupu bude pravděpodobnost výskytu 0,5 vést v průměru k obsazenosti 5 z každých 10 lokalit se stejnými podmínkami prostředí (např. Meynard & Quinn, 2007 ; Elith & Graham, 2009 ; Li et al. 2011 ; Santika, 2011). Jeho aplikací je prezence–absence nahodilým procesem převedena na pravděpodobnost výskytu, jež graduálně odpovídá na environmentální proměnné (Obr. 8.c) (Meynard & Kaplan 2012).



Obr. 8 Srovnání druhové distribuce aplikací prahové metody a pravděpodobnostního přístupu (Meynard & Kaplan 2012).

3. Metodika

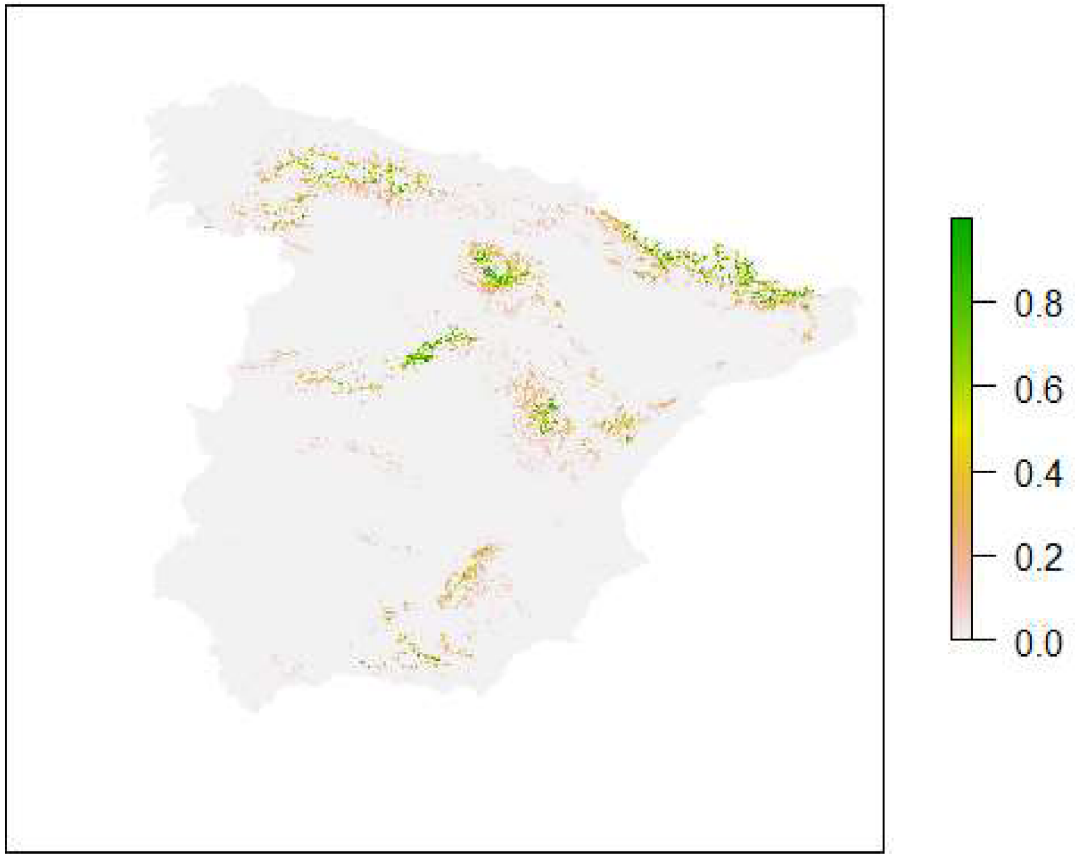
Z důvodu ověření předpokladu, že sampling bias ovlivňuje schopnost modelů správně detekovat tvar reakce druhů na prostředí, byl v rámci použité metodiky zvolen přístup virtuálních druhů (Zurell et al. 2010, Meynard et al. 2019).

3.1 Charakteristika zájmového území

Pro účely této bakalářské práce bylo jako území vybráno Španělské království s rozlohou 505 000 km^2 (4. největší země v Evropě) (Obr. 9). Na severu panuje oceánské podnebí s dostatkem srážek a s mírnou celoroční teplotou oproti jižní části Španělska, kde se území nachází ve středozemním podnebí s výskytem vyšších teplot a nedostatkem srážek. Od severu k jihu po východní straně zaujímá povrch Pyrenejské pohoří, které průměrně dosahuje výšek 2500 - 2900 m n. m.

Obr. 9 Studovaná oblast Španělsko.
(https://commons.wikimedia.org/wiki/File:Spain_in_Europe.svg)

3.2 Generování virtuálního druhu



Obr. 10 Vygenerovaný virtuální druh.

Virtuální druh byl pro území Španělska generován pomocí balíčku 'virtualspecies' (Obr. 10) (Leroy et al. 2018, ver. 1.5) ve statistickém softwaru R (ver. 4.1.0, R Development Core Team) ve třech krocích:

(i) Definování reakce virtuálního druhu na environmentální proměnné,

(ii) zkombinování jednotlivých reakcí a vygenerování pravděpodobnostního rasteru,

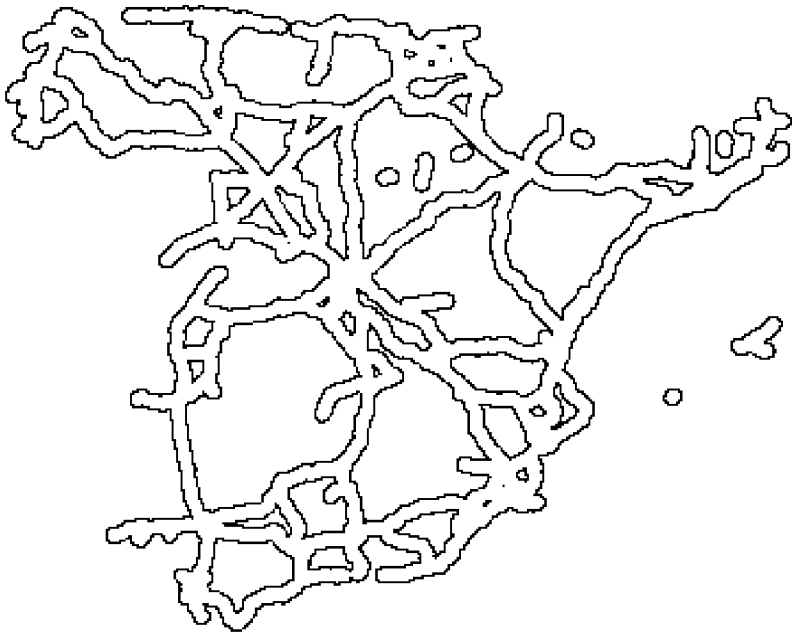
(iii) převedení pravděpodobnostního rasteru na prezence-absenční raster a vysamplování výskytových dat.

Pro vytvoření virtuálního druhu byl použit model nadmořské výšky a pokryvnost lesem (https://centrodedescargas.cnig.es/CentroDescargas/index.jsp). Reakce definovaná u nadmořské výšky (sd = 1600, mean = 300) u pokryvnosti lesem (sd = 20, mean = 80). Díky tomu byl vygenerován virtuální druh s nízkou prevalencí a úzkou nikou (který je náchylnější k sampling bias - viz Visscher 2006 nebo Gábor et al. 2020). Výskytová data byla vygenerována pomocí dvou způsobů:

(i) data byla náhodně generována pro celé území Španělska,

(ii) data byla generována jen poblíž hlavních komunikačních tahů v celém Španělsku.

Pro tento účel byl stažen shapefile vrsty (Obr. 11) silnic a kolem něho byl následně vytvořen buffer ve vzdálenosti 10 km. Pro každý scénář bylo vygenerováno 100 prezenčních záznamů. Díky tomu jsem mohl vygenerovat dataset se sampling bias.



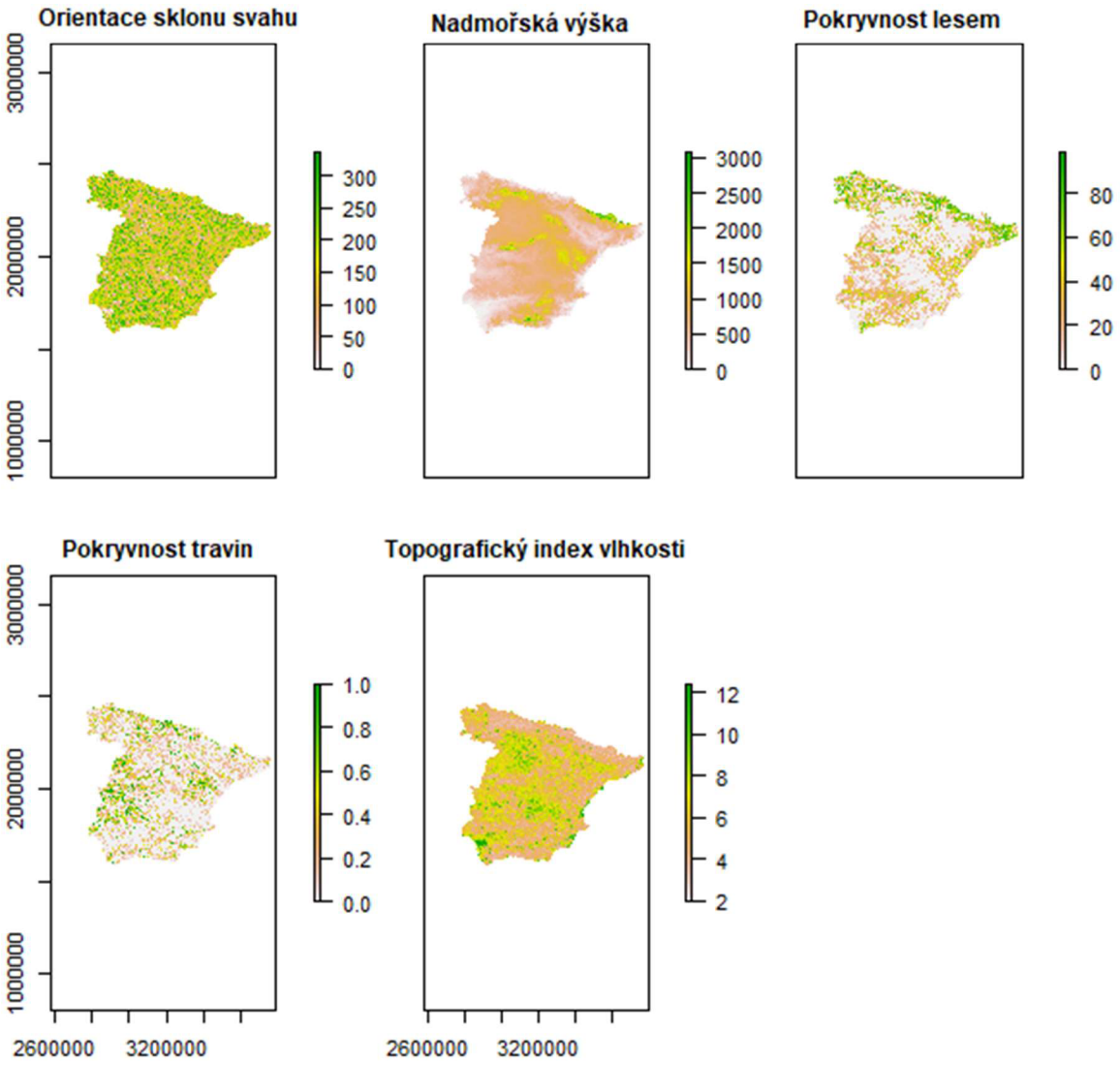
Obr. 11 Shapefile s vrstvou silnic s 10 km buffrem.

3.3 Environmentální proměnné

Naimi et al. (2011, 2014) ukázali, že prostorová autokorelace v environmentálních proměnných negativně ovlivňuje SDM. Proto byly vybrány proměnné prostředí, které zahrnovaly různé stupně prostorové autokorelace.

Pro vytvoření modelů distribuce bylo použito pět environmentálních proměnných (Obr. 13). Dvě z proměnných souvisely s charakteristikami stanovišť: pokrytí travních porostů a pokrytí lesy (http://centrodedescargas.cnig.es/; Španělské národní zeměpisné centrum); a tři se týkaly topografie: nadmořská výška (http://centrodedescargas.cnig.es/; španělský národní zeměpisný střed), topografický index vlhkosti a orientace sklonu svahu. Topografický index vlhkosti a orientace sklonu svahu byly odvozeny z výškového modelu (SAGA-GIS v. 2.1.4; Conrad et al. 2015). Všechny proměnné prostředí byly převzorkovány z původního rozlišení 10 x 10 m na rozlišení 50 x 50 m s použitím středních hodnot původních dat (Moudrý a kol. 2019) pro účely modelování. Zhoršení bylo nutné kvůli omezeným výpočetním možnostem mého počítače.

K identifikaci jakýchkoliv potenciálních problémů multikolinearity mezi proměnnými prostředí byla použita analýza rozptylových inflačních faktorů (VIF; balíček 'usdm', verze 1.1-18). Multikolinearita mezi prediktory může negativně ovlivnit SDM tím, že způsobí nestabilní odhady parametrů a zkreslené statistiky testů (Belsley 1991, Chatfield 1995, Dormann et al. 2013). Všechny hodnoty VIF naznačovaly nízkou multikolinearitu (<3), proto nebyly vyloučeny žádné proměnné (Zuur et al. 2010).



Obr. 12 Pět environmentálních proměnných použitých pro studovanou oblast.

3.4 Modely druhové distribuce

Modely druhové distribuce byly vytvořeny ve statistickém softwaru R (balíček 'sdm' ver. 1.0-98; Naimi & Araújo 2016) pomocí modelovací metody Maxent (Phillips et al. 2006), což je metoda často používaná v ekologických studiích (Linda a kol. 2016, Rodríguez et al. 2019, Santamarinaet et al. 2019, Ancillotto et al. 2020 El-Gabbas et al. 2020, Boral & Moktan 2021, Venne & Curie 2021). Modely byly použity s 10 000 background body a defaultním nastavením, jak je doporučeno (Phillips et al. 2006).

K vyhodnocení prediktivní výkonnosti modelu byly použity různé validační metriky. Byl použit Sørensenův index (SI), doporučený pro hodnocení SDM pomocí výskytů pouze v přítomnosti (Li & Guo 2013, Leroy et al. 2018). Kromě toho byla také vypočítána míra nadhodnocené predikce (OPR, Barbosa et al. 2013) a podhodnocené predikce (UPR, Fielding & Bell 1997), aby se zjistilo, zda poziční nejistota vedla ke konzistentnímu zkreslení nadhodnocené/podhodnocené predikce. Performační metriky byly získany pomocí 5-fold křízové validace.

Z důvodu zjištění účinku sampling bias na ekologickou interpretaci SDM byly porovnány informace u důležitosti jednotlivých proměnných, přičemž vygenerované křivky ukázaly reakci druhu na změnu prostředí (Elith et al. 2005, Murray & Conner 2009). U důležitosti jednotlivých proměnných byla také uložena informace o křivkách, které ukázaly reakci druhu na změnu v prostředí. Celý proces byl opakován celkem 50krát.

3.4.1 Použitý software

Při zpracování bakalářské práce byl použit software R, což je jazyk a prostředí pro statistické výpočty a grafiku. R poskytuje širokou škálu statistických (lineární a nelineární modelování, klasické statistické testy, analýza časových řad, klasifikace, shlukování, …) a grafických technik. Jednou ze silných stránek R je snadnost, s jakou lze vytvářet sofistikované grafy v kvalitě publikace, včetně matematických symbolů a vzorců tam, kde je to potřeba.

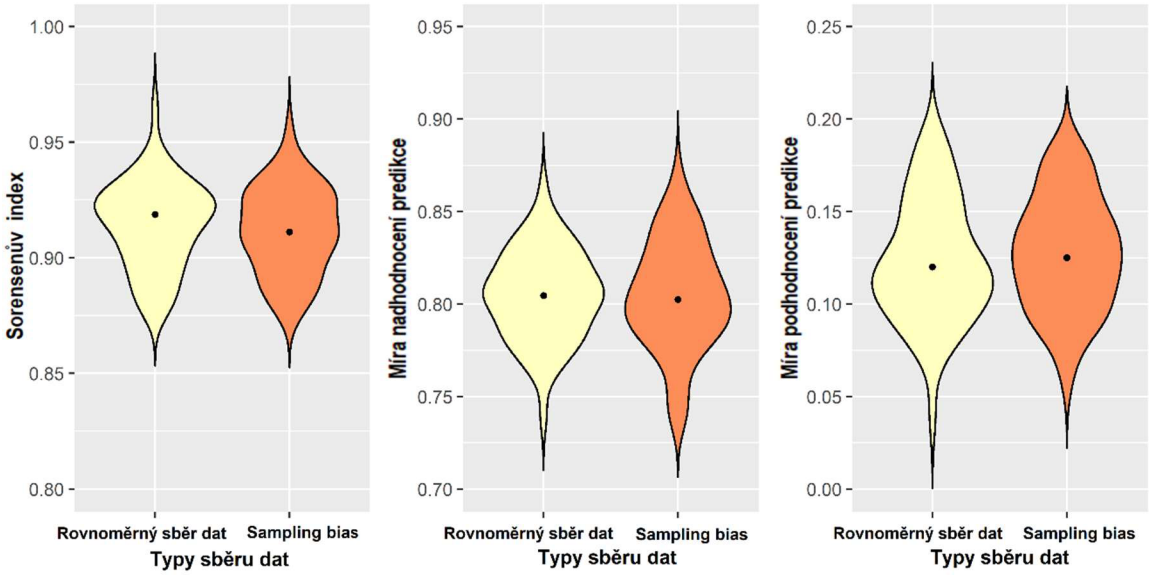
Velká péče v prostředí R byla věnována výchozím nastavením pro drobné volby designu v grafice, ale uživatel si zachovává plnou kontrolu. R je k dispozici jako free download software za podmínek GNU General Public License od Free Software Foundation ve formě zdrojového kódu. Kompiluje a pracuje na široké škále platforem UNIX a podobných systémech (včetně FreeBSD a Linuxu), Windows a MacOS.

Mnoho uživatelů považuje R za statistický systém, ve kterém jsou implementovány statistické techniky. S distribucí R je dodáváno asi osm balíčků a mnoho dalších je

dostupných prostřednictvím rodiny internetových stránek CRAN, které pokrývají velmi širokou škálu moderních statistik (https://www.r-project.org/about.html).

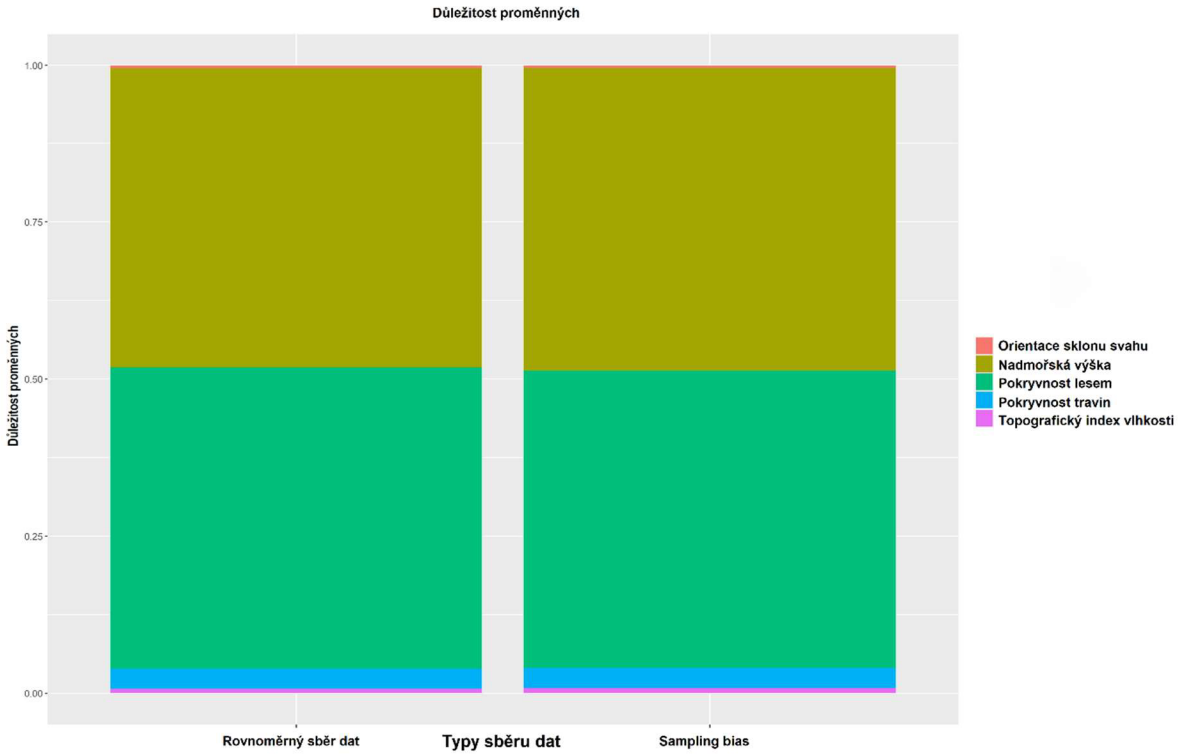
Použitý skript v mé bakalářské práci je uveden v příloze.

4. Výsledky



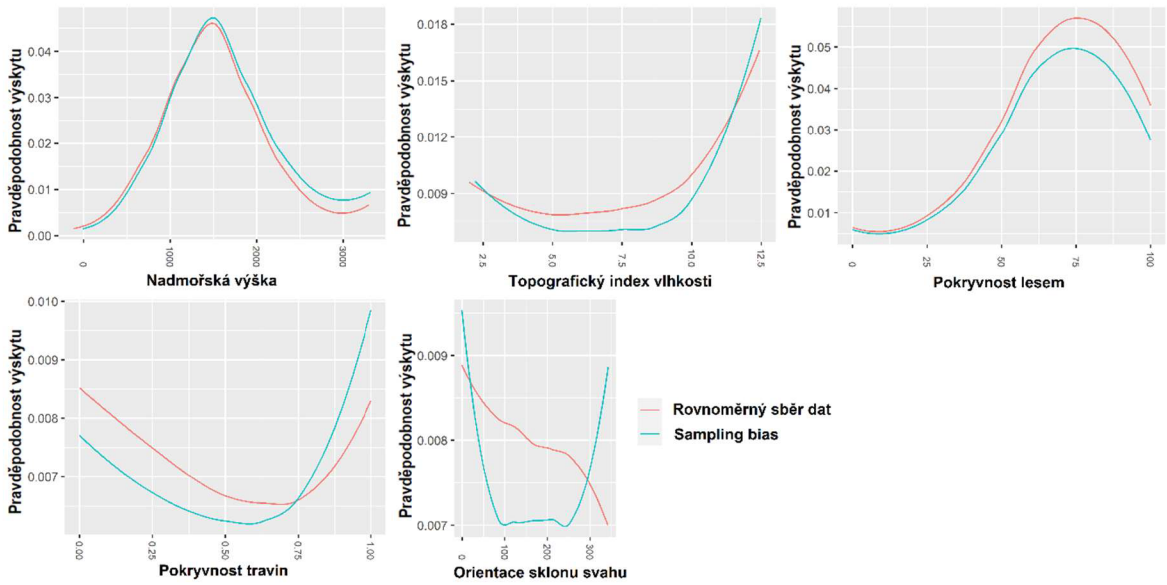
Obr. 13 Porovnání obou typů sběru dat v podobnostních metrikách SI, OPR a UPR.

Sorensenův index ukázal, že model performance je lepší pro modely, kde byla použita rovnoměrně sesbíraná data, pro která SI dosahoval hodnot 0,932, přičemž u sampling bias dosáhl hodnoty pouze 0,91, tedy rovnoměrně sesbíraná data měla u SI vyšší výkonnost nejen díky vyšší přesnosti o 0,021 (Obr. 14). Překvapivé bylo že Opr i Upr byly téměř stejné, ačkoliv random data vycházela vždy o něco lépe. Zároveň došlo k velmi podobnému vysokému nadhodnocení prediktivně vhodné plochy pro oba typy dat (OPR>0,8), ovšem pro rovnoměrně sesbíraná data vycházel lepší výsledek ve větší kumulaci bodů v jeho průměru oproti sampling bias (Obr. 14). Nakonec u sampling bias došlo k většímu podhodnocení 0,125 oproti rovnoměrně sesbíraným datům, kde byla hodnota 0,12, tedy o 0,005 menší než u sampling bias a zároveň měl větší kumulaci bodů v průměru (Obr. 14). Ve všech třech metrikách rovnoměrně sesbíraná data vycházela výsledkově lépe než data se sampling bias. Rozdíly byly minimální a data se sampling bias nijak zásadně nezkreslovala výsledky oproti rovnoměrně sesbíraným datům.



Obr. 14 Vliv sampling bias na environmentální proměnné.

Mezi nejvlivnějšími proměnnými prostředí pro oba typy dat byly nadmořská výška (pro rovnoměrně sesbíraná data ≐ 47,7% a pro sampling bias ≐ 48,2%) a pokryvnost lesu (pro rovnoměrně sesbíraná data ≐ 48,2% a pro sampling bias ≐ 47,2%). Další proměnnou byla pokryvnost travin, kde pro rovnoměrně sesbíraná data vyšla hodnota ≐ 3% a pro sampling bias byla hodnota o něco vyšší ≐ 3,4 %. Minimální vliv na distribuci měly proměnné - topografický index vlhkosti (pro rovnoměrně sesbíraná data ≐ 0,7% a pro sampling bias ≐ 0,8%) a orientace sklonu svahu (pro rovnoměrně sesbíraná data ≐ 0,4 % a pro sampling bias ≐ 0,4 %). Oba typy výskytových dat vedly k velmi podobným výsledkům ve významu proměnných (Obr. 15). Modely tedy ve všech případech správně odvodily nejvlivnější proměnné bez ohledu na sampling bias.



Obr. 15 Vliv sampling bias na pravděpodobnost výskytu v různých environmentálních proměnných.

K největšímu vychýlení křivek odezvy proběhlo u orientace sklonu svahu, kde data se sampling bias vysoce podhodnotila v daných stupních orientace sklonu svahu svoji předpověď oproti rovnoměrně sesbíraným datům. U zbývajících environmentálních proměnných vycházely křivky odezvy velice podobně. U nadmořské výšky měl sampling bias tendenci trochu nadhodnocovat svou předpověď. Topografický index vlhkosti naopak s daty se sampling bias podhodnocoval svou předpověď, než udávala rovnoměrně sesbíraná data. Stejně dopadl i pokryvnost lesa, kde data se sampling bias podhodnocovala svou předpověď. A nakonec pokryvnost travinami, kde sampling bias se snižujícím pokryvem podhodnocoval předpověď, a naopak se zvyšujícím pokryvem svou předpověď nadhodnocoval (Obr. 16). Sampling bias nejvíce negativně zasahoval při orientaci sklonu svahu. V ostatních environmentálních proměnných data se sampling bias měla mizivý vliv na výsledky.

5. Diskuze

Cílem mé bakalářské práce bylo ověřit, jak může sampling bias ovlivnit ekologickou interpretabilitu modelu druhové distribuce. Konkrétně jsem se v této práci zaměřil na vliv sampling bias na schopnost modelu detekovat důležitost environmentálních proměnných a schopnost modelu správně detekovat vliv měnícího se prostředí na druhovou distribuci při použití modelovací techniky Maxent.

Výsledky mé práce ukázaly, že sampling bias negativně ovlivňuje přesnost modelu druhové distribuce. To je v souladu s předchozími studiemi, které dospěly k závěru, že sampling bias snižuje kvalitu SDM. Přesnější údaje o výskytu obecně vedou k lepší výkonnosti SDM (Oria et al. 2014, Tessarolo et al. 2014, Gábor et al. 2020).

Na druhou stranu výsledky ukázaly, že i modely vytvořené s daty negativně ovlivněnými sampling bias mohou být stále ekologicky interpretovatelné. Ačkoliv se mi v porovnání s modely vytvořenými pomocí rovnoměrně generovaných dat nepatrně změnila důležitost environmentálních proměnných a tvar křivek ukazující reakci druhu na změny v prostředí, rozdíly byly pouze nepatrné. To naznačuje, že nízká přesnost modelu nemusí nutně vést k nízké schopnosti odvodit, které proměnné určují distribuci druhu a jak druh na tyto proměnné reaguje. Tyto výsledky jsou důležité pro budoucí studie, protože naznačují, že data se sampling bias není nutné filtrovat a tedy snižovat celkový počet záznamů, které se při modelování dají používat, což jak ukázal Smith et al. (2021) může negativně ovlivnit odhad toho, jakým způsobem se může změnit druhová biodiverzita vlivem klimatických změn.

Ačkoliv je moje práce optimistická k dalšímu využívání sampling bias dat v SDM, je potřeba si uvědomit, ze mé závěry musí být podrobeny dalšímu výzkumu. Budoucí studie by měly ověřit, jak se mohou mé závěry lišit při použití datových sad s menším nebo naopak větším počtem záznamů, při použití jiného rozlišení environmentálních proměnných nebo velikosti studované oblasti či použití jiné modelovací techniky.

6. Závěr

Výsledky této studie jsou zaměřeny na zkoumání vlivu sampling bias na ekologickou interpretabilitu modelů druhové distribuce. Virtuální druh byl vytvořen pomocí balíčku 'virtualspecies a pro vytvoření modelů druhové distribuce byl vybrán algoritmus Maxent. Validace modelů proběhla pomocí metrik Sorensenův index podobnosti, míry nadhodnocené predikce a míry podhodnocené predikce. Ve všech třech metrikách měla data se sampling bias negativní vliv na modely. Při hodnocení vlivu sampling bias na environmentální proměnné modely ve všech případech správně odvodily nejvlivnější proměnné bez ohledu na sampling bias. Při vlivu sampling bias na pravděpodobnost výskytu pro pět použitých environmentálních proměnných, sampling bias nejvíce negativně zasahoval při orientaci sklonu svahu, ačkoliv i přesto se křivka podobala modelu, který byl vytvořen pomocí náhodně vygenerované datové sady. Pro ostatní environmentální proměnné se data se sampling bias výrazněji neprojevila.

Závěrem lze říci, že i data se sampling bias mohou být stále užitečná při tvorbě modelů druhové distribuce, pokud je jejich cílem ekologická interpretace.

7. Seznam použité literatury

ALLOUCHE O., TSOAR A. & KADMON R. 2006: Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology, 43(6), 1223-1232.

Araújo, M., Luoto, M., 2007: "The importance of biotic interactions for modelling species distributions under climate change." Global Ecology and Biogeography, 16.6, 743-753.

Austin, M. P. 2002: Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. – Ecol. Model.157: 101–118.

Bakkenes, M., Alkemade, J. R. M., Leemans, F. I. R., Latour, J. B. 2002: Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050, Global Change Biology, 8(4), 390-407.

Barbosa, A. M., Real, R., Munoz, A. R., Brown, J. A. 2013: New measures for assessing model ekvilibrium and prediction mismatch in species distribution models, Biodiversity Letter, 19(10), 1333-1338.

Begon M., Harper J., Townsend C., 1992: Ecology: individuals, populations and communities, Univerzita Palackého, Olomouc.

BCCVL, ©2015: Maxent – algorithm for modelling species distribution (online) [cit. 2022.02.01], dostupné z https://support.bccvl.org.au/support/solutions/articles/6000083216-ma- xent

Bombi, P., D'Amen, M., 2012: Scaling down distribution maps from atlas data: a test of different approaches with virtual species. Journal of Biogeography 39(4), 640-651.

Boria, R. A., Olson, L. E., Goodman, S. M., Anderson, R. P. 2014: Spatial filtering to reduce sampling bias can improve the performance of ecological niche models, Ecological Modelling, 275, 73-77.

Busby, J., 1991: BIOCLIM – a bioclimate analysis and prediction system. Plant protection quarterly, 56-87.

Carpenter, G., Gillison, A. N., Winter, J. 1993: DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals, Biodiversity and Conservation, 2, 667-680.

Conrad, O., et al. 2015: System for Automated Geoscientific Analyses (SAGA) v. 2.1.4, Geosci. Model Dev., 8, 1991-2007.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitao P. J., Munkemuller, T., McClean, C., Osborne, P. E., Schroder, B., Skidmore, A. K., Zurell, D., Lautenbach, Sven. 2013: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Wildlife Biology, 36(1), 27-46.

Dvorský, M., Macek, M., Kopecký, M., Wild, J., & Doležal, J. 2017: Niche asymmetry of vascular plants increases with elevation. *Journal of Biogeography*, 44(6), 1418-1425.

Elith, J., Graham, C., Anderson, R., Dudík, M., Ferrier, S., Guisan, A., Li, J., 2006: Novel methods improve prediction of species' distributions from occurrence data. Ecography, 129-151.

Elith, J. & Graham, C. H. 2009: Do They? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models, Ecography, 32, 66-77.

Elith, J., Leathwick J. R. 2009: Species Distribution Models: Ecological Explanation and Prediction Across Space and Time, The Annual Review of Ecology, Evolution, and Systematics, 40,677-97.

Erickson, K. D. & Smith, A. B. 2021: Accounting for imperfekt detection in data from museums and herbaria when modelling species distributions: combining and contrasting data-level versus model-level bias correction, Ecography, 44(9).

Fielding, A. H. & Bell, J.F. 1997: A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models, Enironmental Conservation, 24, 38-49.

Gábor, L., Moudrý, V., Barták, V., & Lecours, V. 2020: How do species and data characteristics affect species distribution models and when to use environmental filtering?. International Journal of Geographical Information Science, 34(8), 1567-1584.

Gábor, L., Moudrý, V., Lecours, V., Malavasi, M., Barták, V., Fogl, M & Václavík, T. 2020: The effect of positional error on fine scale species distribution models increases for specialist species. Ecography, 43(2), 256-269.

Gaffney, P. M. 1975: Roots of the Niche Concept. The American Naturalist, 109(968),490.

Gaubert, P., Papes. M., Peterson, A. T. 2006: Natural history collections and the conservation of poorly known taxa: Ecological niche modelling in central African rainforest genets, Biological Conservation, 130(1), 106-117.

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., Tottrup, A. P. 2016: What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements, Diversity and Distributions, 22(11), 1139-1149.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A., Loiselle, B. A. and 2008: The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45: 239-247.

Grinnell, J. 1917: The Niche-Relationships of the California Thrasher. The Auk, 34(4), 427–433.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., Mccarthy, M. A., Tingley, R., Wintle, B. A. 2015: Is my species distribution model fit for purpose? Matching data and models to applications, Global Ecology and Biography, 24(3), 276-292.

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G., Korber, J. H. 2007: Biotic interactions improve prediction of boreal bird distribution at macro-scales, Global Ecology and Biogeography, 16(6), 754-763.

Hirzel, A. H., Helfer, V., Metral, F., 2001: Assessing habitat-suitability models with a virtual species, Ecological Modelling, 145(2-3), 111-121.

Hirzel, A. H., Hausser, J., Perrin, N., Chessel, D. 2002: Ecological-Niche Factor Analysis: How to Compute Habitat-Suitability Maps without Absence Data, Ecology, 83, 2027-2036.

Hirzel, A. H., Le Lay, G. 2008: Habitat suitability modelling and niche theory, Journal of Applied Ecology, 45(5), 1372-1381.

Jarošík V., 1987: Ekologie, Učební text, Karlova univerzita, Přírodovědecká fakulta, 54 s.

Jiménez-Valverde, A. 2012: Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling, Global Ecology and Biogeography, 21(4), 498-507.

Kienast, F., Degenhardt, B., Weilenmann, B., Wäger, Y., Buchecker, M., 2012: GIS-assisted mapping of landscape suitability for nearby recreation, Landscape and Urban planning, 105(4), 385-399.

Leitão, P. J., Moreira, F., & Osborne, P. E. 2011: Effects of geographical data sampling bias on habitat models of species distributions: a case study with steppe birds in southern Portugal. International Journal of Geographical Information Science, 25(3), 439-454.

Leroy, B., Meynard, Ch. N., Bellard, C. & Courchamp, F. 2015: Virtual species, an R package to generate virtual species distributions, Ecography 38, 001-009.

Leroy B, Delsol R, Hugueny B, et al. 2018: Without quality presence–absence data, discriminationmetrics such as TSS can be misleading measures of model performance, Journal Biogeography,45:1994–2002.

Lobo, J. M., Jiménez-Valverde, A., Real, R. 2008: AUC: a misleading measure of the performance of predictive distribution models, Global Ecology and Biogeography, 17(2), 145-151.

Lobo, J. M., Jiménez-Valverde, A., Hortal, J. 2010: The uncertain nature of absences and their importance in species distribution modelling, Ecography, 33(1), 103-114.

McPherson, J. M., Jetz, W., Rogers, D. J. 2004: The effects of species range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact, Journal of Applied Ecology, 41(5), 811-823.

Macdonald, H. 2004: Geologic Puzzles: Morrison Formation, ***Starting Point***, z http://serc.carleton.edu/introgeo/interactive/examples/morrisonpuzzle.html

Meynard, Ch. N. & Kaplan, M. D. 2012: Using virtual species to study species distributions and model performance, Journal of Biogeography, 40(1), 1-8.

Meynard, Ch. N. & Kaplan, M. D. 2012: The effect of a gradual response to the environment on species distribution modelling performance. Ecography, 35(6), 499-509.

Meynard, Ch. N., Leroy, B. & Kaplan, D. M. 2019: Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing?, Ecography, 42, 2021-2036.

Miller, J. 2010: Species Distribution Modelling. Geography Compass, 4(6), 490-509.

Moudrý V., 2015: Modelling species distributions with simulated virtuál species. Journal of Biogeography 42(8), 1365-1366.

Moudrý, V., Lecours, V., Gdulová, K., Gábor, L., Moudrá, L., Kropáček, J., Wild, J., 2018: On the use of global DEMs in ecological modelling and the accuracy of new bare-earth DEMs, Ecological Modelling, 383, 3-9.

Moudrý, V., & Devillers, R. 2020: Quality and usability challenges of global marine biodiversity databases: An example for marine mammal data. Ecological Informatics, 56, 101051.

Murray, K., & Conner, M. M. 2009: Methods to quantify variable importance: implications for the analysis of noisy ecological data. Ecology, 90(2), 348-355.

Naimi, B., Skidmore, A. K., Groen, T. A., & Hamm, N. A. 2011: Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. Journal of Biogeography, 38(8), 1497-1509

Naimi, B., Araújo, M. B., 2016: SDM: a reproducible and extensible R platform for species distribution modelling, Wildlife Biology, 39(4), 368-375.

NASA, ©2020: Species Distribution Modelling Data (online) [cit. 2022.01.18], dostupné z https://earthdata.nasa.gov/learn/pathfinders/biodiversity/species-distribution.

Peirce, C. S. 1884: The numerical measure of the success of prediction,Science,4(93), 453–454.

Peterson, A., 2003: Predicting the geography of species' invasions via ecological niche modeling. The quarterly review of biology 78(4), 419-433.

Peterson, A. T. 2008: Phylogeography is not enough: The need for multiple lines of evidence. Frontier of Biogeography, 1(1).

Phillips, S., Anderson, R., Schapire, R., 2006: Maximum entropy modeling of species geographic distributions. Ecological modelling 190.3-4, 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, Ecological Applications, 39(1), 181-197.

Pulliam, H., 2000: On the relationship between niche and distribution. Ecology letters 3.4, 349-361.

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, Maiorano L. 2016: Performance tradeoffs in target-group bias correction for species distribution models, Ecography, 40(9), 1076-1087.

Rissler, L. J., Apodaca, J. J. 2007: Adding More Ecology into Species Delimitation: Ecological Niche Models and Phylogeography Help Define Cryptic Species in the Black Salamander, Systematic Biology, 56(6), 924-942.

Santika, T. 2011: Assesing the effect of prevalence on the predictive performance of species distribution models using simulated data, Global Ecology and Biogeography, 20(1), 181-192.

Scott, J. M., Heglund, P. J., Morrison, M. L., Wall, W. A. 2002: Predicting Species Occurences: Issues of Accuracy and Scale, ISBN: 1-55963-787-0.

Smith, T. B., Mila, B., Girman, J., Kimura, J. 2000: Genetic evidence for the effect of a postglacial population expansion on the phylogeography of a North American songbird, Royal Society, 267(1447).

Smith, A. B. & Santos, M. J. 2020: Testing the ability of species distribution models to infer variable importance, Ecography, 43(12), 1801-1813.

Smith, A. B., Murphy, S. J., Henderson, D., & Erickson, K. D. 2021: Imprecisely georeferenced specimen data provide unique information on species' distributions and environmental tolerances: Don't let the perfect be the enemy of the good. *bioRxiv*.

Sillero, N., 2011: What does ecological modelling model? A proposed classification of ecoloical niche models based on their underlying methods, Ecological Modelling, 222(8), 1343-1346.

Soberón, J., Peterson, A. T., 2005: Interpretation of models of fundamental ecological niches and species distributional areas, Biodiversity Informatics, 2, 1-10.

Soberón, J., Nakamura, M. 2009: Niches and distribution areas: Concepts, methods and assumptions, PNAS, 106(2).

Soberón, J. M. 2010, Niche and area of distribution modeling: a population ecology perspective. Ecography, 33: 159-167.

Studijní texty předmětu Z0025 Ekologie a životní prostředí, Elektronická učebnice, Masarykova univerzita, 2013, 124 s.

Stockwell, D., Peters, D. G. 1999: The GARP modelling system: Problems and solutions to automated spatial prediction, International Journal of Geographical Information Science, 13, 143-158.

Suarez, A., Tsutsui, N. D. 2004: The Value of Museum Collections for Research and Society, Bioscience, 54(1),66-74.

Syfert, M. M., Smith, M. J., Coomes, D.A. 2013: The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models, PLOS ONE, 8(7).

The R Foundation ©2022: What is R? (online) [cit. 2022.01.20], dostupné z https://www.r-project.org/about.html

Varela, S., Anderson, R. P., Valdés, R. G., Fernández-González, F. 2014: Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models, Wildlife Biology, 37(11), 1084-1091.

Venne, S., & Currie, D. J. 2021: Can habitat suitability estimated from MaxEnt predict colonizations and extinctions?. *Diversity and Distributions*, *27*(5), 873-886.

Wikipedie ©2022: Španělsko (online) [cit. 2022.02.20], dostupné z https://cs.wikipedia.org/wiki/%C5%A0pan%C4%9Blsko

WorldClim, ©2020: Bioclimatic variables (online) [cit. 2022.01.10], dostupné z https://www.worldclim.org/data/bioclim.html

Zimmermann, N. E., Edwards, T. C., Jr, Graham, C. H., Pearman, P. B. and Svenning, J. C. 2010: New trends in species distribution modelling. Ecography, 33: 985-989.

Zurell, D. Berger, U. Cabral, J. S. 2010: The virtual ecologist approach: simulating data and observers. Oikos, 119(4), 622-635.

Zuur, A. F., Ieno, E. N., & Elphick, C. S. 2010: A protocol for data exploration to avoid common statistical problems. Methods in ecology and evolution, 1(1), 3-14.

Seznamy

7.1 Obrázky

7.2 Tabulky

8. Příloha

Krok 1: Instalace a nahrání packages

```
library(dismo)
library(rJava)
library(glm2)
library(gam)
library(raster)
library(virtualspecies)
library(ggplot2)
library(sdm)
library(maptools)
library(dplyr)
library(tidyverse)
library(sf)
library(cowplot)
```

Krok 2: Nahrání funkcí

```
sdm.package.evaluation <- function (fit.model){
 th <- mean(getEvaluation(fit.model, stat= "threshold", opt = 2)[,2])

 cm1 <- as.table(sdm:::.cmx(o = as.vector(fit.model@models$occ$maxent$`1`@evaluation$test.dep@observed),
                  p = as.vector(ifelse(fit.model@models$occ$maxent$`1`@evaluation$test.dep@predicted[] >= th, 1, 0))))

 cm2 <- as.table(sdm:::.cmx(o = as.vector(fit.model@models$occ$maxent$`2`@evaluation$test.dep@observed),
                  p = as.vector(ifelse(fit.model@models$occ$maxent$`2`@evaluation$test.dep@predicted[] >= th, 1, 0))))

 cm3 <- as.table(sdm:::.cmx(o = as.vector(fit.model@models$occ$maxent$`3`@evaluation$test.dep@observed),
                  p = as.vector(ifelse(fit.model@models$occ$maxent$`3`@evaluation$test.dep@predicted[] >= th, 1, 0))))

 cm4 <- as.table(sdm:::.cmx(o = as.vector(fit.model@models$occ$maxent$`4`@evaluation$test.dep@observed),
                  p = as.vector(ifelse(fit.model@models$occ$maxent$`4`@evaluation$test.dep@predicted[] >= th, 1, 0))))

 cm5 <- as.table(sdm:::.cmx(o = as.vector(fit.model@models$occ$maxent$`5`@evaluation$test.dep@observed),
                  p = as.vector(ifelse(fit.model@models$occ$maxent$`5`@evaluation$test.dep@predicted[] >= th, 1, 0))))

 eval <- getEvaluation(fit.model, stat= c("AUC", "Kappa"))

 perf <- rbind(data.frame(performance(cm1)), (performance(cm2)), data.frame(performance(cm3)),
          data.frame(performance(cm4)),data.frame(performance(cm5)))

 data.frame(TPR = mean(perf$TPR), TNR = mean(perf$TNR), FPR = mean(perf$FPR), FNR = mean(perf$FNR), Sensitivity = mean(perf$Sensitivity),
        Specificity = mean(perf$Specificity), AUC = mean(eval$AUC), Kappa = mean(eval$Kappa),TSS = mean(perf$TSS),
        Jaccard = mean(perf$Jaccard), Sorensen = mean(perf$Sorensen), F_measure = mean(perf$F_measure), OPR = mean(perf$OPR),
        UPR = mean(perf$UPR))

}
```

```
performance <- function (confusion.matrix) {
 tp <- confusion.matrix[1,1]
 fp <- confusion.matrix[1,2]
 fn <- confusion.matrix[2,1]
 tn <- confusion.matrix[2,2]
 TPR <- tp / (tp+fn)
 TNR <- tn / (tn+fp)
 FPR <- fp / (fp+tn)
 FNR <- fn / (fn+tp)
 Sensitivity <- TPR
 Specificity <- TNR
 TSS = Sensitivity + Specificity - 1
 Jaccard = TPR/(FNR + TPR + FPR)
 Sorensen = 2*TPR/(FNR + 2*TPR + FPR)
 F_measure= 2 * Jaccard
 OPR = fp/(tp+fp)
 UPR = 1 - Sensitivity
 data.frame(TPR = TPR, TNR = TNR, FPR = FPR, FNR = FNR, Sensitivity = Sensitivity, Specificity = Specificity,
         TSS = TSS, Jaccard = Jaccard, Sorensen = Sorensen, F_measure = F_measure, OPR = OPR, UPR = UPR)
}

extract_curves <- function(rc, nfolds = 5){
 rc %>%
  map(~ mutate(.x,
           response = rowMeans(.[,2:(nfolds + 1)]),
           variable = names(.x)[1]) %>%
       rename(value = 1) %>%
       select(-c(2:(nfolds + 1)))) %>%
  bind_rows()
}
```

Krok 3: Nahrání dat (environmentální proměnné, shp. silnic a virtuální druh)

```
Bioall <- stack("aspect.tif", "dem.tif", "forest.tif", "grassland.tif", "twi.tif")

sfdata <- readShapeSpatial("data/silnice2.shp", proj4string=CRS("+proj=longlat"))

Bioall.virtualSpecies <- stack("forest.tif", "dem.tif")
```

Krok 4: Korelace mezi environmentálními proměnnými

```
Bioall.metrix <- na.omit(as.matrix(Bioall))
cor <- cor(Bioall.metrix);cor
```

Krok 5: Tvorba virtuálního druhu s rozdílnou šířkou niky

```
my.parametrs <- formatFunctions(forest = c(fun = 'dnorm', mean = 80, sd = 20),
                        dem = c(fun = 'dnorm', mean = 1600, sd = 300))

virtual.species <- generateSpFromFun(raster.stack = Bioall.virtualSpecies,
                             parameters = my.parametrs,
                             species.type = "multiplicative",
                             plot = T)
```

Step 6: Převedení rastru na prezence-absenční rastr

```
PA.raster <- convertToPA(virtual.species, alpha = -0.05, beta = 0.3, plot = T)

perf <- numeric()

var.importance <- numeric()
response.curves <- numeric()
var.importance <- numeric()
```

Krok 7: Samplování obou typů dat (random sampling a sampling bias)

```
randomSampling <- sampleOccurrences(PA.raster, n = 100, type = "presence only", plot = T)

samplingBias <- sampleOccurrences(PA.raster, n = 100, type = "presence only", bias = "polygon",
                          bias.strength = 6, bias.area = sfdata, plot = T)
```

Krok 8: Příprava dat pro SDM

```
DATA.Prep <- cbind(randomSampling$sample.points, samplingBias$sample.points[,1:2])

 sp0 <- na.omit(data.frame(occ=DATA.Prep$Observed, DATA.Prep[,1:2]))
 coordinates(sp0) <- ~x+y
 sp1 <- na.omit(data.frame(occ=DATA.Prep$Observed, x = DATA.Prep[,5], y = DATA.Prep[,6]))
 coordinates(sp1) <- ~x+y
```

Krok 9: Tvorba modelu Maxent

```
d0 <- sdmData(train = sp0, predictors = Bioall, bg=list(n=10000,method='gRandom',remove=TRUE))
d1 <- sdmData(train = sp1, predictors = Bioall, bg=list(n=10000,method='gRandom',remove=TRUE))


m0 <- sdm (occ~., data = d0, methods = "maxent", replication =  "cv", cv.folds=5)
m1 <- sdm (occ~., data = d1, methods = "maxent", replication =  "cv", cv.folds=5)

perf <- rbind(perf,
            data.frame(sdm.package.evaluation(m0), data = "RandomSampling"),
            data.frame(sdm.package.evaluation(m1), data = "SamplingBias"))


 response.curves <- rbind(response.curves,
                   data.frame(extract_curves(getResponseCurve(m0)@response, nfolds = 5), data = "RandomSampling"),
                   data.frame(extract_curves(getResponseCurve(m1)@response, nfolds = 5), data = "SamplingBias"))


 var.importance <- rbind(var.importance,
                  data.frame(getVarImp(m0)@varImportanceMean$corTest[,1:2], getVarImp(m0)@varImportanceMean$AUCtest[,1:2], data = "RandomSampling"),
                  data.frame(getVarImp(m1)@varImportanceMean$corTest[,1:2], getVarImp(m1)@varImportanceMean$AUCtest[,1:2], data = "SamplingBias"))
```

Krok 10: Grafické výstupy

```
var.importance2 <- var.importance %>%
 group_by(variables, data) %>%
 summarize(mean_cor = mean(corTest))

ggplot(var.importance2, aes(x = data, y = mean_cor, fill = variables)) +
 geom_bar(stat = "identity", position = "fill") +
 theme(
 plot.title = element_text(hjust = 0.5, size=11, face="bold"),
 axis.title.y=element_text(size=10, face="bold"),
 axis.title = element_text(size=9),
 axis.title.x=element_text(size=10, face="bold"),
 legend.title = element_blank()) +
 labs(title="Variable Importance",
     x="Different type of data sampling",
     y="Variable importance")

ggsave("var_importance.png", dpi=300, height = 25, width = 35, units = "cm")
```

Model performance

```
sorensen <- ggplot(perf, aes(as.factor(data), Sorensen, fill = data)) +
  geom_violin(trim=FALSE, color="black") +
  scale_fill_brewer(palette="RdYlGn", direction=-1) +
  stat_summary(fun=median, geom="point", size=1, color = "black") +
  scale_y_continuous(limits=c(0.8, 1), breaks = c(0.75, 0.8, .85, .9, .95, 1)) +
  theme_grey() +
  theme(
    plot.title = element_text(hjust = 0.5, size=11, face="bold"),
    axis.title.y=element_text(size=10, face="bold"),
    axis.title = element_text(size=9),
    axis.title.x=element_text(size=10, face="bold"),
    legend.position = "",
    legend.title = element_blank()) +
  labs(y="Sorensen index", x="Different type of data sampling")
opr <- ggplot(perf, aes(as.factor(data), OPR, fill = data)) +
  geom_violin(trim=FALSE, color="black") +
  scale_fill_brewer(palette="RdYlGn", direction=-1) +
  stat_summary(fun=median, geom="point", size=1, color = "black") +
  scale_y_continuous(limits=c(0.8, 1), breaks = c(0.75, 0.8, .85, .9, .95, 1)) +
  theme_grey() +
  theme(
    plot.title = element_text(hjust = 0.5, size=11, face="bold"),
    axis.title.y=element_text(size=10, face="bold"),
    axis.title = element_text(size=9),
    axis.title.x=element_text(size=10, face="bold"),
    legend.position = "",
    legend.title = element_blank()) +
  labs(y="Over prediction rate", x="Different type of data sampling")
upr <- ggplot(perf, aes(as.factor(data), UPR, fill = data)) +
  geom_violin(trim=FALSE, color="black") +
  scale_fill_brewer(palette="RdYlGn", direction=-1) +
  stat_summary(fun=median, geom="point", size=1, color = "black") +
  scale_y_continuous(limits=c(0.8, 1), breaks = c(0.75, 0.8, .85, .9, .95, 1)) +
  theme_grey() +
  theme(
    plot.title = element_text(hjust = 0.5, size=11, face="bold"),
    axis.title.y=element_text(size=10, face="bold"),
    axis.title = element_text(size=9),
    axis.title.x=element_text(size=10, face="bold"),
    legend.position = "",
    legend.title = element_blank()) +
  labs(y="Under prediction rate", x="Different type of data sampling")

plot.all <- plot_grid(sorensen, opr, upr, nrow = 1, align = "hv")
```

Response curves

```
aspect <- ggplot(response.curves[ which(response.curves$variable=='aspect'), ], aes(x=value, y=response, color = data)) +
 geom_smooth(method="loess", se=F, size = 0.5) +
 theme_grey() +
 theme(
  plot.title = element_text(hjust = 0.5, size=11, face="bold"),
  axis.title.y=element_text(size=10, face="bold"),
  axis.title = element_text(size=9),
  axis.title.x=element_text(size=10, face="bold"),
  axis.text.x = element_text(angle = 270, hjust=1, vjust = 0.5, size = 7),
  legend.title = element_blank()) +
 labs(x="Aspect",
     y="Occurrence probability",
     fill="Data Accuracy")
forest <- ggplot(response.curves[ which(response.curves$variable=='forest'), ], aes(x=value, y=response, color = data)) +
 geom_smooth(method="loess", se=F, size = 0.5) +
 theme_grey() +
 theme(
  plot.title = element_text(hjust = 0.5, size=11, face="bold"),
  axis.title.y=element_text(size=10, face="bold"),
  axis.title = element_text(size=9),
  axis.title.x=element_text(size=10, face="bold"),
  axis.text.x = element_text(angle = 270, hjust=1, vjust = 0.5, size = 7),
  legend.title = element_blank(),
  legend.position = "none") +
 labs(x="Amount of Forest",
     y="Occurrence probability",
     fill="Data Accuracy")
grassland <- ggplot(response.curves[ which(response.curves$variable=='grassland'), ], aes(x=value, y=response, color = data)) +
 geom_smooth(method="loess", se=F, size = 0.5) +
 theme_grey() +
 theme(
  plot.title = element_text(hjust = 0.5, size=11, face="bold"),
  axis.title.y=element_text(size=10, face="bold"),
  axis.title = element_text(size=9),
  axis.title.x=element_text(size=10, face="bold"),
  axis.text.x = element_text(angle = 270, hjust=1, vjust = 0.5, size = 7),
  legend.title = element_blank(),
  legend.position = "none") +
 labs(x="Grassland",
     y="Occurrence probability",
     fill="Data Accuracy")
dem <- ggplot(response.curves[ which(response.curves$variable=='dem'), ], aes(x=value, y=response, color = data)) +
 geom_smooth(method="loess", se=F, size = 0.5) +
 theme_grey() +
 theme(
  plot.title = element_text(hjust = 0.5, size=11, face="bold"),
  axis.title.y=element_text(size=10, face="bold"),
  axis.title = element_text(size=9),
  axis.title.x=element_text(size=10, face="bold"),
  axis.text.x = element_text(angle = 270, hjust=1, vjust = 0.5, size = 7),
  legend.title = element_blank(),
  legend.position = "none") +
 labs(x="Elevation",
     y="Occurrence probability",
     fill="Data Accuracy")
twi <- ggplot(response.curves[ which(response.curves$variable=='twi'), ], aes(x=value, y=response, color = data)) +
 geom_smooth(method="loess", se=F, size = 0.5) +
 theme_grey() +
 theme(
  plot.title = element_text(hjust = 0.5, size=11, face="bold"),
  axis.title.y=element_text(size=10, face="bold"),
  axis.title = element_text(size=9),
  axis.title.x=element_text(size=10, face="bold"),
  axis.text.x = element_text(angle = 270, hjust=1, vjust = 0.5, size = 7),
  legend.title = element_blank(),
  legend.position = "none") +
 labs(x="TWI",
     y="Occurrence probability",
     fill="Data Accuracy")

plot.response <- plot_grid(dem, twi, forest, grassland, aspect, nrow = 2)
```

Variable importance

```
var.importance2 <- var.importance %>%
  group_by(variables, data) %>%
  summarize(mean_cor = mean(corTest))

ggplot(var.importance2, aes(x = data, y = mean_cor, fill = variables)) +
  geom_bar(stat = "identity", position = "fill") +
  theme(
  plot.title = element_text(hjust = 0.5, size=11, face="bold"),
  axis.title.y=element_text(size=10, face="bold"),
  axis.title = element_text(size=9),
  axis.title.x=element_text(size=10, face="bold"),
  legend.title = element_blank()) +
  labs(title="Variable Importance",
       x="Different type of data sampling",
       y="Variable importance")
```