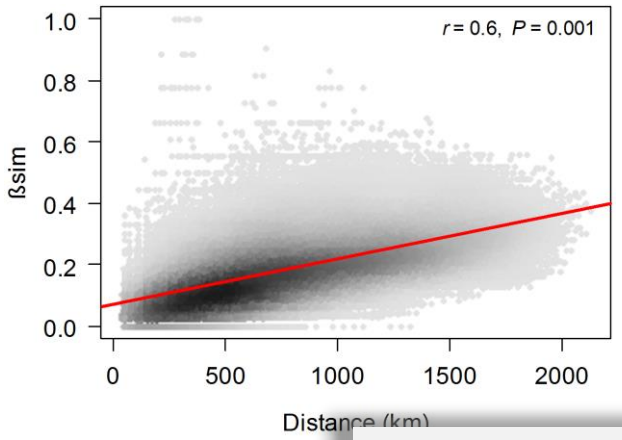
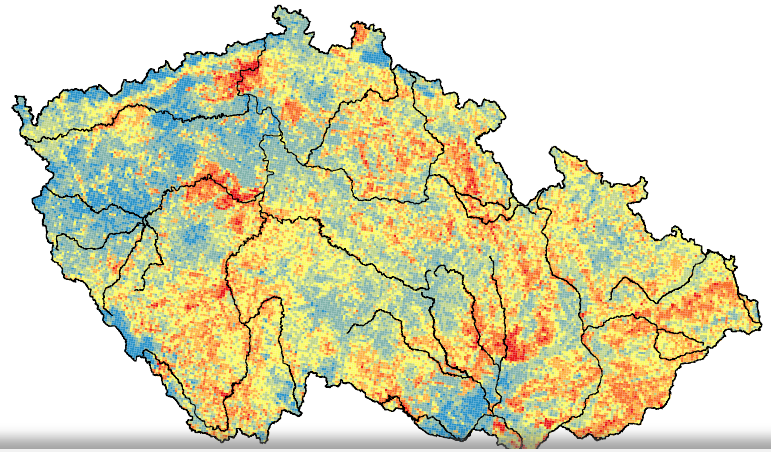
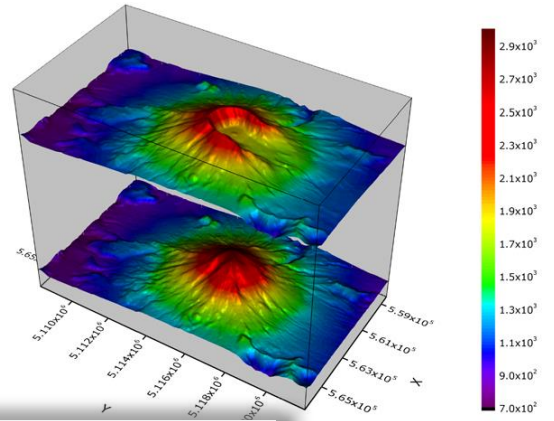
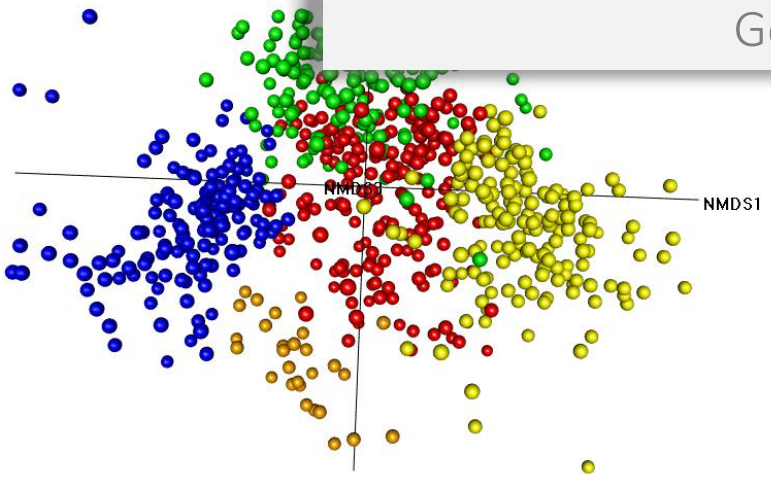


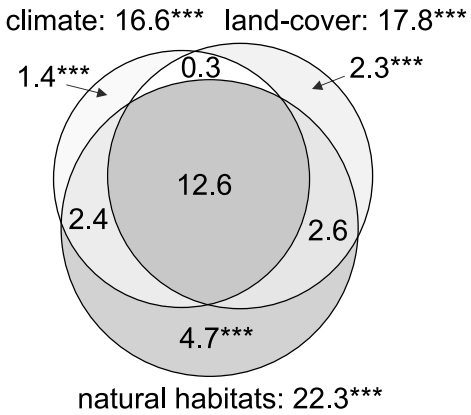




Metody fyzické geografie 3:
Biogeografie & ekologie

Jan Divíšek
Geografický ústav & Ústav botaniky a zoologie





Metody fyzické geografie 3 – 22. 5. 2017

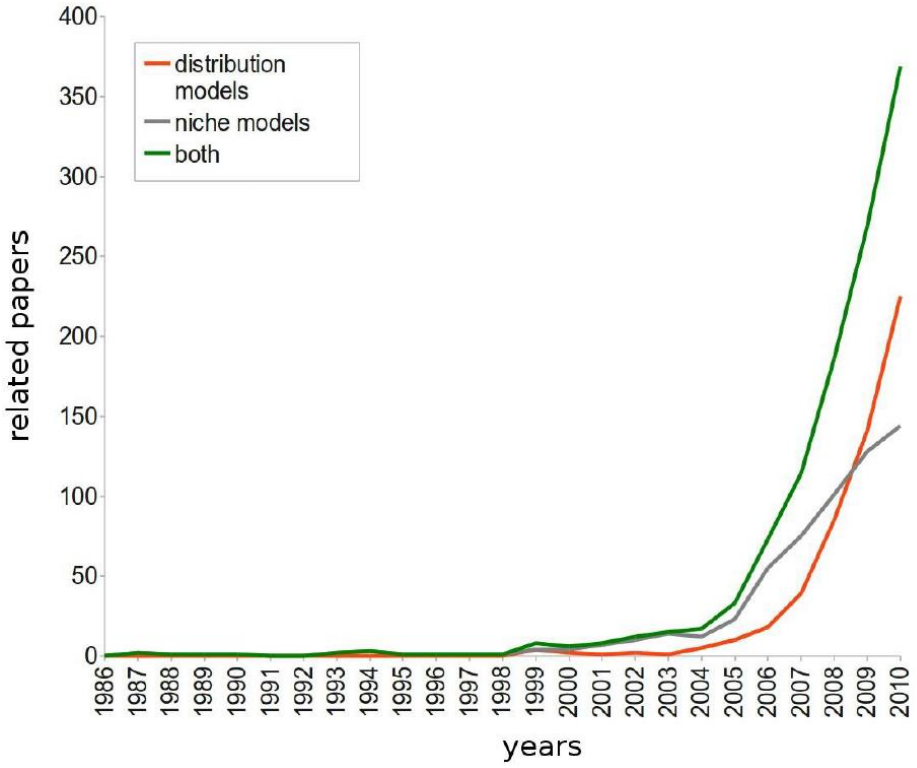
- Teoretická část
  - Modelování rozšíření druhů
  - Machine-learning methods
    - CART
    - Random Forest
    - MaxEnt
  - Prostorová autokorelace

- Praktická část
  - zítra

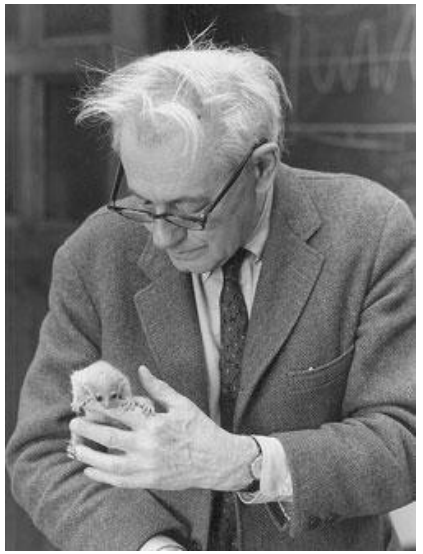
Modelování v biogeografii a ekologii

- Postup využívající numerické metody k předpovězení geografického rozšíření druhu na základě pozorovaného vztahu k environmentálním faktorům
- Geografická reprezentace rozšíření (1/0) nebo pravděpodobnosti výskytu duhu

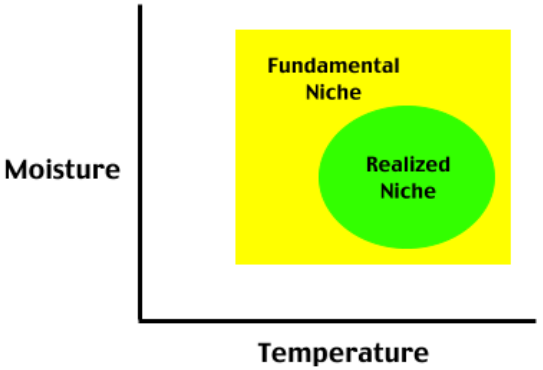
- Alternativní názvy
  - *Environmental niche modelling (ENM)*
  - *Species distribution modelling (SDM)*
  - *Climate envelope modelling*
  - *Habitat suitability modelling*



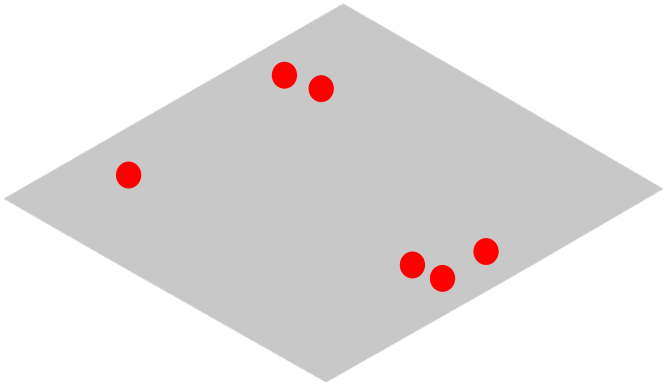
Teoretické základy

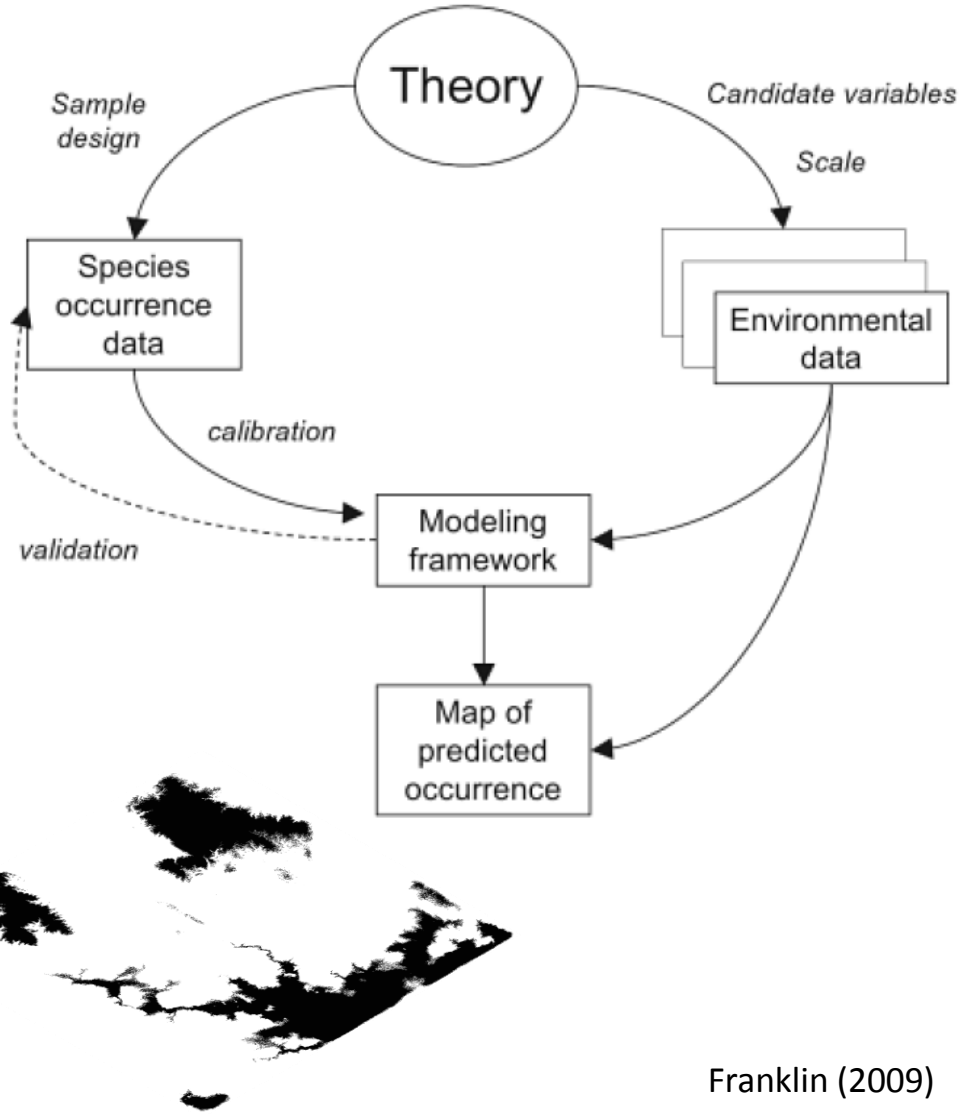


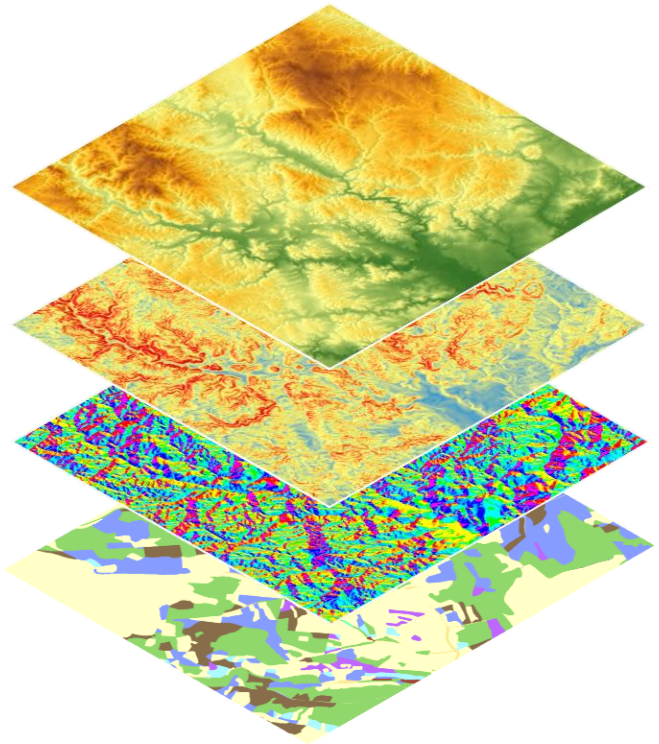
- Modelování rozšíření druhů staví na Hutchinsonově (1957) teorii fundamentální a realizované niky
- Neexistuje všeobecný konsenzus o tom, co modely vlastně vyjadřují
  - Většina modelů nezahrnuje druhové interakce a další faktory omezující rozšíření druhu → ***fundamentální nika*** (např. Soberón & Peterson, 2005)
  - Modely jsou založeny na pozorovaných prezencích (a případně absencích) druhu → ***realizovaná nika*** (např. Guisan & Zimmermann, 2000)
- Mezidruhové interakce
  - Pozitivní – součást fundamentální niky (rozšiřují prostor existence organismu)
  - Negativní – součást realizované niky (limitují prostor existence organismu)
- Modelování probíhá v určitém měřítku (rozlišení) → i když se dva druhy vyskytují ve stejném kvadrátu nemusí spolu přicházet do kontaktu



Obecný postup modelování

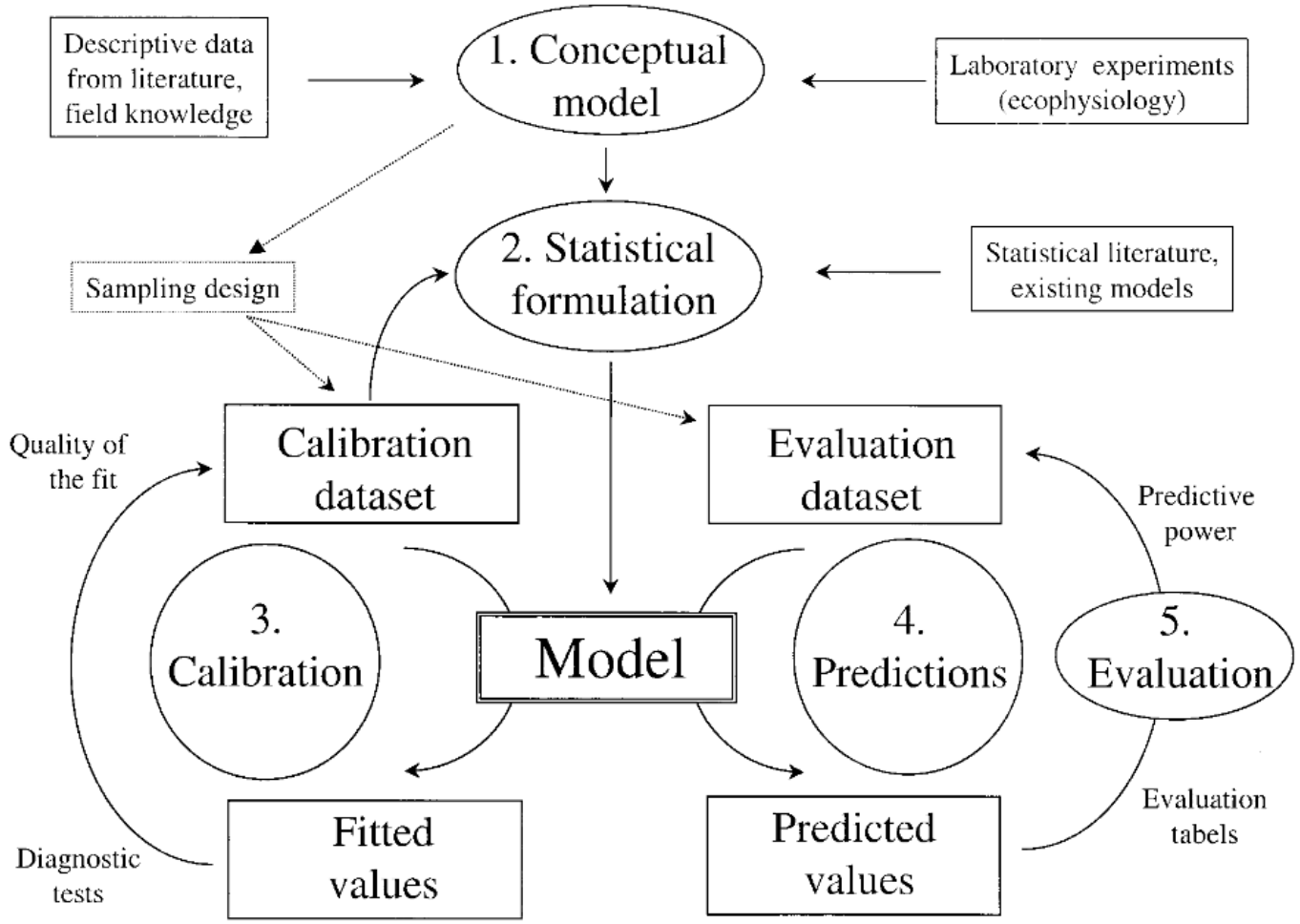






Franklin (2009)

Obecný postup modelování



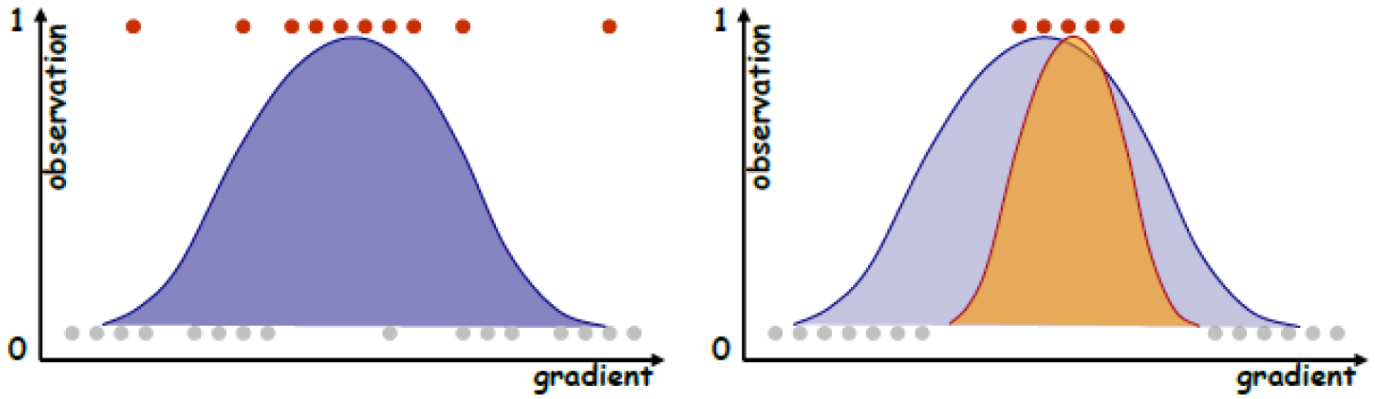
Guisan & Zimmerman (2000)

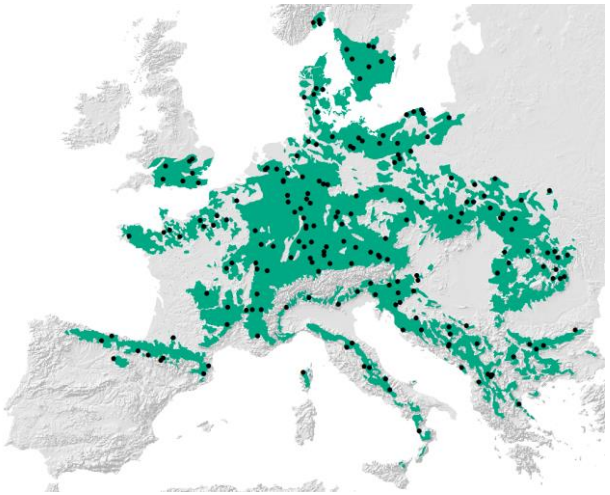
Klíčové kroky

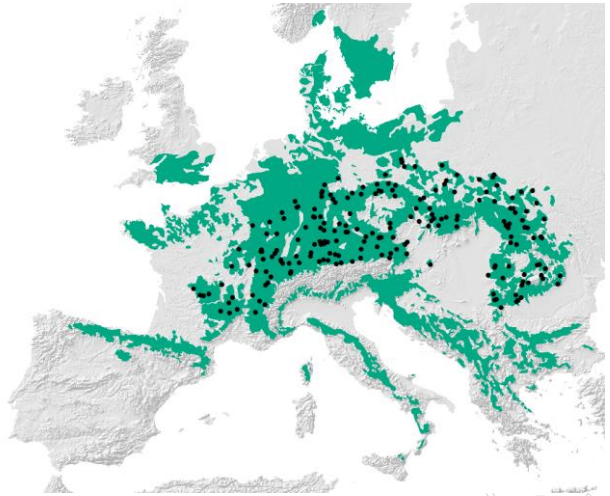
1. Teorie, hypotézy, otázky…
2. Data o rozšíření druhu (velikost vzorku, sampling bias, absenční data…)
3. Environmentální data (dostupnost, rozlišení…)
4. Modelovací metoda
5. Testovací a validační metody
6. Finální projekce modelu do prostoru

Data o výskytu druhu – velikost vzorku

- Čím více, tím lépe?
- Reprezentativní pokrytí
  - Areál
  - Environmentální gradient
- Důležitější extrémy než průměr







Diversity and Distributions, (Diversity Distrib.) (2008) **14**, 763–773

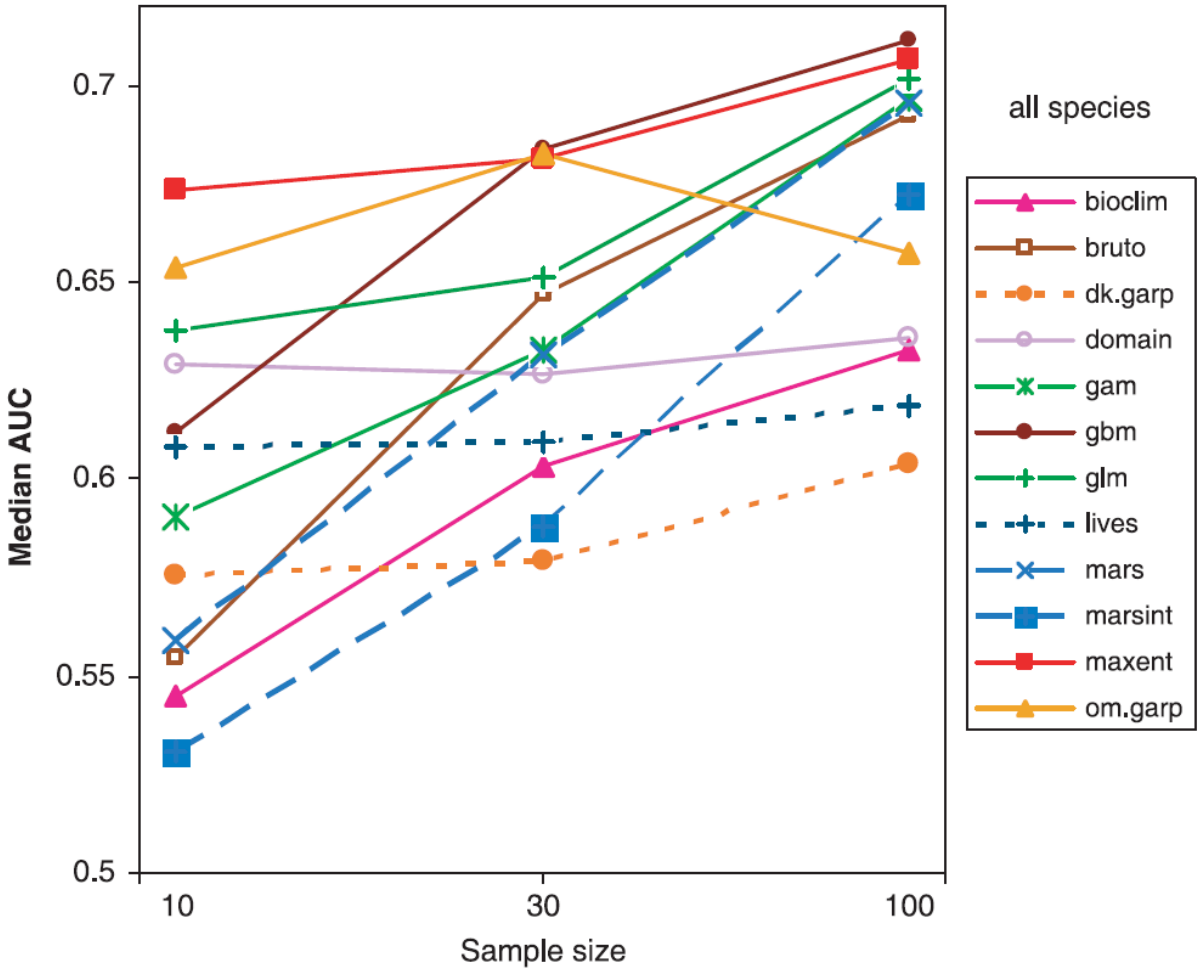
BIODIVERSITY RESEARCH

Effects of sample size on the performance of species distribution models

M. S. Wisz[1*], R. J. Hijmans[2], J. Li[3], A. T. Peterson[4], C. H. Graham[5], A. Guisan[6] and NCEAS Predicting Species Distributions Working Group†

- Porovnání 12 modelovacích metod
- 46 druhů z 6 různých oblastí
- 3 velikosti vzorků (10, 30 a 100)

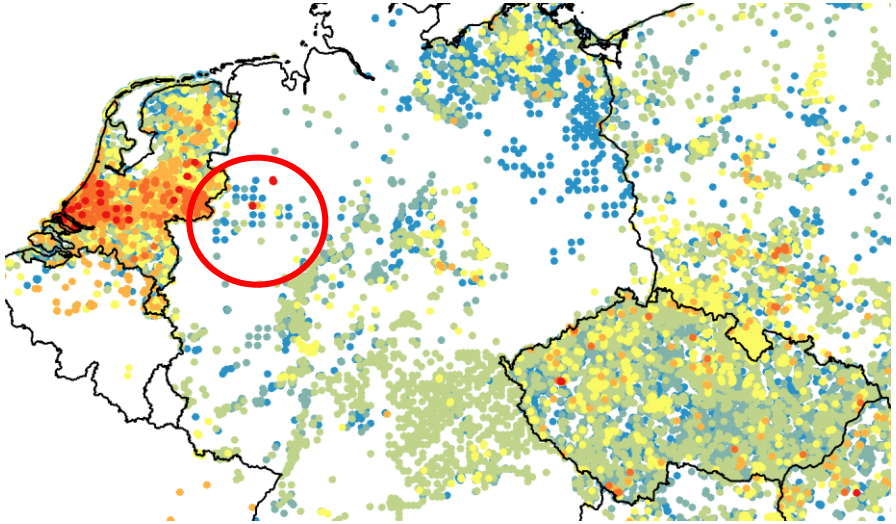
- Se snižujícím se počtem vzorků klesala přesnost modelů
- Nejméně citlivý byl MaxEnt
- Žádná metoda nefungovala dobře s $n < 30$



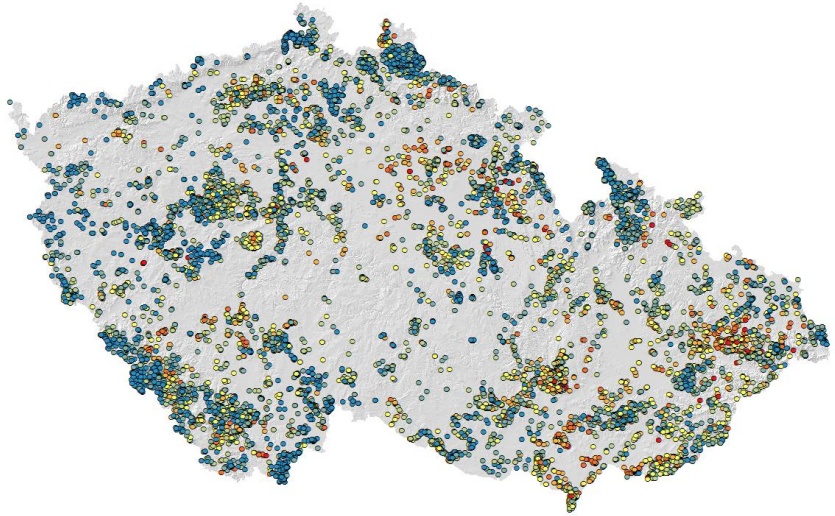
Data o výskytu druhu – sampling bias

- Přesnost lokalizace
  - Lokality zaměřené GPS vs. data z gridů
- Oversampling × undersampling
  - Některým územím je věnována vyšší pozornost než jiným (preferenční snímkování)
- Pseudo-replikace
  - Vzdálenost vzorků nižší než rozlišení enviro. dat
- Autokorelace
  - Téměř vždy
  - Pokud nezůstává v reziduích modelu, je to OK
  - Pokud ano, obtížné testování vlivu env. proměnných
    - Chybějící důležitý prediktor

Trávníky z EVA



Lesní snímky z ČNFD



Journal of Applied Ecology 2008, **45**, 239–247 doi: 10.1111/j.1365-2664.2007.01408.x

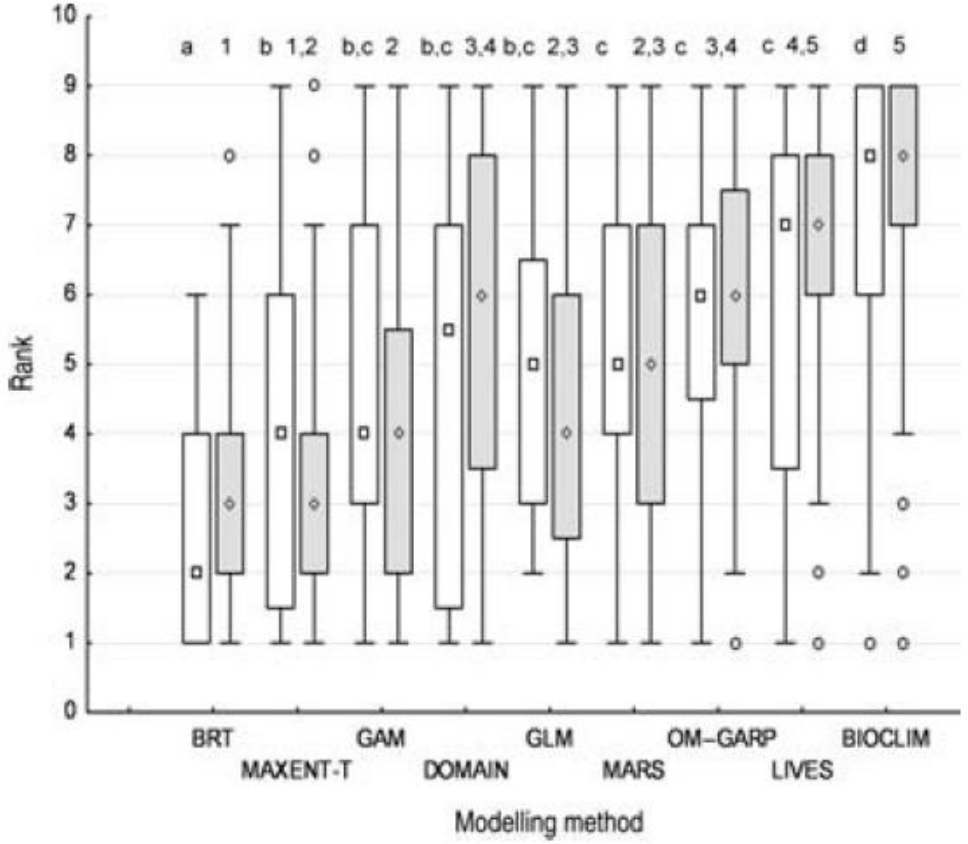
The influence of spatial errors in species occurrence data used in distribution models

Catherine H. Graham[1]*, Jane Elith[2], Robert J. Hijmans[3]†, Antoine Guisan[4], A. Townsend Peterson[5], Bette A. Loiselle[6] and The Nceas Predicting Species Distributions Working Group‡

- 10 modelovacích metod
- 40 druhů ve 4 odlišných regionech

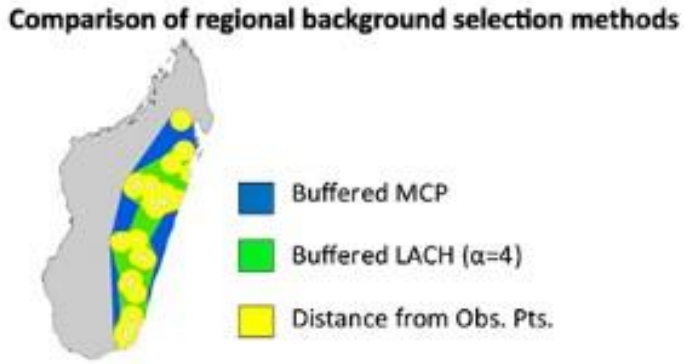
- Chyba lokalizace
  - Posun každé souřadnice o číslo náhodně vybrané z normálního rozdělení s průměrem 0 a SD = 5 km

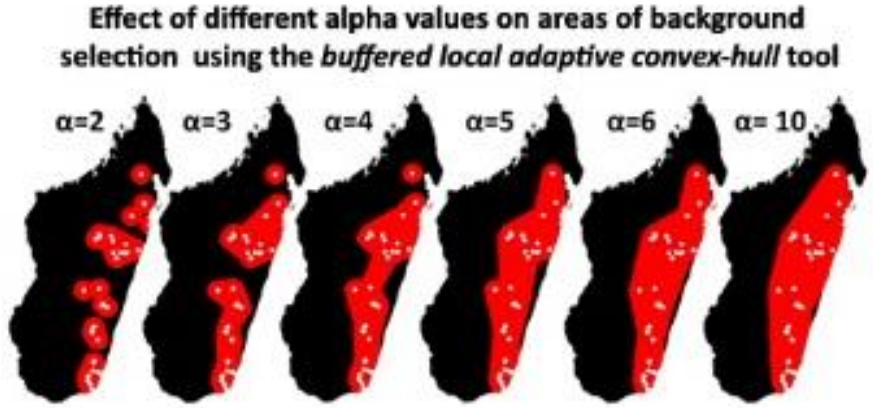
- Chyba v lokalizaci vzorků snížila přesnost modelu ve 3 ze 4 regionů
- I přes chybu v lokalizaci bylo možné pro většinu druhů postavit relativně přesné modely
- MaxEnt a Boosted Regression Trees byly nejméně závislé na nepřesnostech v lokalizaci vzorků



Absenční data

- Ne vždy dostupné nebo použitelné
- Některé metody vyžadují:
  - jen prezenční data (MaxEnt)
  - prezenční a absenční data (GLM, Random Forests)
- Pokud nejsou dostupné, používají se tzv. *pseudo-absences* (*background points*)
  - Náhodně vygenerované body v daném regionu
  - Jejich počet a rozmístění může zásadně ovlivnit výsledek modelu
  - Pokud nemáme reprezentativní vzorek areálu druhu doporučuje se omezit prostor pro generování pseudo-absences





BMC Ecology

BioMed Central

Research article

Open Access

Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data

Mary S Wisz*[1] and Antoine Guisan[2]

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2012, **3**, 327–338 doi: 10.1111/j.2041-210X.2011.00172.x

Selecting pseudo-absences for species distribution models: how, where and how many?

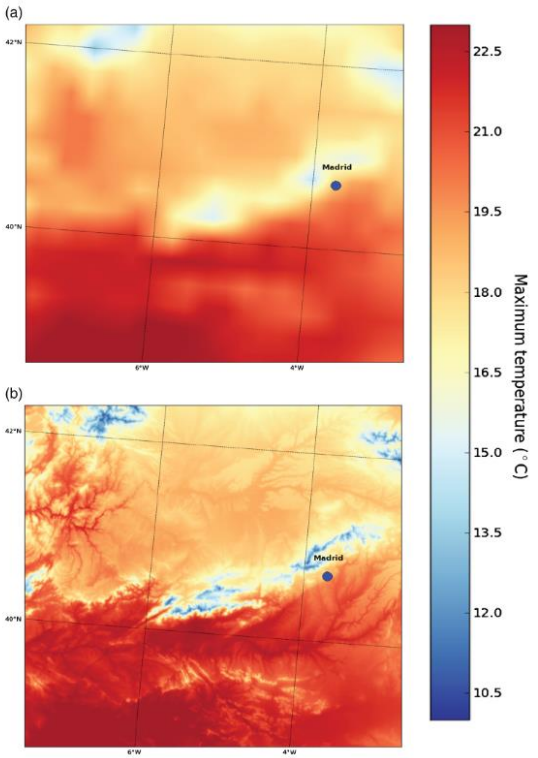
Morgane Barbet-Massin[1]*, Frédéric Jiguet[1], Cécile Hélène Albert[2,3] and Wilfried Thuiller[3]

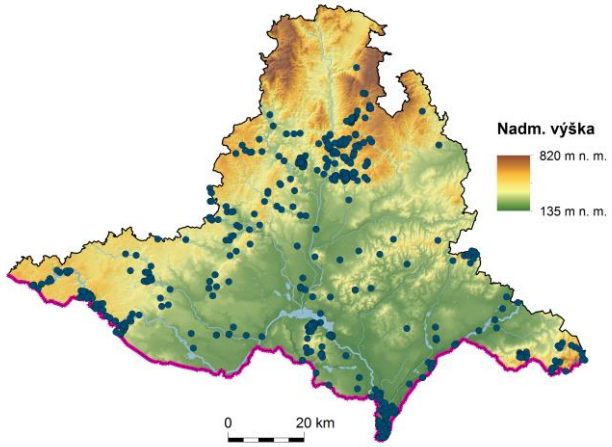
Table 1. How to choose pseudo-absences according to the modelling technique from results of this study, the bold criteria being the most important for the considered modelling technique. 'Same as number of presences, at least 10 runs when less than 1000 PA' means that when more than 300 presence points were considered, 1000 PA should be selected and when 100 or less presences points were considered, a minimum of 10 runs with 100 PA gave the best results

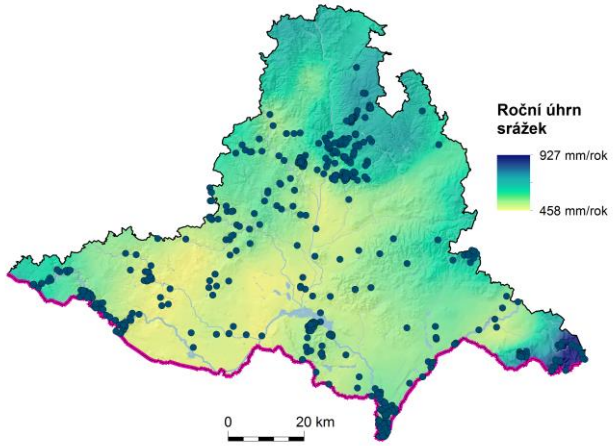
	Method for selecting pseudo-absences	Number of pseudo-absences
GLM, GAM	**'random' performs consistently well, excepted when presences are climatically biased for which '2°far' is the best method**	10 000 PA or a minimum of 10 runs with 1000 PA with an equal weight for presences and absences
MARS	'random' performs consistently well, except when presences are climatically biased for which '2°far' is the best method	**A minimum of 10 runs with 100 PA**
MDA	'2°far' performs consistently better with few presences, 'SRE' performs better with a large number of presences; 'random' performs consistently well with spatially biased presences	**A minimum of 10 runs with 100 PA with an equal weight for presences and absences**
CTA, BRT, RF	'2°far' performs consistently better with few presences, 'SRE' performs better with a large number of presences	**Same as number of presences, 10 runs when less than 1000 PA with an equal weight for presences and absences**

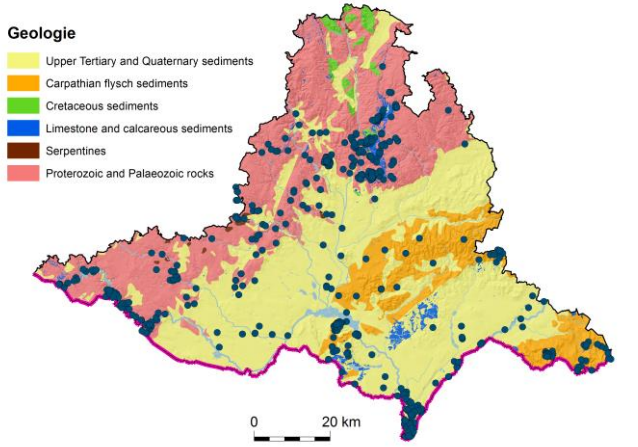
Environmentální data

- Výběr environmentálních dat vždy závisí na:
  - otázkách, které řeším
  - relevanci vzhledem biologickým datům
  - dostupnosti a jejich kvalitě
  - prostorovém měřítku, resp. rozlišení
    - Používat plochojevná zobrazení









Journal of Vegetation Science **27** (2016) 1308–1322

SYNTHESIS

What we use is not what we know: environmental predictors in plant distribution models

Heidi K. Mod, Daniel Scherrer, Miska Luoto & Antoine Guisan

- Relevance env. prediktorů a frekvence jejich použití v SDMs
- 200 studií zamřených na modelování (2010-2015)

- Většina studií nepoužila některé důležité ekofyziologické environmentální proměnné (vlhkost, půdní pH, živiny atp.)
- Počet používaných relevantních prediktorů stagnuje posledních 15 let

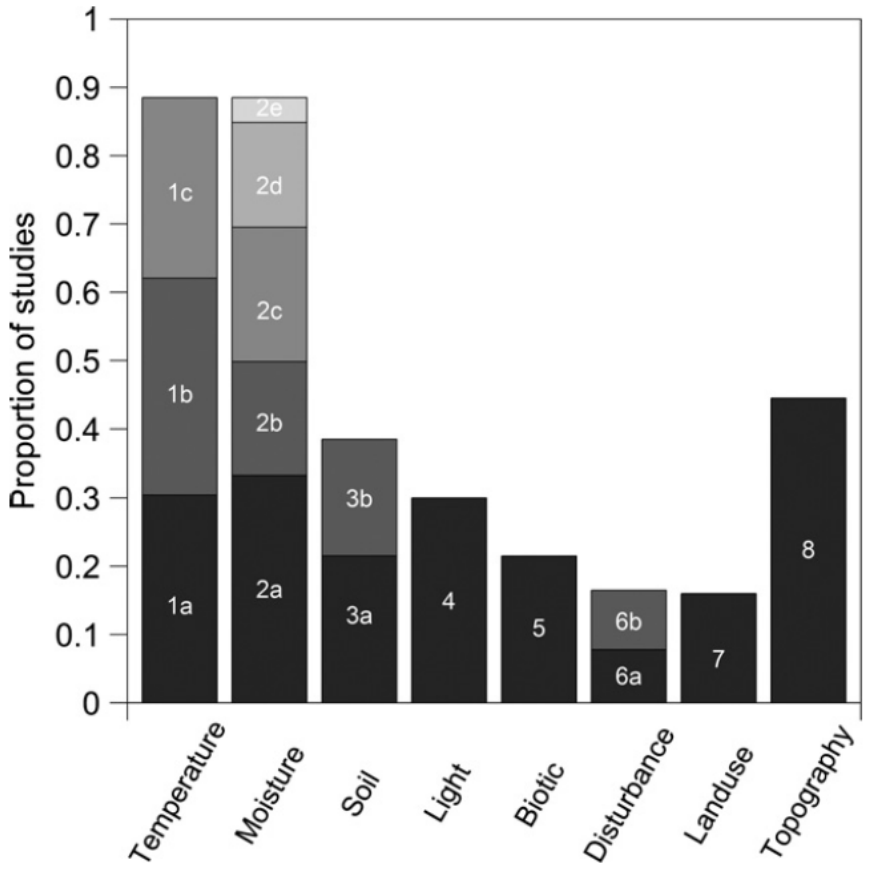


Fig. 3. Proportion of studies in which each predictor class was used: 1a mean temperature; 1b extreme temperature; 1c seasonality of temperature; 2a mean precipitation; 2b extreme precipitation; 2c seasonality of precipitation; 2d water balance; 2e soil moisture; 3a pH/bedrock; 3b nutrients; four radiation; five biotic interactions; 6a natural disturbances; 6b human disturbances; seven land use; eight topography.

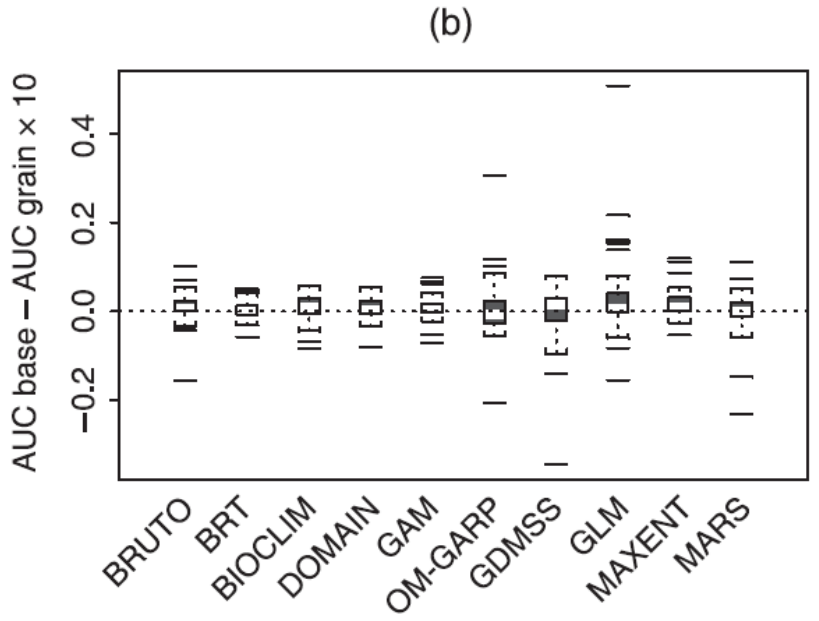
Diversity and Distributions, (Diversity Distrib.) (2007) 13, 332–340

BIODIVERSITY RESEARCH

Sensitivity of predictive species distribution models to change in grain size

Antoine Guisan[1]*, Catherine H. Graham[2], Jane Elith[3], Falk Huettmann[4] and the NCEAS Species Distribution Modelling Group†

- 10 modelovacích metod
- 50 druhů ve 5 odlišných regionech
- 10× snížené rozlišení env. proměnných



	Mean AUC	Rank base	SD AUC	Mean AUC 10×	Rank 10×	SD AUC 10×	*P*-value (AUC diff.)
(b) Techniques							
BRUTO	0.6928	4	0.1306	0.6859	2	0.1291	*0.0067*
BRT	0.7040	1	0.1321	0.6993	1	0.1299	0.2060
BIOCLIM	0.6391	9	0.1074	0.6313	10	0.1134	*0.0078*
DOMAIN	0.6383	10	0.1248	0.6337	9	0.1262	0.1248
GAM	0.6939	3	0.1338	0.6844	3	0.1287	**0.0004**
OM-GARP	0.6605	8	0.1284	0.6542	8	0.1129	0.8963
GDMSS	0.6726	7	0.1390	0.6731	6	0.1216	0.1476
GLM	0.6838	5	0.1352	0.6546	7	0.1447	**0.0049**
MAXENT	0.6954	2	0.1241	0.6793	4	0.1216	**0.0008**
MARS	0.6744	6	0.1583	0.6760	5	0.1543	0.4602

Article

Progress in Physical Geography
2014, Vol. 38(1) 79–96

DOI: 10.1177/0309133313512667
ppg.sagepub.com
SAGE

Very high resolution environmental predictors in species distribution models: Moving beyond topography?

Jean-Nicolas Pradervand*
University of Lausanne, Switzerland

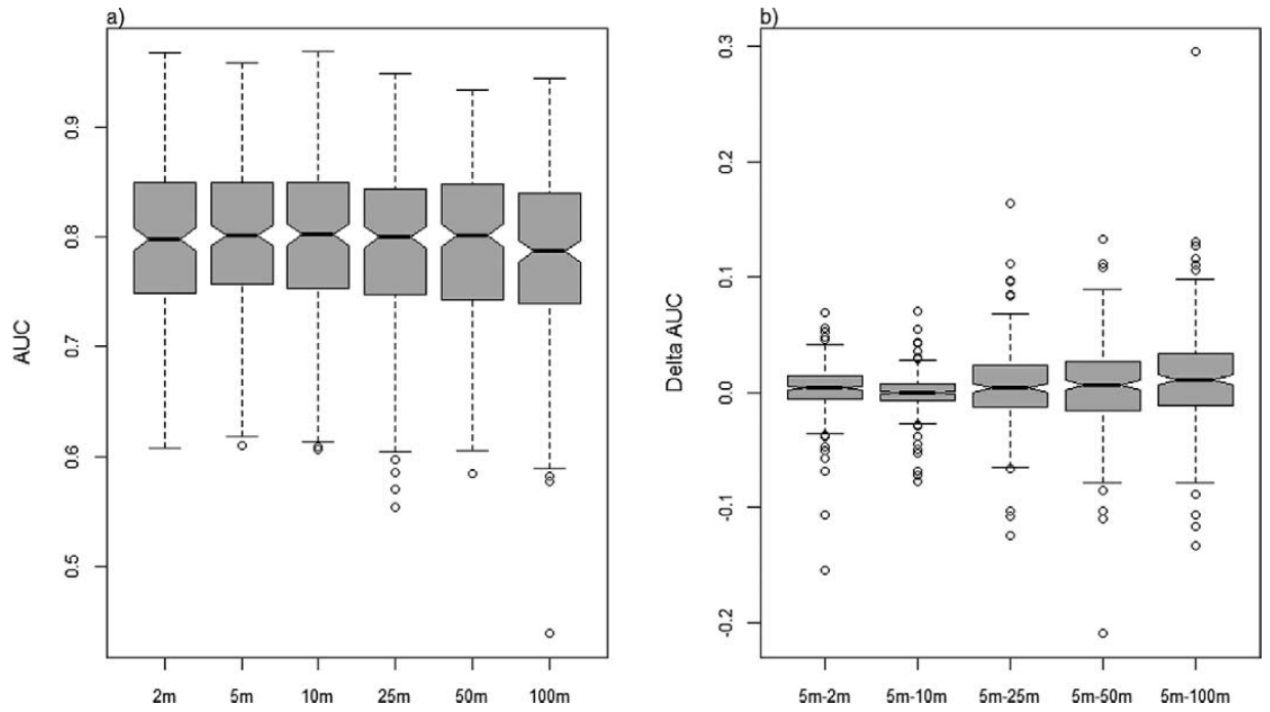


Figure 1. (a) Boxplots of the AUC values representing the mean of the three modelling techniques across all species, for the six different resolutions examined. (b) Difference (delta) in the mean AUC values across all species of the means for each species across the three modelling techniques. Differences are always expressed between the 5 m models (best AUCs) and all other resolutions (2, 10, 25, 50 and 100 m). On average, models at the different resolutions show roughly no difference.

Metody

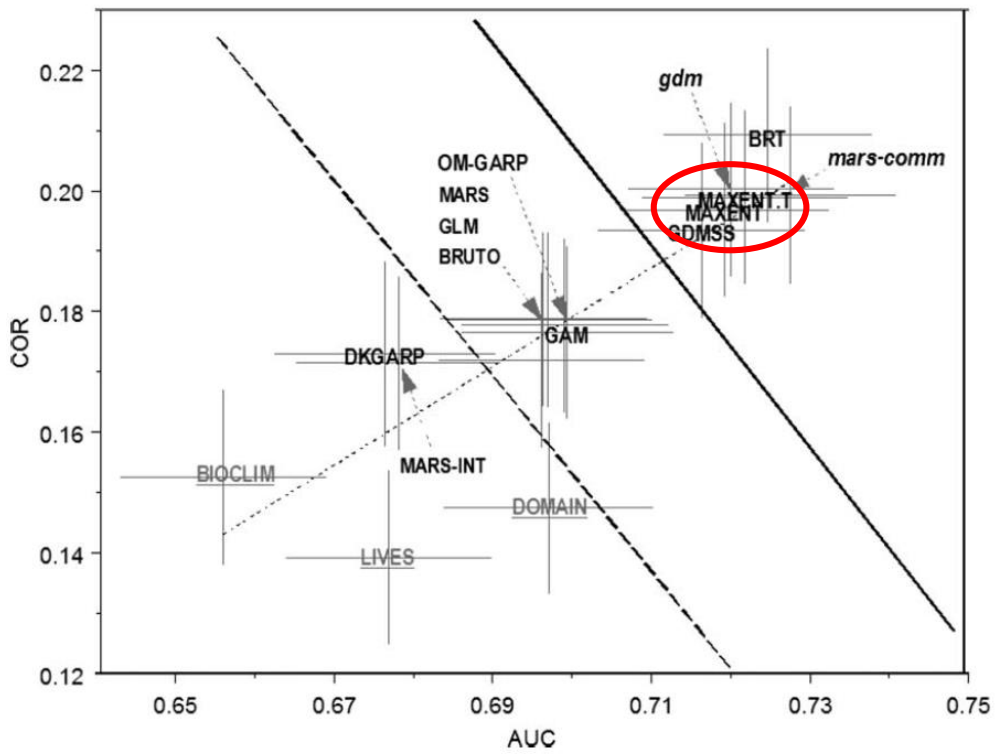
- Základní 3 kategorie modelovacích metod

1) **Profile techniques** (jednoduché metody využívající např. environmentální vzdálenosti mezi vzorky)
   - BIOCLIM
   - DOMAIN
   - Ecological Niche Factor Analysis (ENFA)
2) **Regression-based techniques**
   - Generalized Linear Models (GLM)
   - Generalized Additive Models (GAM)
   - Multivariate Adaptive Regression Splines (MARS)
3) **Machine-learning techniques**
   - Boosted Regression Trees (BRT)
   - **Random Forests (RF)**
   - Support Vector Machines (SVM)
   - **MaxEnt**

ECOGRAPHY 29: 129–151, 2006

Novel methods improve prediction of species' distributions from occurrence data

Jane Elith*, Catherine H. Graham*, Robert P. Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J. Hijmans, Falk Huettmann, John R. Leathwick, Anthony Lehmann, Jin Li, Lucia G. Lohmann, Bette A. Loiselle, Glenn Manion, Craig Moritz, Miguel Nakamura, Yoshinori Nakazawa, Jacob McC. Overton, A. Townsend Peterson, Steven J. Phillips, Karen Richardson, Ricardo Scachetti-Pereira, Robert E. Schapire, Jorge Soberón, Stephen Williams, Mary S. Wisz and Niklaus E. Zimmermann



Methods in Ecology and Evolution
British Ecological Society
Methods in Ecology and Evolution 2013, **4**, 327–335
doi: 10.1111/2041-210x.12022

The impact of modelling choices in the predictive performance of richness maps derived from species-distribution models: guidelines to build better diversity models

Blas M. Benito[1]*, Luis Cayuela[2] and Fabio S. Albuquerque[1]

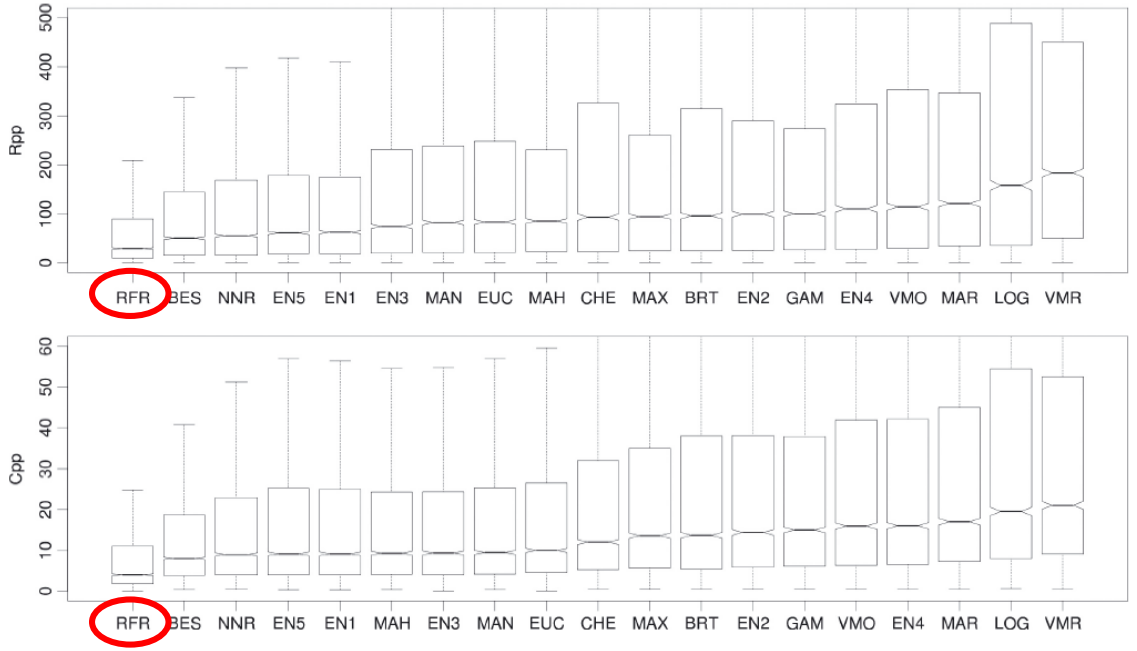


Fig. 5. Ordered boxplots showing the performance of each S-SDM series in terms of species richness (Rpp, plot a) and species composition (Cpp, plot b). Lower values indicate better predictive performances. The y axis was clipped to the higher quartiles to provide a better visualization.

Machine-learning methods

Klasifikační a regresní stromy (*Classification And Regression Trees*, CART)

- Rozhodovací strom – sada hierarchicky uspořádaných rozhodovacích pravidel
- Podle typu závislé proměnné je dělíme na:
  - Klasifikační
  - Regresní
- Postupně dělí závislou proměnnou tak, aby její hodnoty uvnitř uzlu byly co nejhomogennější a zároveň mezi uzly co nejrozdílnější
- Homogenitu počítá pomocí kriteriální statistiky
  - Minimální kvadratická chyba pro regresní stromy $Q(\mathrm{T})$

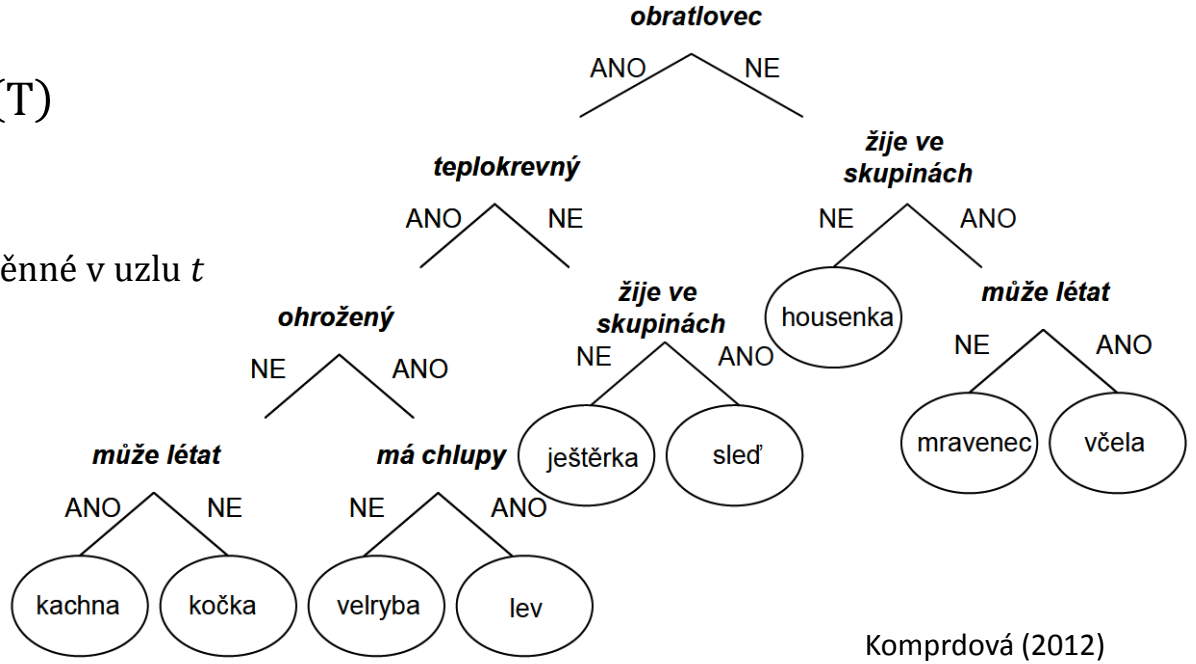
$$y_t = \frac{1}{N_t}\sum y_{i(t)} \qquad Q_t(\mathrm{T}) = \frac{1}{N_t}\sum_{i=1}^{N_t}(y_i - \bar{y}_t)^2$$

N_t je počet pozorování v uzlu t; $y_{i(t)}$ jsou hodnoty závislé proměnné v uzlu t

  - Gini index pro klasifikační stromy

$$GI = \sum_{c=1}^{J} p_{tc}(1 - p_{tc}) = 1 - \sum_{c=1}^{J} p_{tc}^2$$

p_{ct} je pravděpodobnost kategorie c v uzlu t



Komprdová (2012)

Algoritmus růstu stromu CART

1. Rozděl soubor na trénovací a testovací. Tento poměr se určuje na základě počtu pozorování a účelu studie
2. Najdi nejlepší rozdělení každého z prediktorů
   a) *Pro spojité proměnné* - seřaď hodnoty každého prediktoru (spojitého nebo ordinálního) od nejmenší po největší → pro všechna možná dělení závislé proměnné na dva dceřiné uzly spočítej kriteriální statistku → rozdělení (dělící hodnota), pro které je kriteriální statistika nejmenší se použije pro rozdělení závislé proměnné
   b) *Pro kategoriální proměnné* – vyzkoušej všechny možnosti rozdělení závislé proměnné pomocí kategorií vysvětlující proměnné a pro každé dělení spočítej kriteriální statistiku → rozdělení (dělící hodnota), pro které je kriteriální statistika nejmenší se použije pro rozdělení závislé proměnné
3. Rozděl soubor na dva dceřiné uzly $t1$ a $t2$ podle hodnoty prediktoru vybrané v kroku 2
4. Opakuj krok 2 a 3, dokud není dosaženo některého z pravidel pro zastavení růstu stromu
5. Použij testovací soubor k ověření vhodné velikosti stromu, a pokud je strom příliš velký, prořež strom (*prune tree*)

Kritéria pro zastavení růstu stromu (*stopping rules*)

1. Terminální uzel obsahuje pouze jedno pozorování
2. Všechna pozorování v uzlu mají stejnou hodnotu všech prediktorů (s použitím vybraných prediktorů již nelze závislou proměnnou dále dělit)
3. Všechna pozorování v uzlu mají stejnou hodnotu závisle proměnné (uzel je zcela homogenní)

Definovaná kritéria

1. Maximální počet větvení daného stromu
2. Maximální počet pozorování v koncovém uzlu
3. Frakce pozorování v uzlu, která již nemůže být oddělena
4. Velikost chyby v potenciálních dceřiných uzlech (uzel se nerozdělí, pokud střední kvadratická chyba (MSE) nebo procento nesprávně klasifikovaných vzorků v důsledku rozdělení překročí určitou hranici)

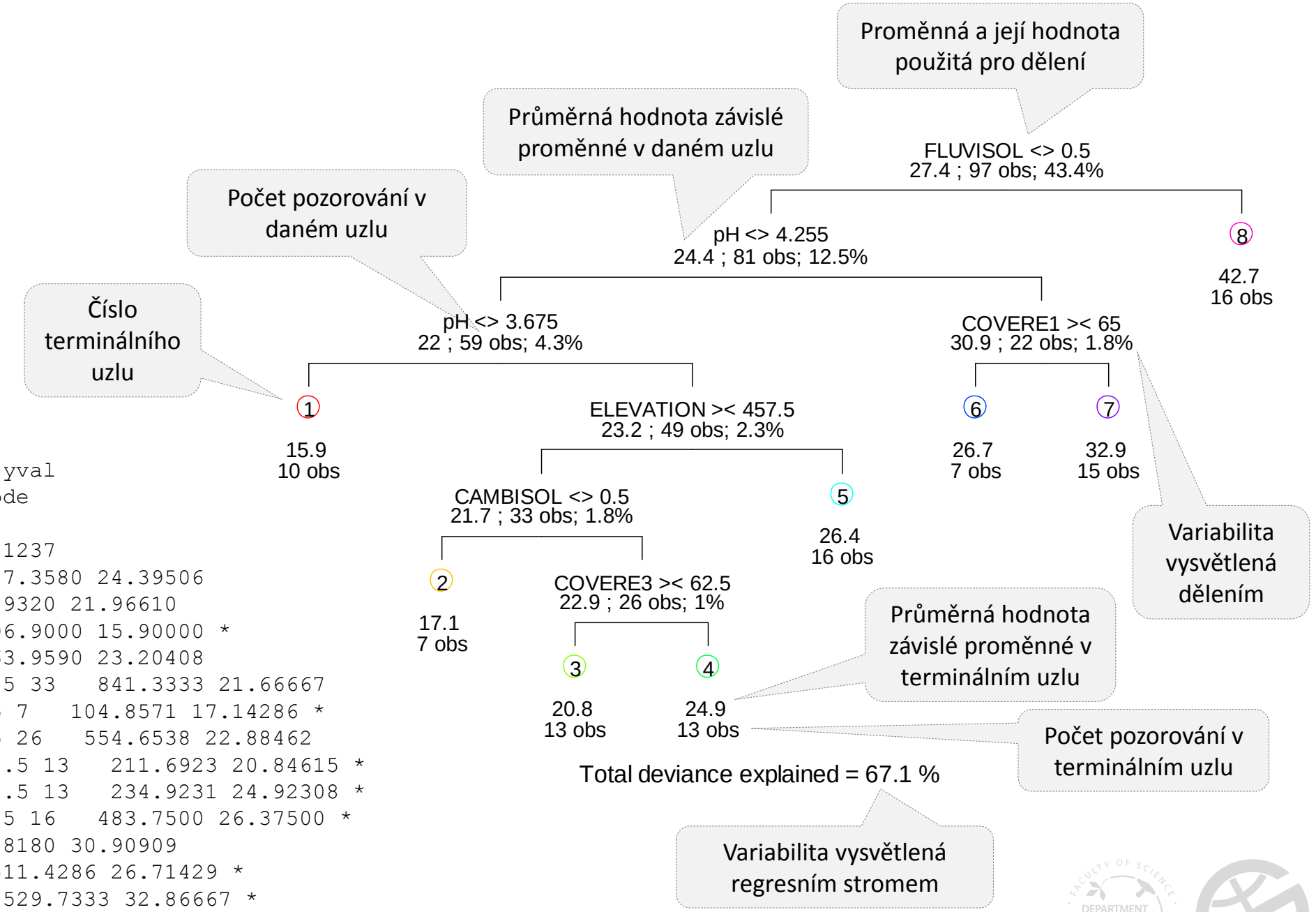
CART v R

`rpart {rpart}`

```
n= 97

node), split, n, deviance, yval
      * denotes terminal node

 1) root 97 10293.5100 27.41237
   2) FLUVISOL< 0.5 81  4807.3580 24.39506
     4) pH< 4.255 59  2303.9320 21.96610
       8) pH< 3.675 10   296.9000 15.90000 *
       9) pH>=3.675 49  1563.9590 23.20408
        18) ELEVATION>=457.5 33   841.3333 21.66667
          36) CAMBISOL< 0.5 7   104.8571 17.14286 *
          37) CAMBISOL>=0.5 26   554.6538 22.88462
            74) COVERE3>=62.5 13   211.6923 20.84615 *
            75) COVERE3< 62.5 13   234.9231 24.92308 *
        19) ELEVATION< 457.5 16   483.7500 26.37500 *
     5) pH>=4.255 22  1221.8180 30.90909
      10) COVERE1>=65 7   511.4286 26.71429 *
      11) COVERE1< 65 15   529.7333 32.86667 *
   3) FLUVISOL>=0.5 16  1015.4380 42.68750 *
```



Vegetační snímky z transektů v údolí Vltavy. Zelený & Chytrý (2007)

Výběr optimálního stromu

- K určení optimální velikosti stromu lze použít kritérium složitosti stromu (*cost-complexity criterium)*
  - Složitost stromu = jeho velikost (počet terminálních uzlů)

$$C_\alpha(T_1) = DT_1 + \alpha|T_1|$$

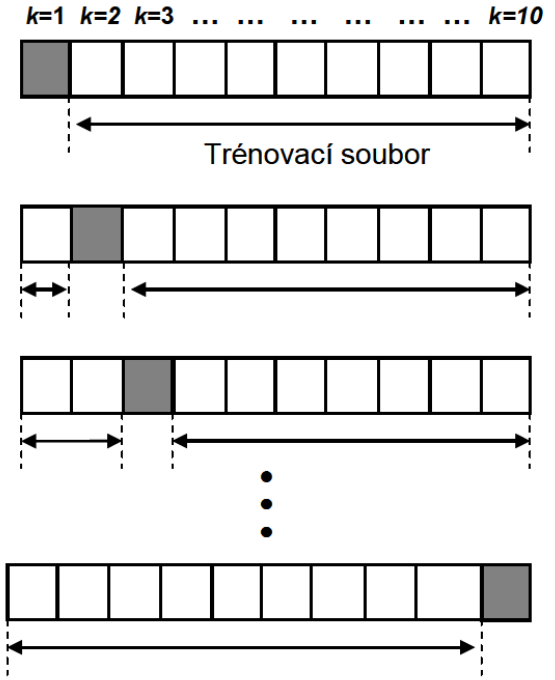
T_1 je počet terminálních uzlů stromu; DT_1 je chyba stromu T_1

Parametr $\alpha \geq 0$ vyjadřuje kompromis mezi velikostí stomu a jeho přesností

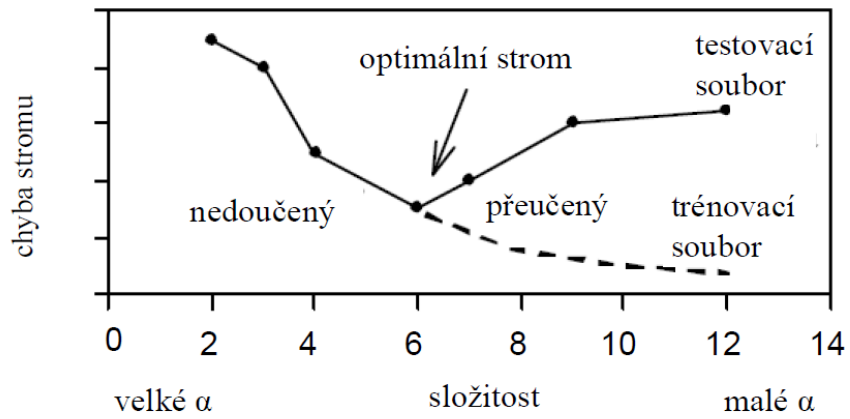
K odhadu α se používá křížová validace

Křížová validace (*Cross-validation*)

- Pozorování jsou rozdělena do *k* nezávislých podsouborů
- Jeden podsoubor se vždy použije pro testování (pozorování nejsou použita při tvorbě modelu)
- Ostatní podsoubory, tj. *k*-1 skupin se použije pro tvorbu modelu
- Celkem je vytvořeno *k* modelů otestovaných na *k* testovacích souborech
- Vybereme strom s největší přesností, ale zároveň rozdíl v chybě mezi testovacím a trénovacím souborem musí být co nejmenší
- Přesnost stromu
  - Klasifikační strom – podíl správně zařazených pozorování
  - Regresní stromu – koeficient determinance (R^2)
- Chyba regresního
  - Pro trénovací soubor $e(t) = 1 - R^2_{tren}$
  - Pro testovací soubor $e'(t) = 1 - R^2_{test}$

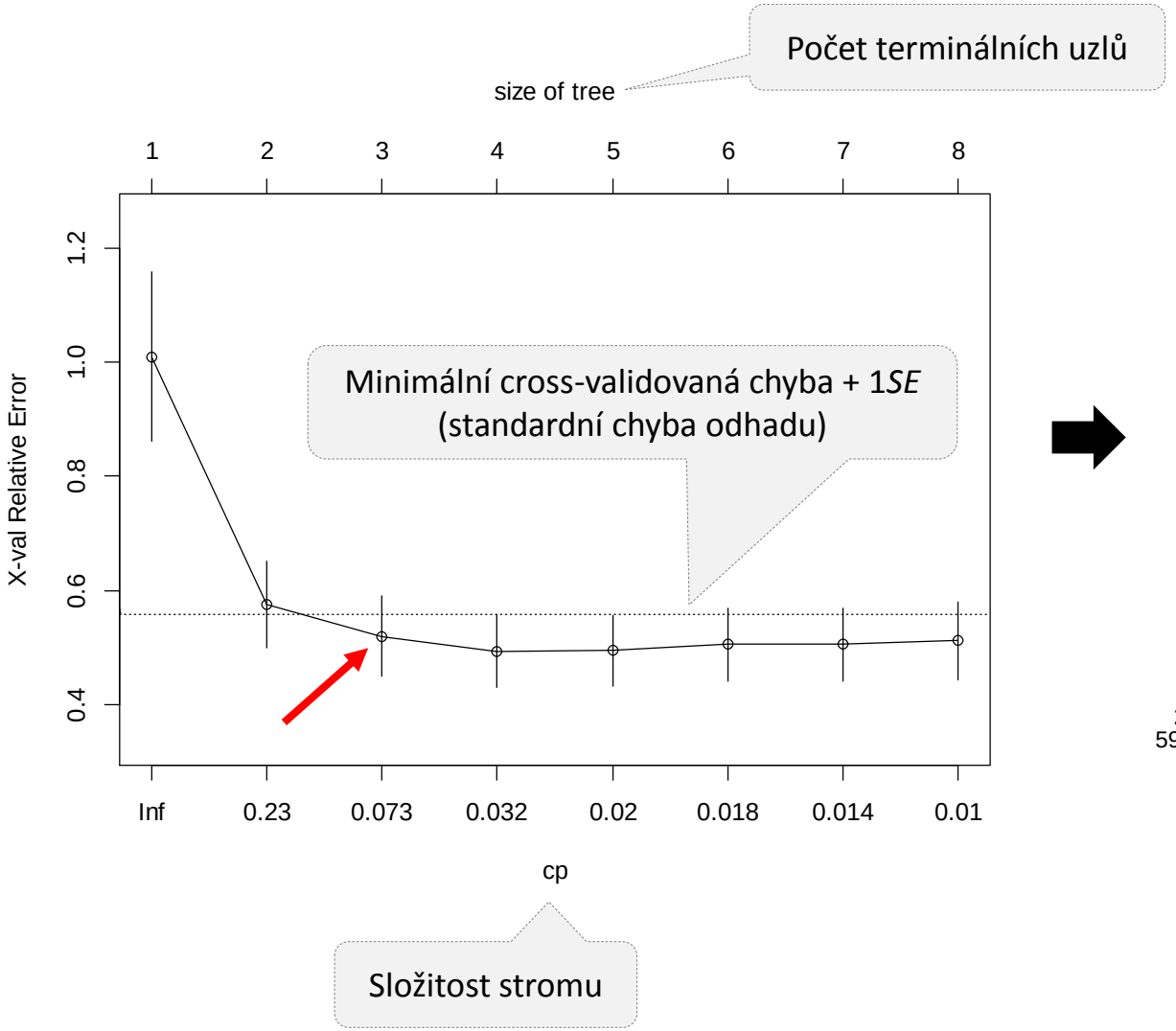


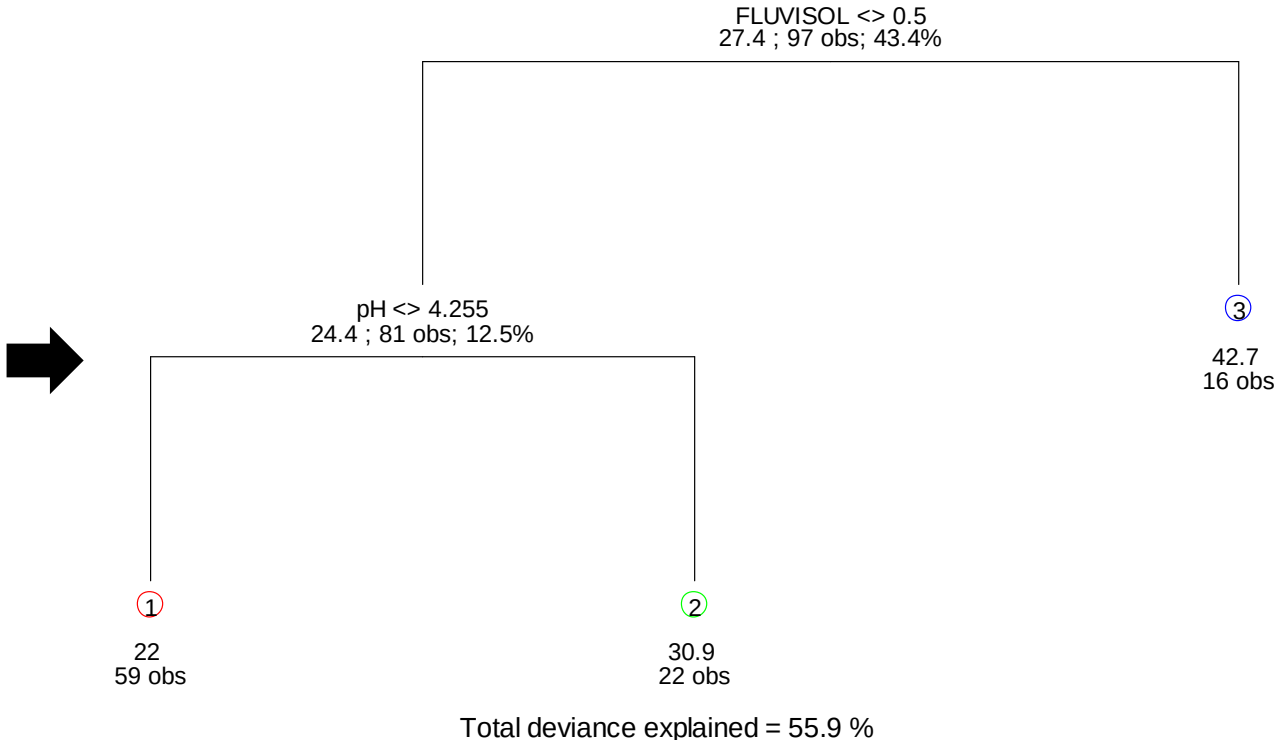
Obr. 2.11 *Princip rozdělení souboru na testovací a trénovací pro k = 10. Tmavá políčka označují testovací soubor.*



Obr. 2.12 *Hledání optimálního stromu. Složitost stromu na ose x je reprezentována počtem terminálních uzlů [6].*

Křížová validace (*Cross-validation*)





Vegetační snímky z transektů v údolí Vltavy. Zelený & Chytrý (2007)

Výhody a nevýhody CART

Výhody

- Neklade žádné podmínky na typ rozdělení závisle proměnné ani prediktorů
- Závisle proměnná i prediktory mohou být všech typů (kategoriální, ordinální i spojité)
- Je možné použít korelované prediktory, protože strom roste hierarchicky a pro dělení se vybírá vždy jen jeden prediktor (mj. ze všech možných korelovaných)
- Výsledky přesnosti stromu lze snadno porovnat s výsledky jiných modelů (R^2)
- Snadné grafické znázornění v podobě grafu se stromovou strukturou, z čehož plyne jednoduchá interpretace získaných výsledků

Nevýhody

- Nestabilita - malá změna v datech způsobí změny v rozhodovacích pravidlech uvnitř uzlů, což může vést ke změně výsledných klasifikací/predikcí
- Vzhledem k nestabilitě je nutná opatrnost při interpretaci stromu
- Stromy jsou nevhodné pro malý počet vzorků a velký počet kategorií závisle proměnné
- Měření přesnosti stromu je výrazně závislé na krosvalidačním mechanizmu a dalších parametrech při validaci modelu ve fázi učení (např. pravidla pro zastavení růstu stromu)

Landscape classification of the Czech Republic based on the distribution of natural habitats

Klasifikace krajiny České republiky na základě rozšíření přírodních biotopů

Jan Divíšek[1,2,3], Milan Chytrý[3], Vít Grulich[3] & Lucie Poláková[4]

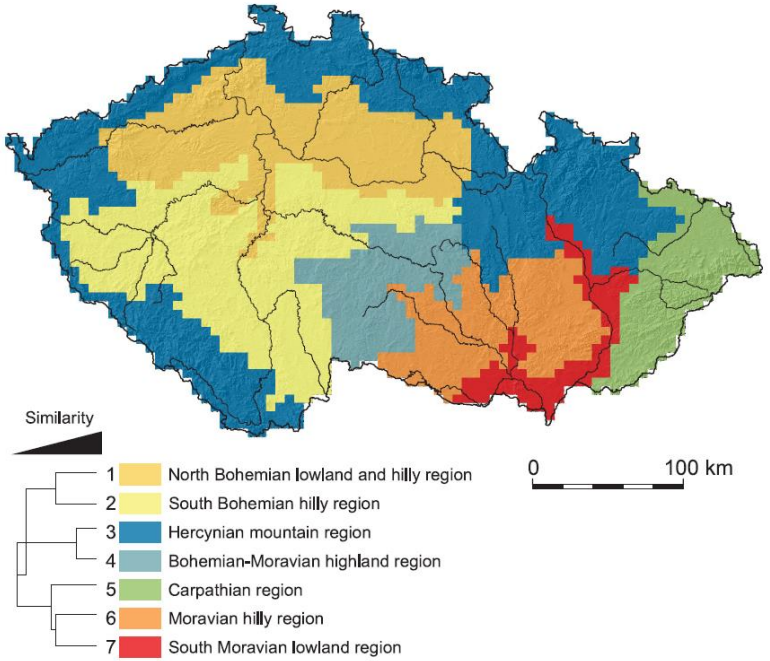


Fig. 7. – Regions of the Czech Republic based on the spatially constrained clustering with the optimal number of seven clusters according to cross-validation procedure. Reversal in the dendrogram is due to spatial constraints.

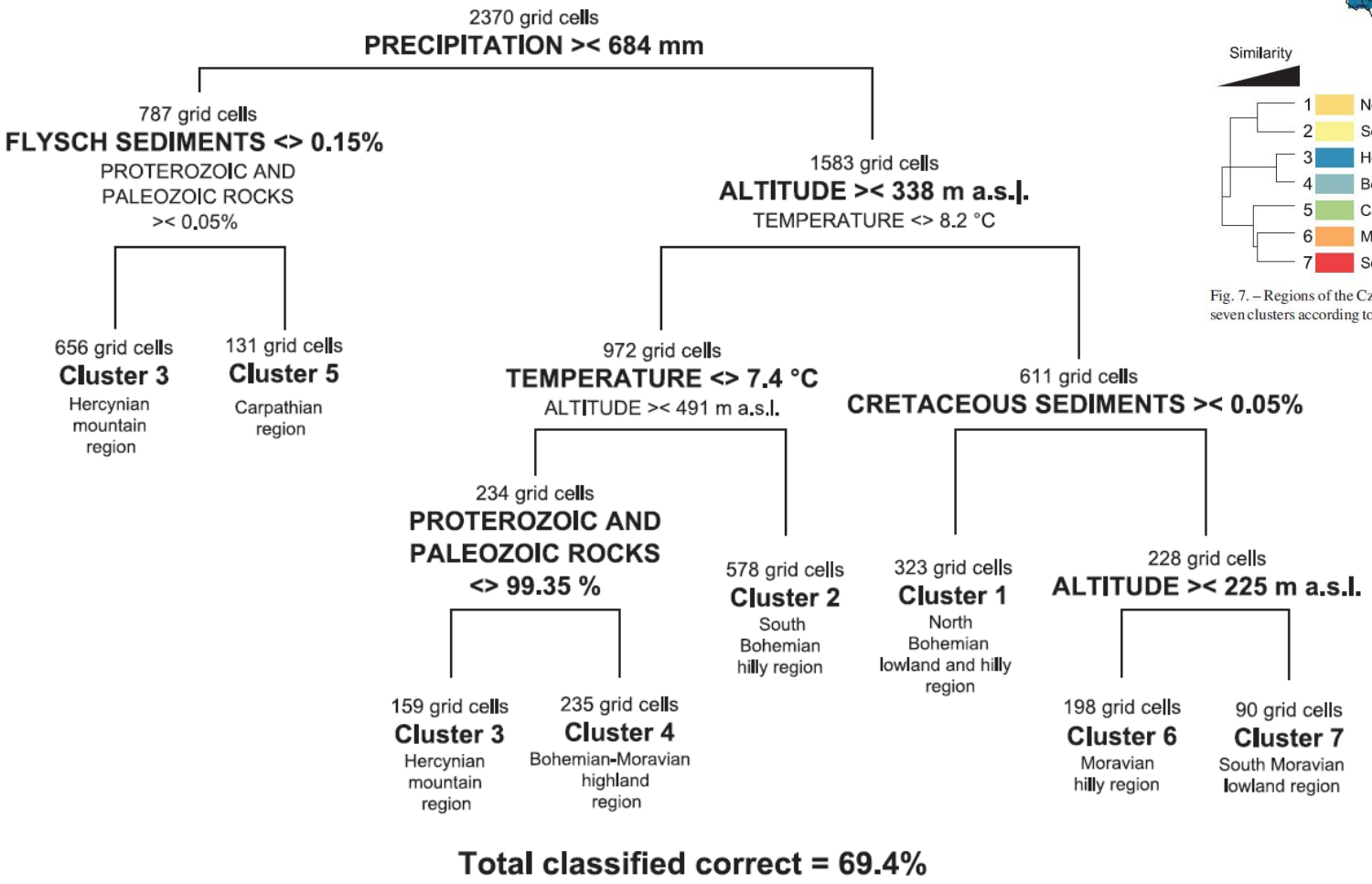


Fig. 9. – Classification tree describing the separation of the seven regions (clusters) resulting from spatially constrained clustering in terms of abiotic factors. See Fig. 5 for details.

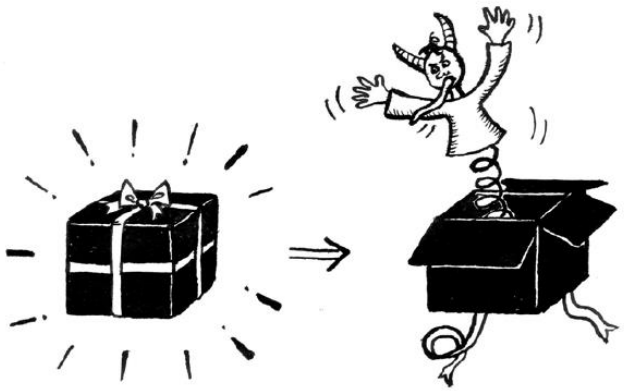
Náhodné lesy

Random Forests

Náhodné lesy (*Random Forest*)

- Metoda je založena na CART → kombinací více stromů vzniká náhodný les
- Náhodné lesy odstraňují problémy spojené s CART, zejména jejich nestabilitu
- Jsou však složitější a méně přehledné → někdy (dříve) považováno za tzv. „black-box“
- Původně pro velké soubory s velkým množstvím prediktorů
- Lze použít (stejně jako CART) pro klasifikaci a regresi
- V biogeografii se dnes často používají pro prostorové modelování rozšíření druhů
  → velice efektivní

`randomForest {randomForest}`



Princip výpočtu náhodného lesa

1. Vytvoř bootstrapový podsoubor L_i o velikosti N → trénovací soubor
   - bootstrapový výběr = náhodný výběr
2. Vyber náhodně m prediktorů
3. Vytvoř strom T_i na bootstrapovém souboru L_i pouze s použitím m náhodně vybraných prediktorů
   - pro klasifikaci je hodnota $m = \sqrt{p}$ a minimální velikost uzlu je 1
   - pro regresi je hodnota $m = p/3$ a minimální velikost koncového uzlu je 5
4. Pomocí vytvořeného stromu predikuj *oob* (*out-of-bag, out of bootstrap sample*) data (testovací soubor) → výpočet chyby stromu
5. Opakuj kroky 1-4 až do konečného počtu stromů v lese (500 ale lze nastavit uživatelem)
6. Spočítej celkový výsledek klasifikace/predikce celého lesa většinovým hlasováním/průměrováním

Výsledek náhodného lesa v R

`randomForest {randomForest}`

```
Call:
 randomForest(x = env, y = div, ntree = 200, mtry = 7, importance = T, nPerm = 999)
               Type of random forest: regression
                     Number of trees: 200
No. of variables tried at each split: 7

          Mean of squared residuals: 54.87076
                    % Var explained: 49.44
```

Počet stromů v lese

Počet proměnných k dispozici pro každé dělení

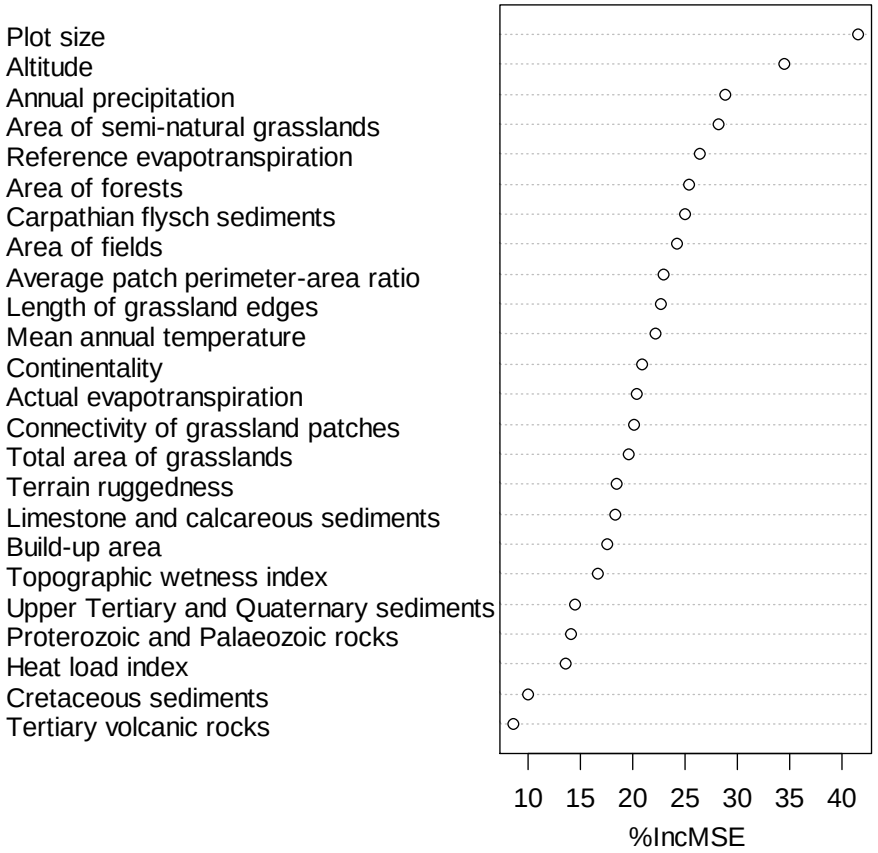
Vysvětlená variabilita

Průměr ze čtverců reziduálních hodnot

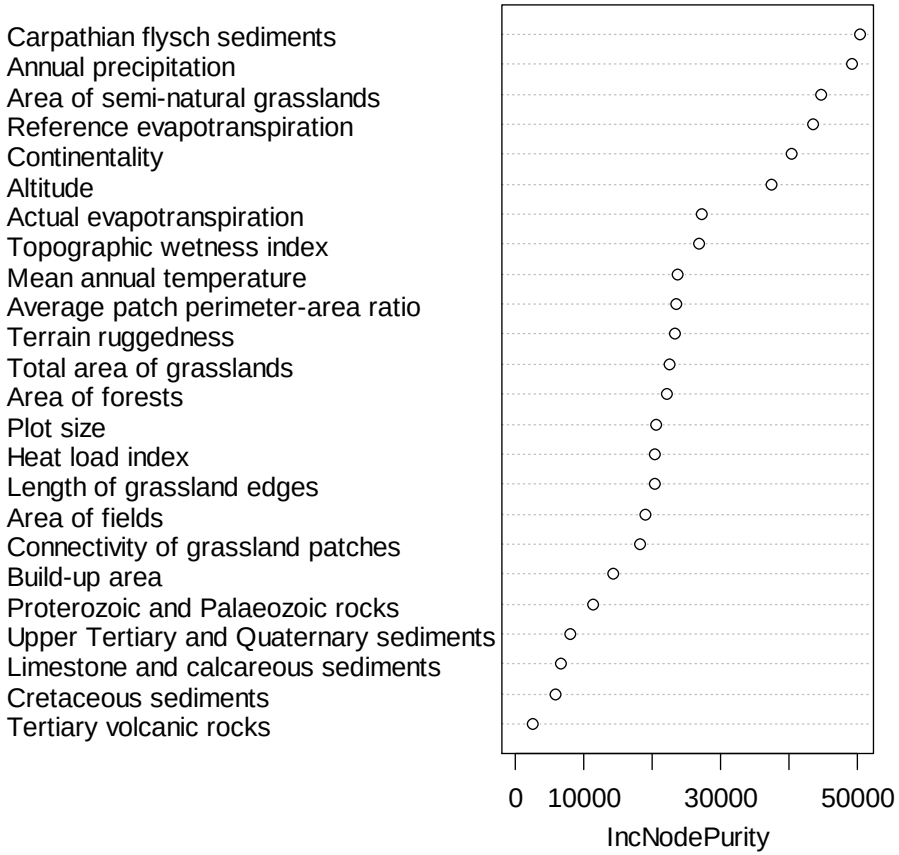
`importance {randomForest}`
`varImpPlot {randomForest}`

Důležitost proměnných v modelu

Mean decrease in accuracy



Mean decrease in node impurity



O kolik permutace (znáhodnění) dané proměnné zhorší predikci testovacích dat (*oob* dat). Čím nižší predikční schopnost (větší predikční chyba MSE) modelu tím významnější proměnná.

Měří jak dobře je vysvětlující proměnná schopna dělit závislou proměnnou. Čím homogennější shluky oddělí tím je významnější.

Důležitost proměnných v modelu

Mean decrease in accuracy (%)

1. Pro každý strom je spočítána chyba na testovacích datech (*oob* data)
   - Pro klasifikací se měří podílem chybně klasifikovaných vzorků (*misclassification rate*)
   - Pro regresi MSE
2. Testovaná proměnná se zamíchá a opět se spočítá predikční chyba
3. Spočítá se rozdíl predikčních chyb pro daný strom
4. Rozdíly se zprůměrují přes všechny stromy a podělí se směrodatnou odchylkou

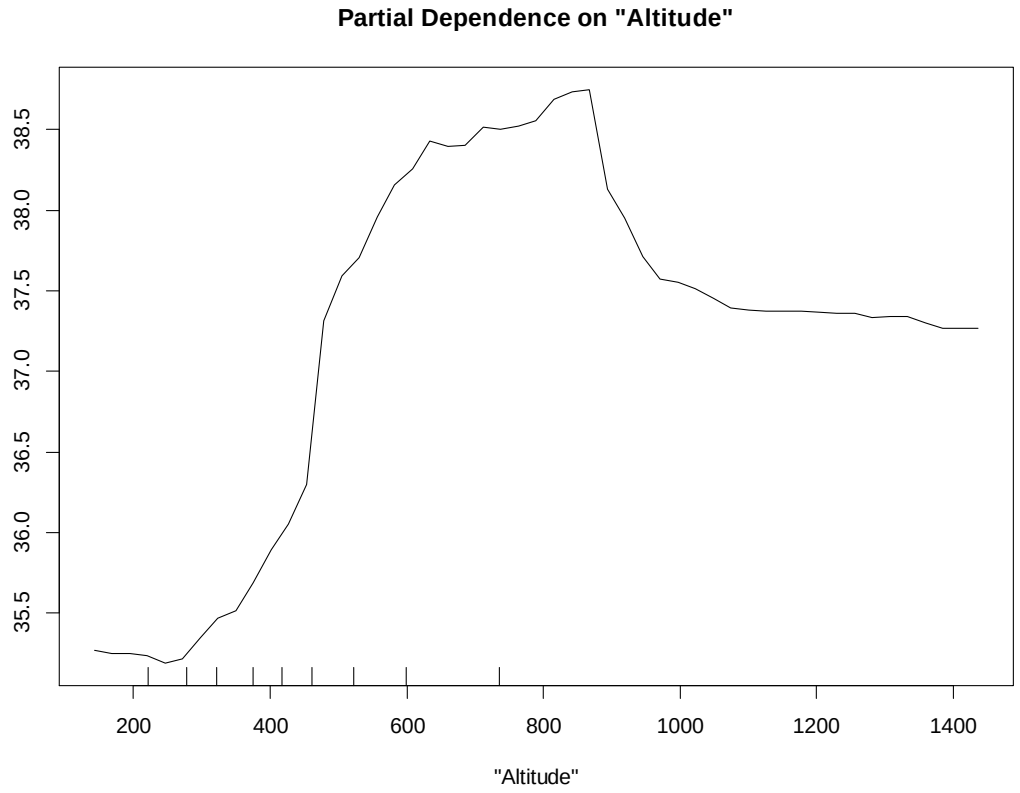
Mean decrease in node impurity

1. Vždy, když je proměnná použita pro dělení se spočítá o kolik poklesne míra „heterogenity“ uzlu
   - Pro klasifikaci se měří Gini indexem
   - Pro regresi reziduální sumou čtverců (RSS)

Partial dependence plot

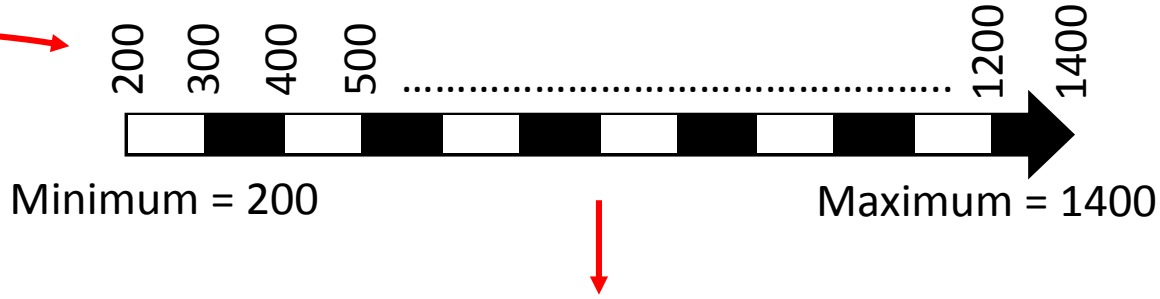
`partialPlot {randomForest}`

- Ukazuje marginální efekt vybrané proměnné, tj. vztah závislé a vysvětlující proměnné
  - Jedná se o vztah mezi závislou a vysvětlující proměnnou za situace, kdy vlivy všech ostatních proměnných jsou zprůměrovány



Princip výpočtu

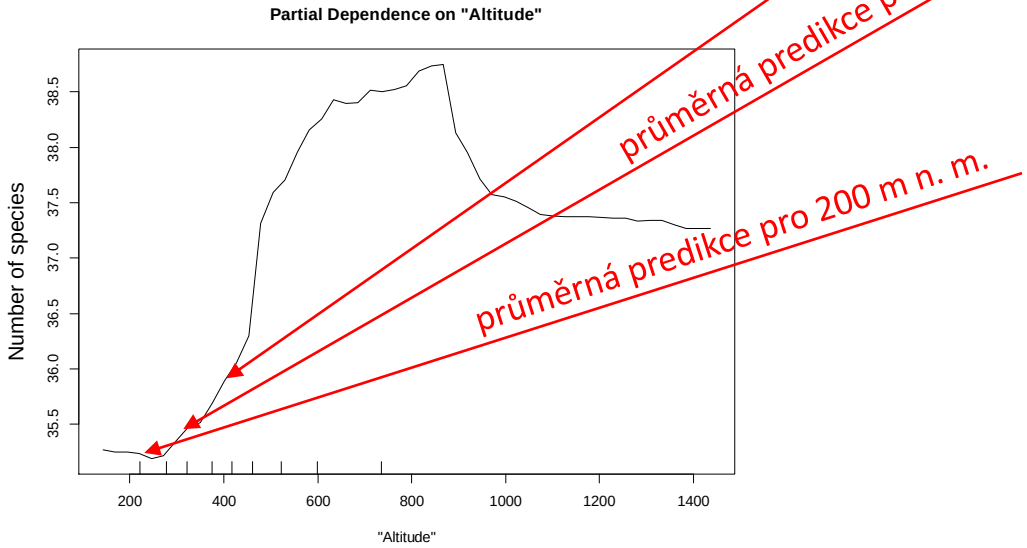
	Temp	Prec	Altitude	pH	Slope
Samp1	6.5	435	253	6.1	3.9
Samp2	5.9	869	560	5.9	5.8
Samp3	6.2	501	280	6.2	2.4
Samp4	7.1	467	200	6.6	1.2
…	…	…	…	…	…
SampN	2.1	1102	1400	5.2	9.5



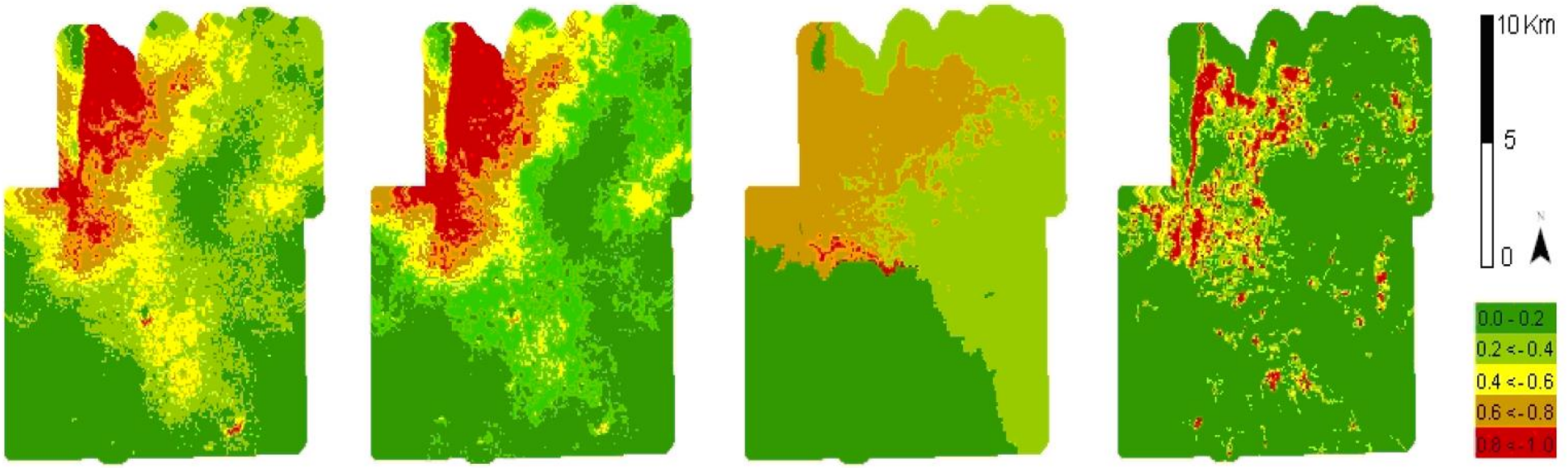
	Temp	Prec	Altitude	pH	Slope
Samp1	6.5	435	400	6.1	3.9
Samp2	5.9	869	400	5.9	5.8

	Temp	Prec	Altitude	pH	Slope	2.4
Samp1	6.5	435	300	6.1	3.9	1.2
Samp2	5.9	869	300	5.9	5.8	…

	Temp	Prec	Altitude	pH	Slope	2.4	9.5
Samp1	6.5	435	200	6.1	3.9	1.2	
Samp2	5.9	869	200	5.9	5.8	…	
Samp3	6.2	501	200	6.2	2.4	9.5	
Samp4	7.1	467	200	6.6	1.2		
…	…	…	…	…	…		
SampN	2.1	1102	200	5.2	9.5		



Random Forest in species distribution modelling



Additive Logistic Model (GAM)

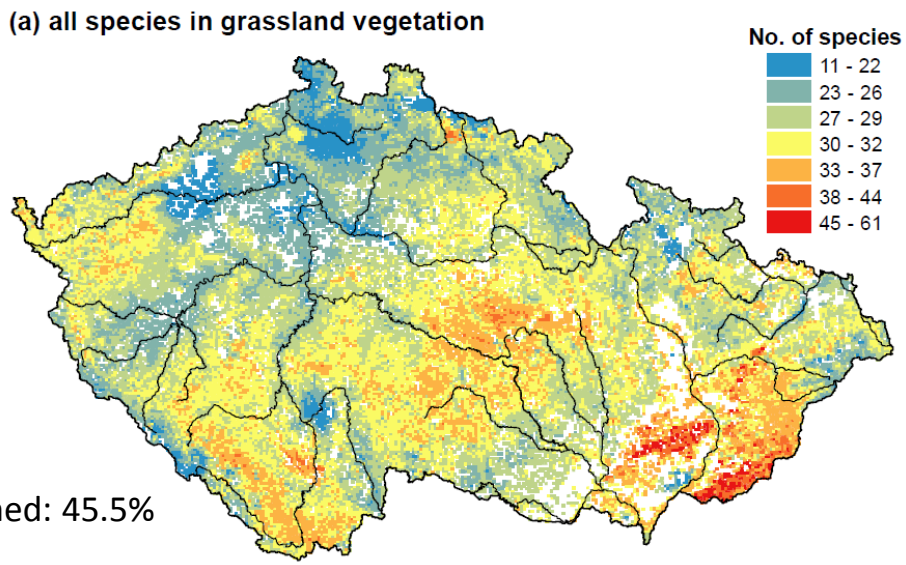
Logistic Model (GLM)

Classification Tree Model

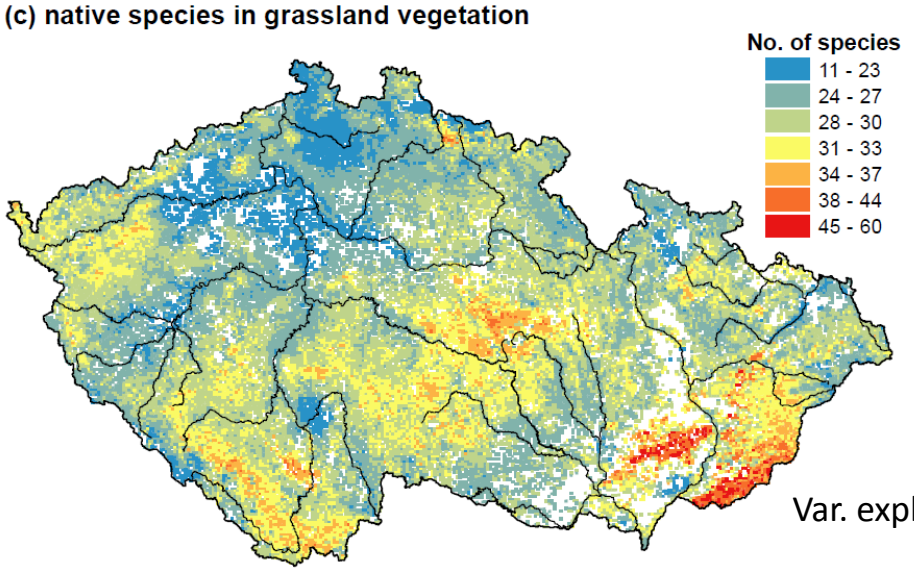
Random Forest Model

Edwards et al., ECOCHANGE, Lausanne, Sep 2009

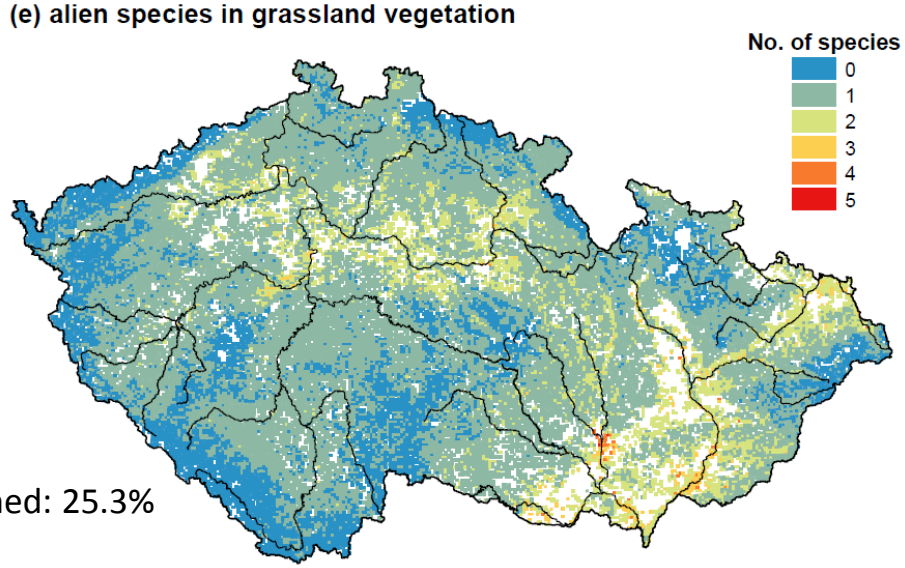
Predikce druhové bohatosti travinné vegetace ČR



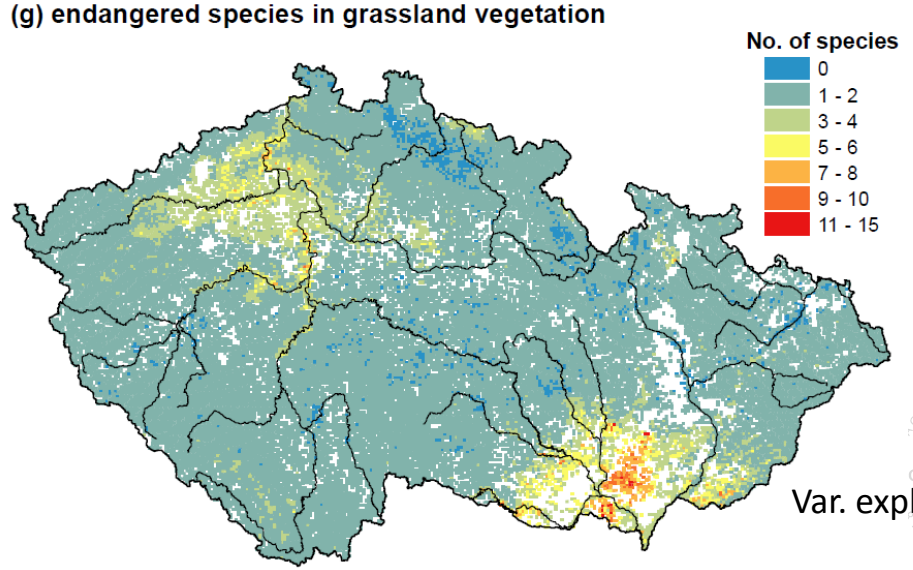
Var. explained: 45.5%



Var. explained: 50.1%

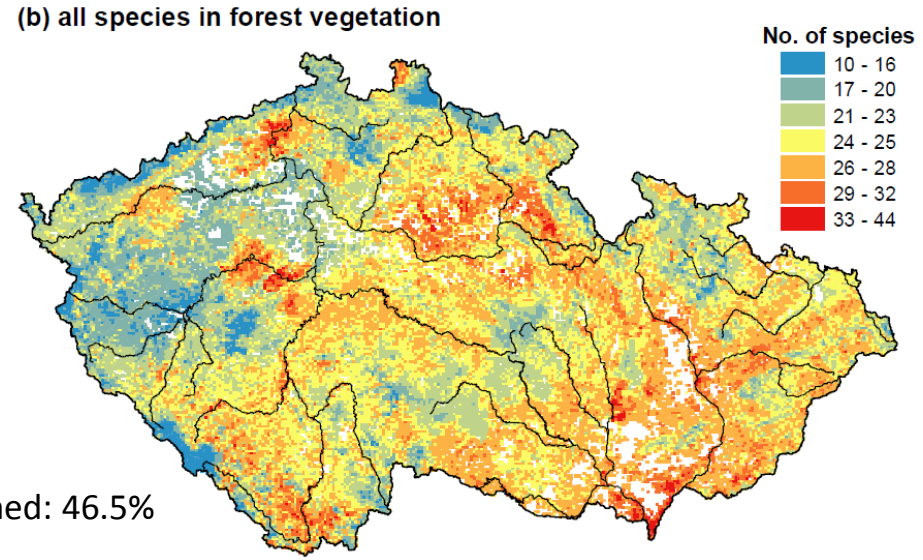


Var. explained: 25.3%

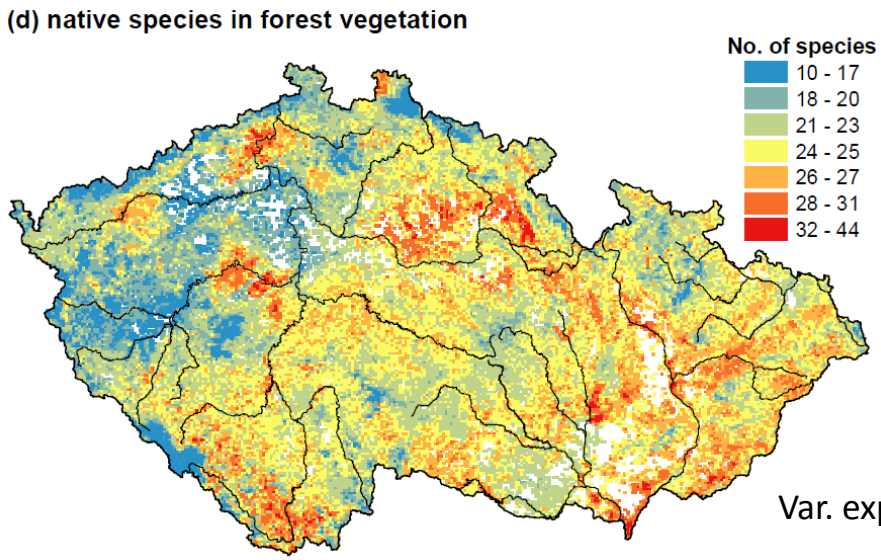


Var. explained: 53.8%

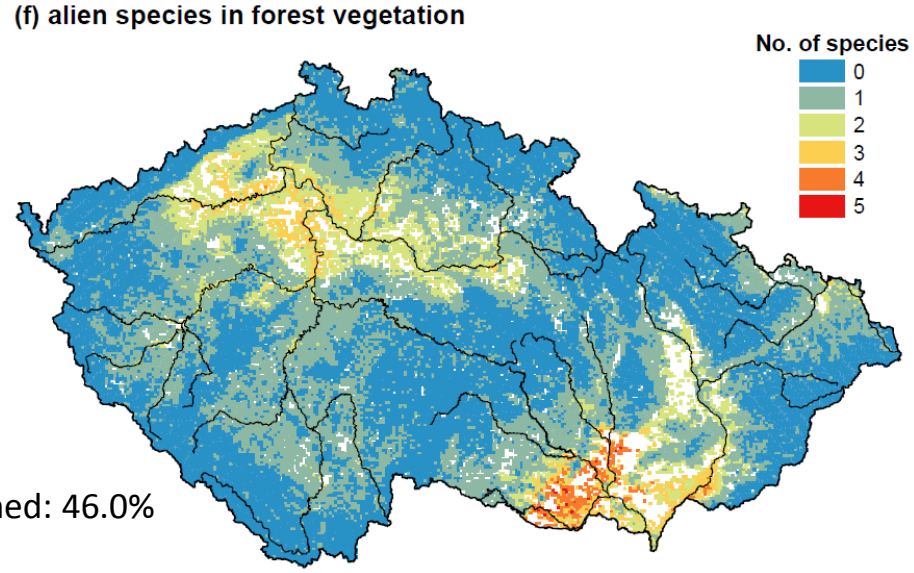
Predikce druhové bohatosti lesní vegetace ČR



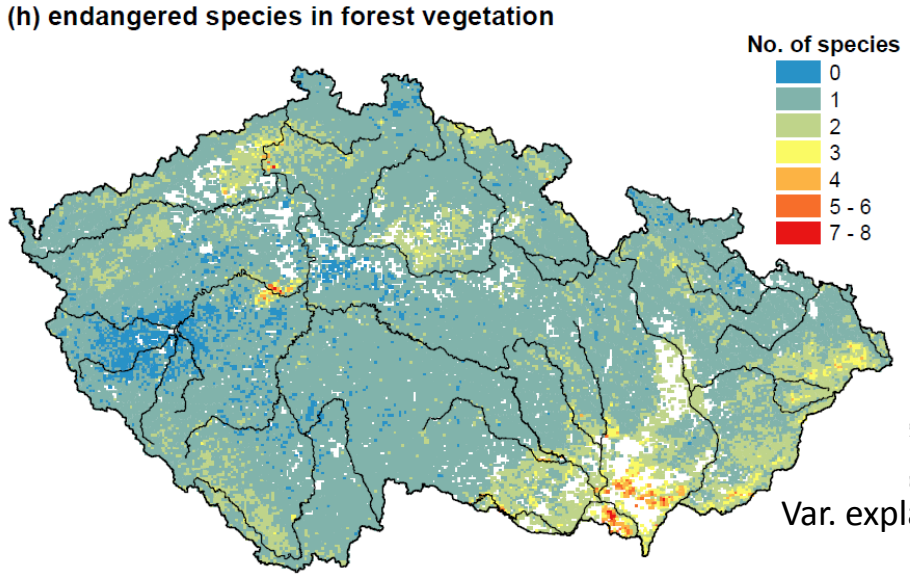
Var. explained: 46.5%



Var. explained: 45.8%

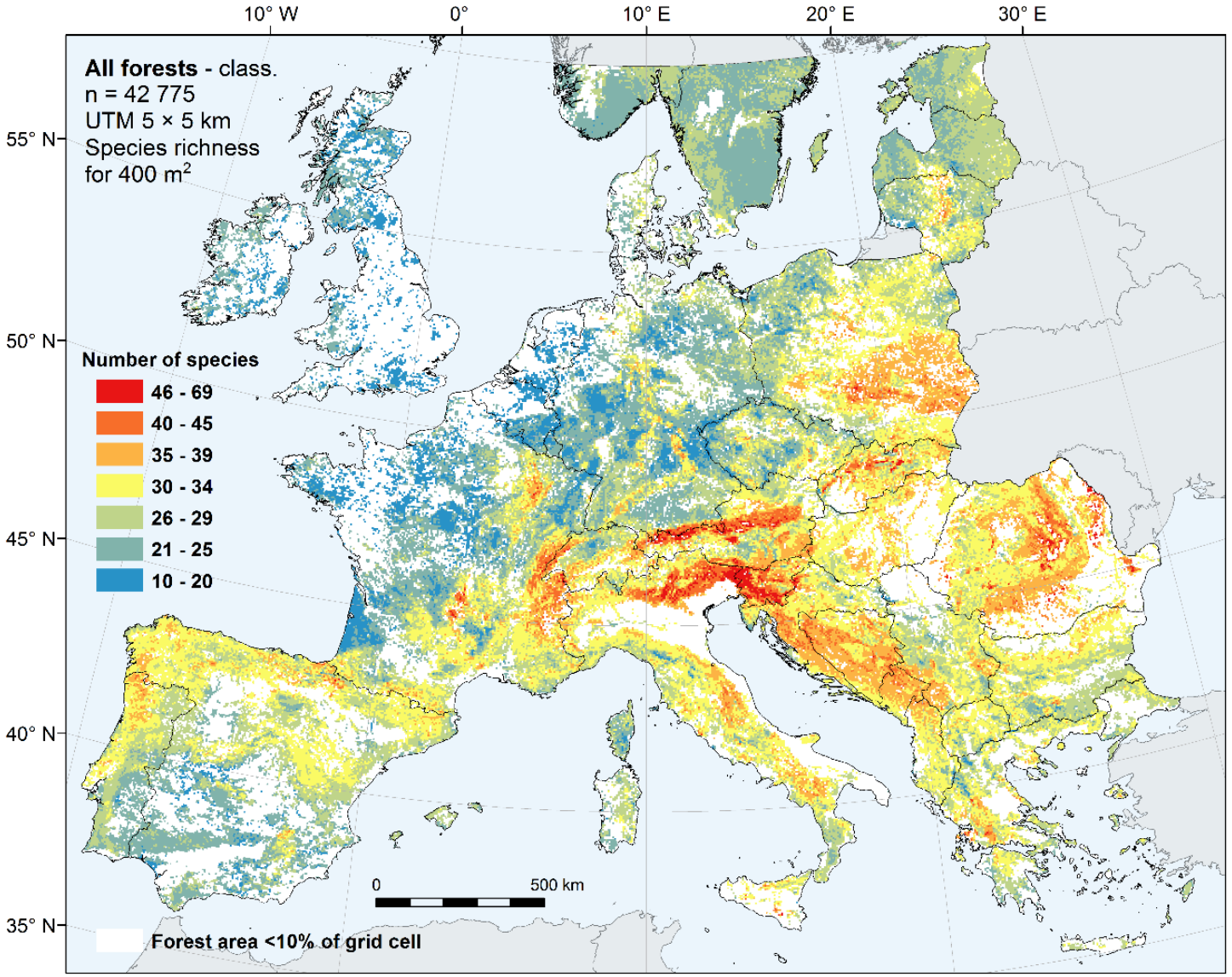


Var. explained: 46.0%



Var. explained: 34.3%

Predikce druhové bohatosti lesní vegetace Evropy



Var. explained: 44.2%

Večeřa et al. (in prep.)

Maximum entropy modeling

MaxEnt

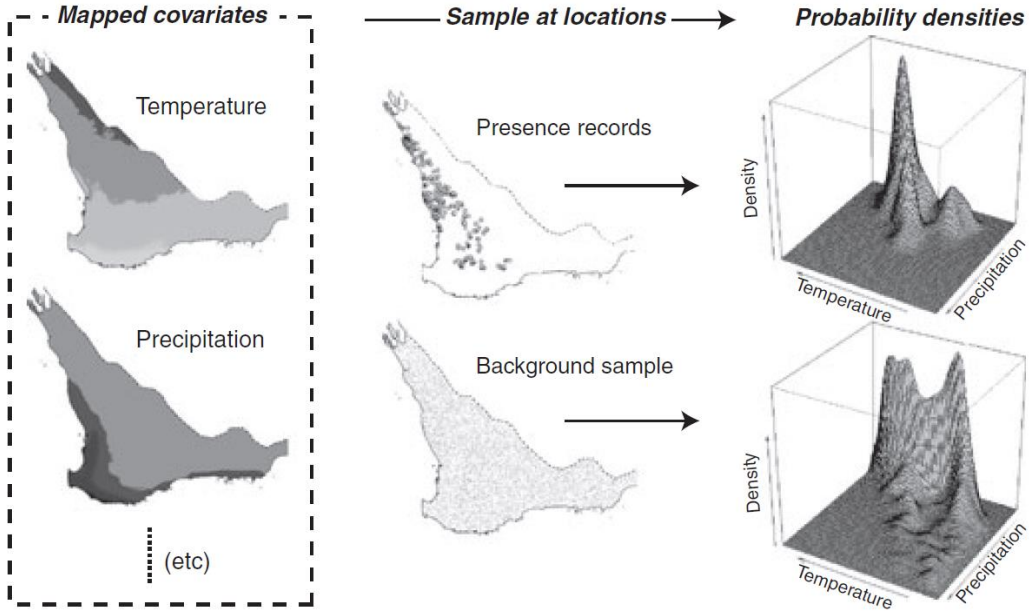
- Metoda „strojového učení“ a software uvedený v r. 2004 pro modelování **prezenčních** dat
  - „*The model minimizes the relative entropy between two probability densities (one estimated from the presence data and one, from the landscape) defined in covariate space*“ (Elith et al. 2011)
- JAVA program, R knihovna dismo, SDMtoolbox pro ArcGIS

Diversity and Distributions, (Diversity Distrib.) (2011) **17**, 43–57

BIODIVERSITY RESEARCH

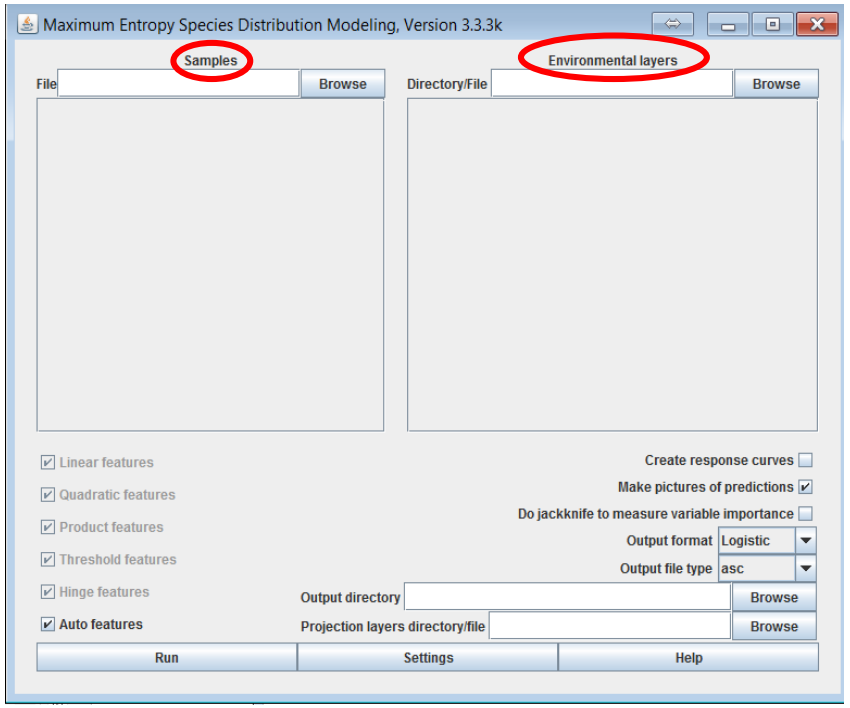
A statistical explanation of MaxEnt for ecologists

Jane Elith[1]*, Steven J. Phillips[2], Trevor Hastie[3], Miroslav Dudík[4], Yung En Chee[1] and Colin J. Yates[5]



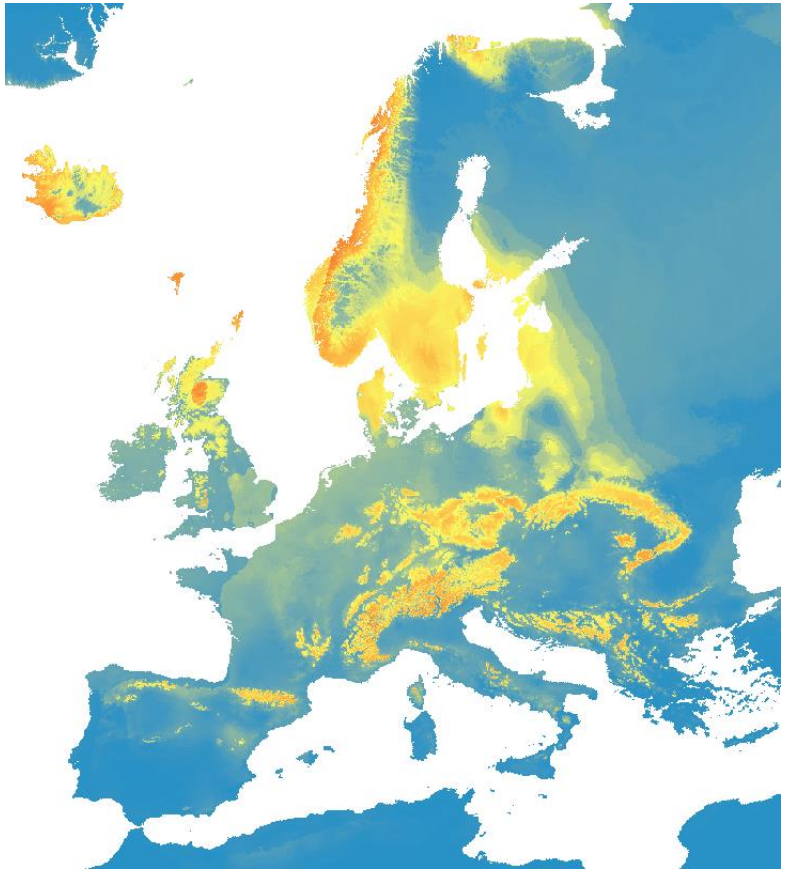
Data

- Prezenční data
  - Souřadnice X a Y lokalit výskytu druhu (csv formát)
- Absenční data
  - Využívá tzv. **background points** (pseudo-absences), které charakterizují environmentální podmínky studované oblasti
  - Background points (zpravidla 10 000) jsou náhodně vybírané z celé studované oblasti, nebo z předdefinovaného prostoru (bias file)
- Environmentální data
  - Rastrové vrstvy (ascii) ve **stejném rozlišení a rozměru**

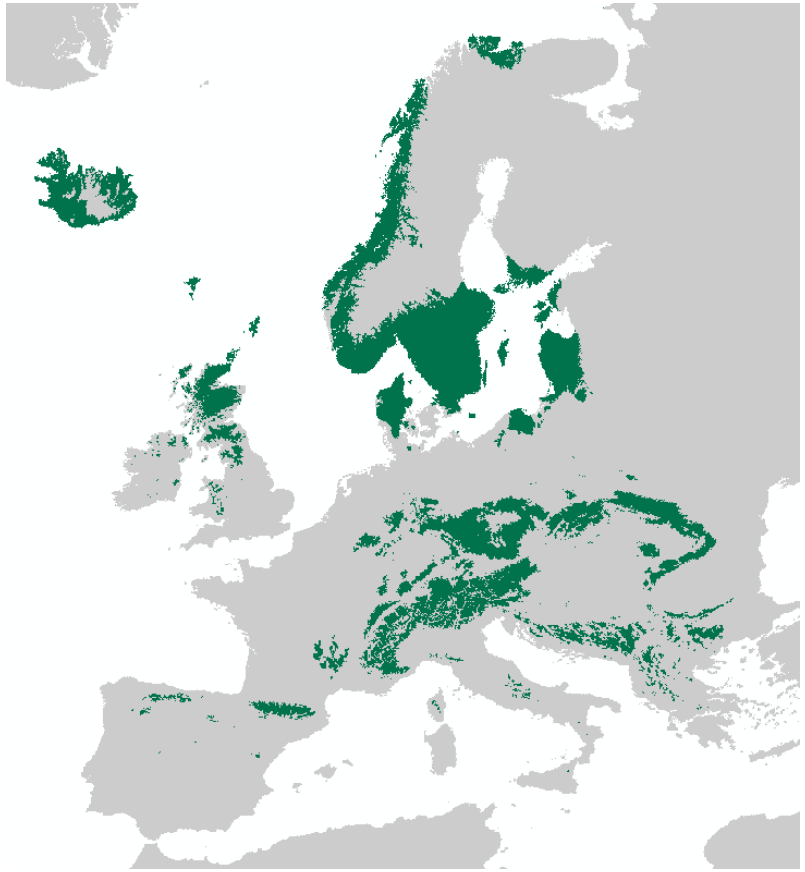


Výstup MaxEntu

Pravděpodobnostní (0…0.5….1)



Kategoriální (0/1)

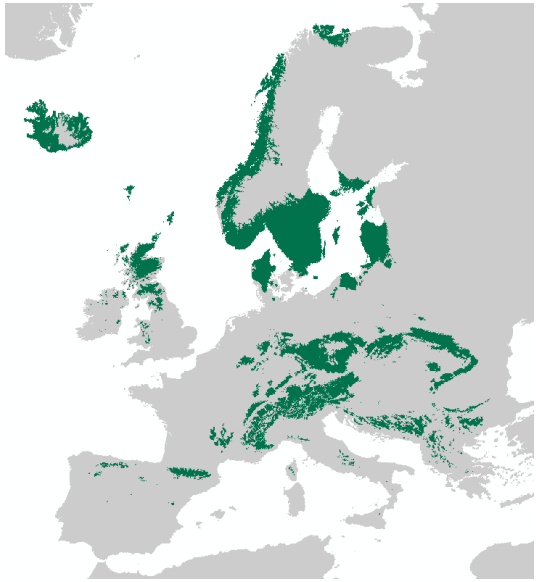


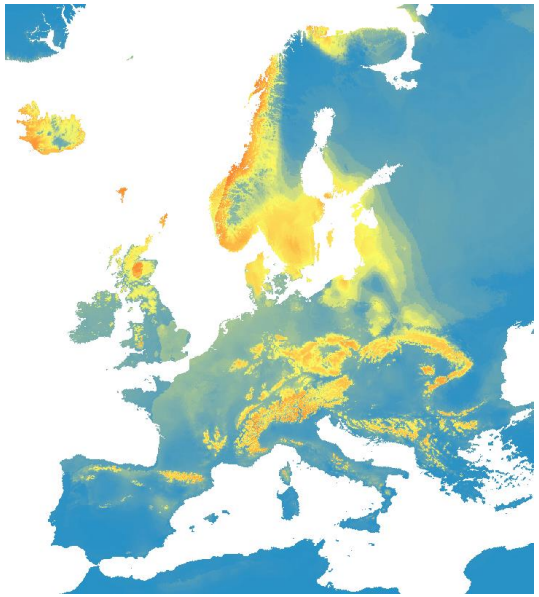
Evaluace modelu

- Pro kategoriální model
  - Threshold-dependent measures (např. Kappa)

Table 9.3. *Threshold-dependent accuracy measures for species presence–absence models based on the error matrix, where n is the total number of observations used for validation; n = TP + TN + FP + FN (Table 9.2). These accuracy measures can be calculated for any probability threshold used to define categorical predictions, except for the true skill statistic which, by definition, is based on the probability threshold for which the sum of sensitivity and specificity is maximized*

Measure	Calculation
Sensitivity	$TP/(TP+FN)$
False negative rate	1 – Sensitivity
Specificity	$TN/(TN+FP)$
False positive rate	1 – Specificity
Percent correct classification	$(TP + TN)/n$
Positive predictive power	$TP/(TP+FP)$
Odds ratio	$(TP \times TN)/(FP \times FN)$
Kappa	$\frac{[(TP+TN)-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}{[n-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}$
True skill statistic	1 – maximum (Sensitivity + Specificity)

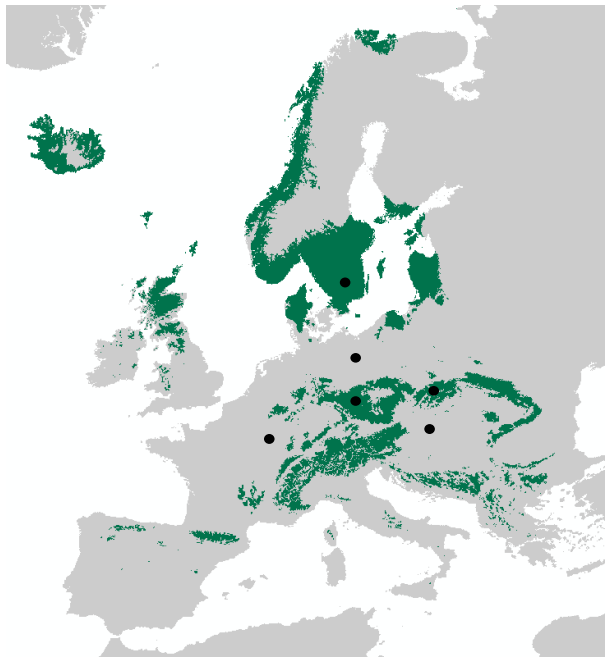




- Pro pravděpodobnostní model
  - Threshold-independent measures (např. AUC)

Evaluace kategoriálního modelu

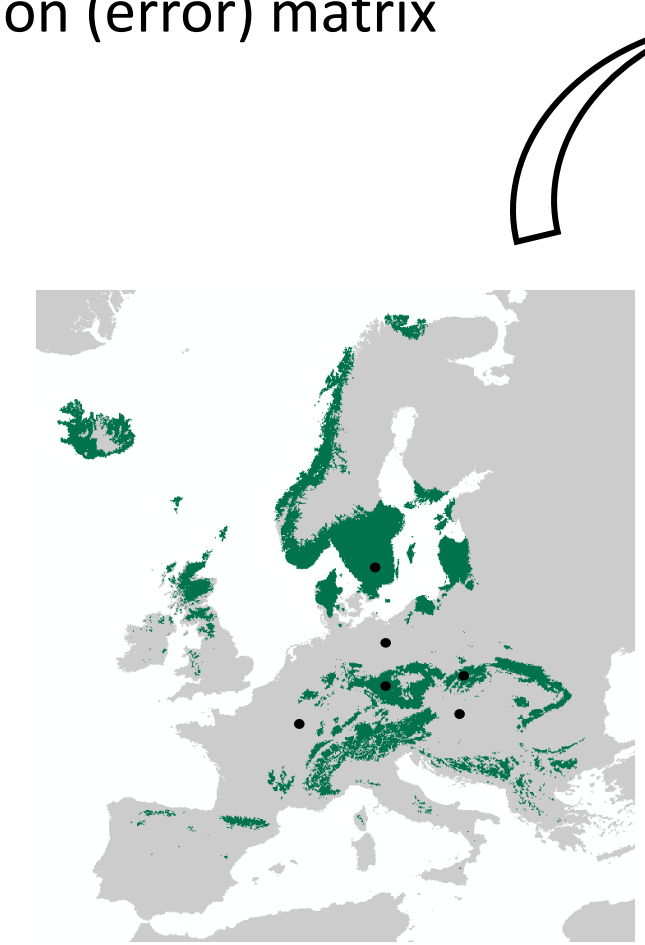
- The confusion (error) matrix

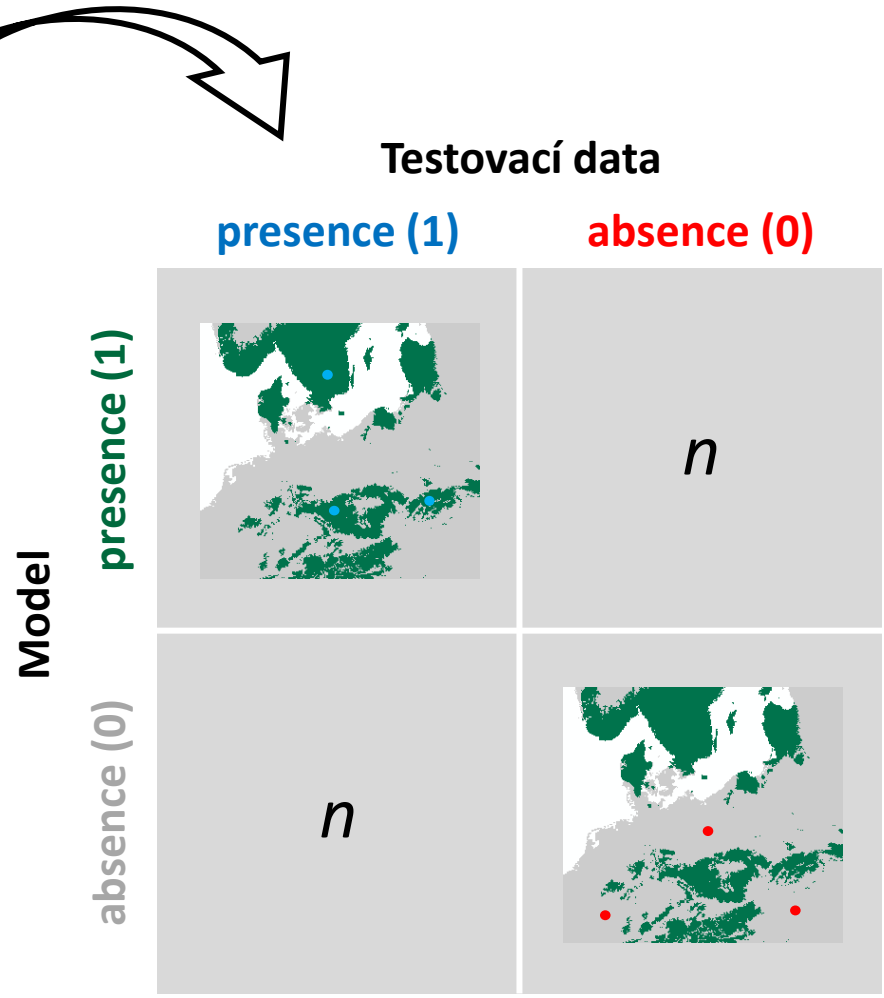


Model	Testovací data: presence (1)	Testovací data: absence (0)
presence (1)	*n*	*n*
absence (0)	*n*	*n*

Evaluace kategoriálního modelu

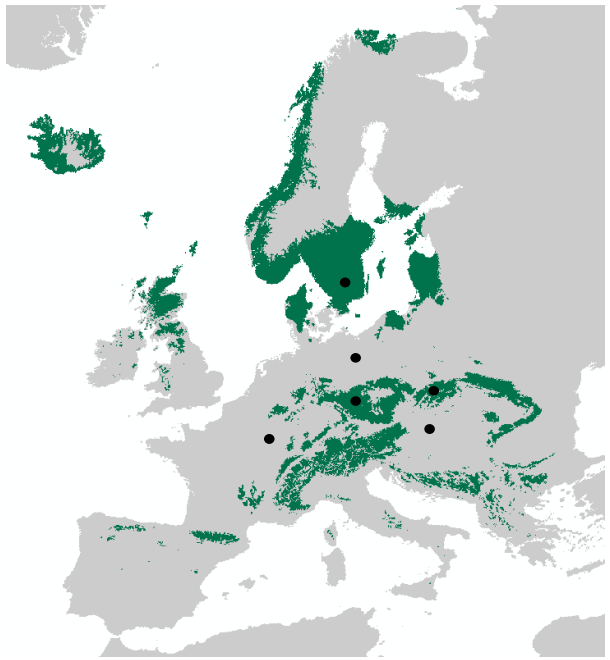
- The confusion (error) matrix

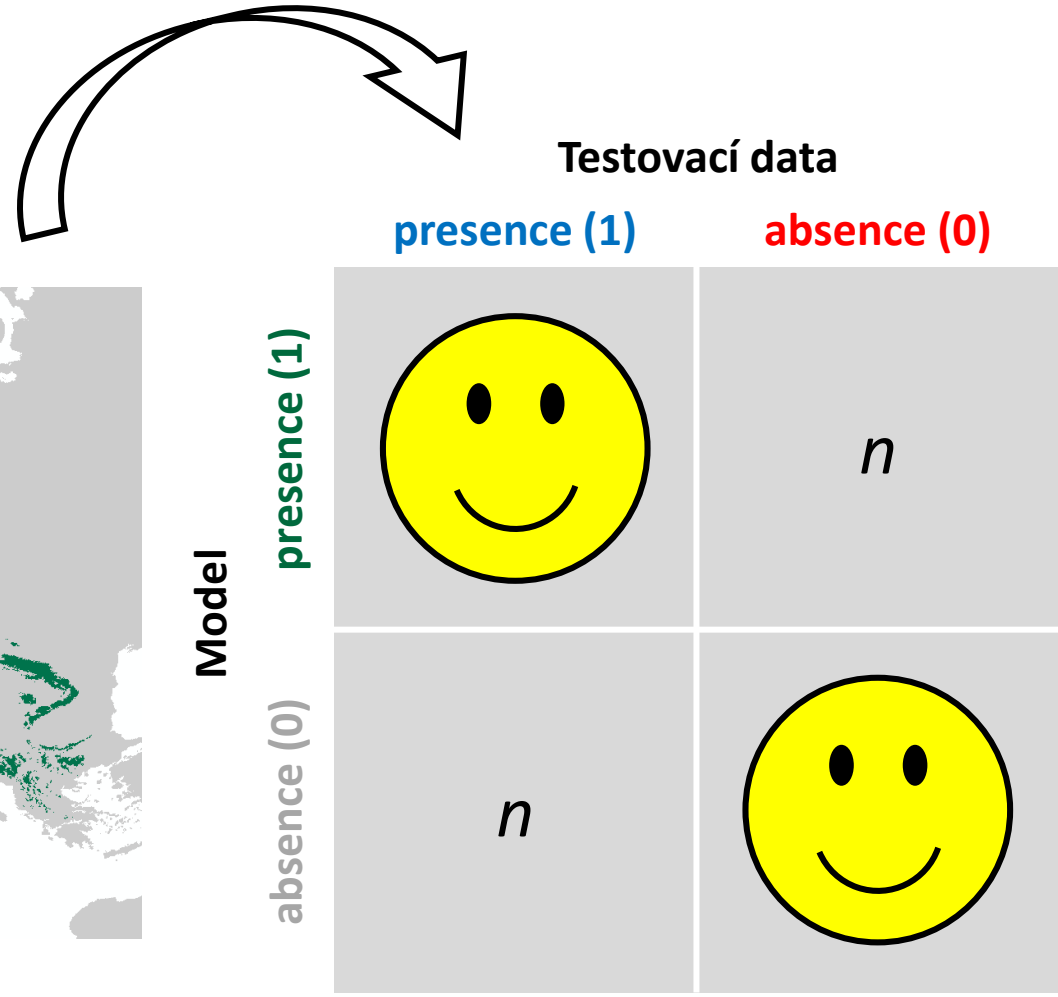




Evaluace kategoriálního modelu

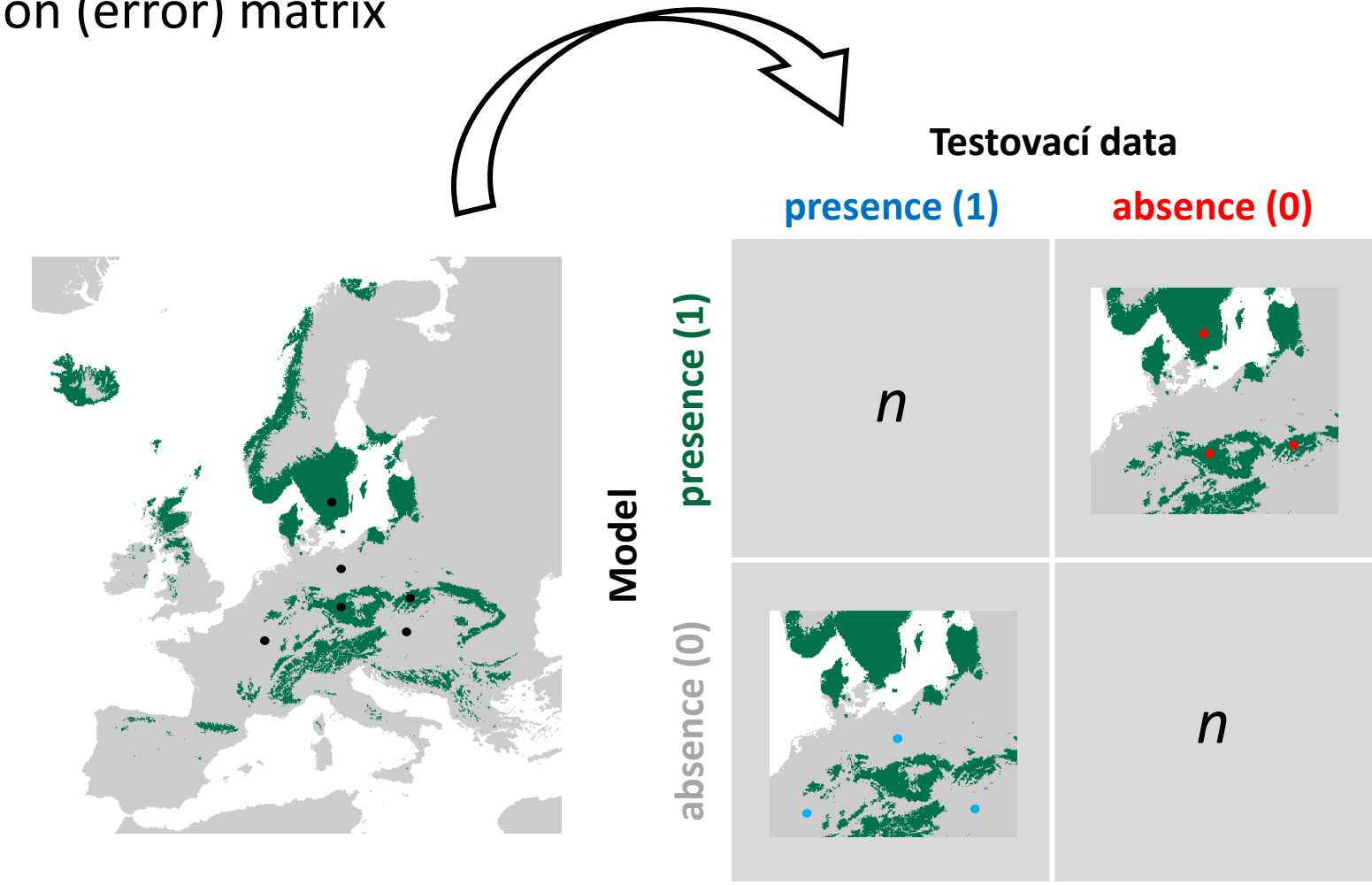
- The confusion (error) matrix





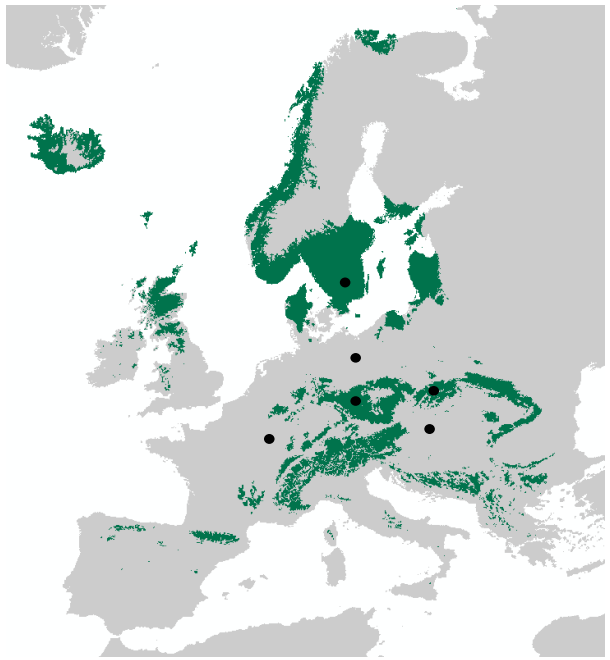
Evaluace kategoriálního modelu

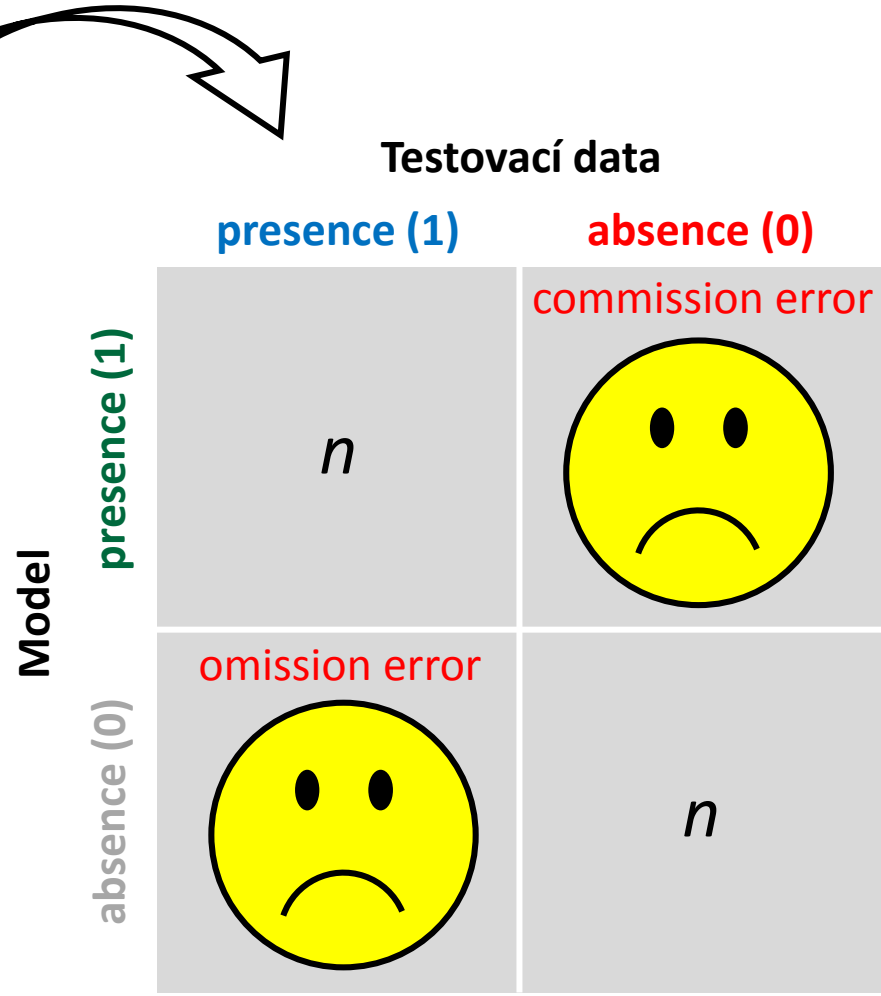
- The confusion (error) matrix



Evaluace kategoriálního modelu

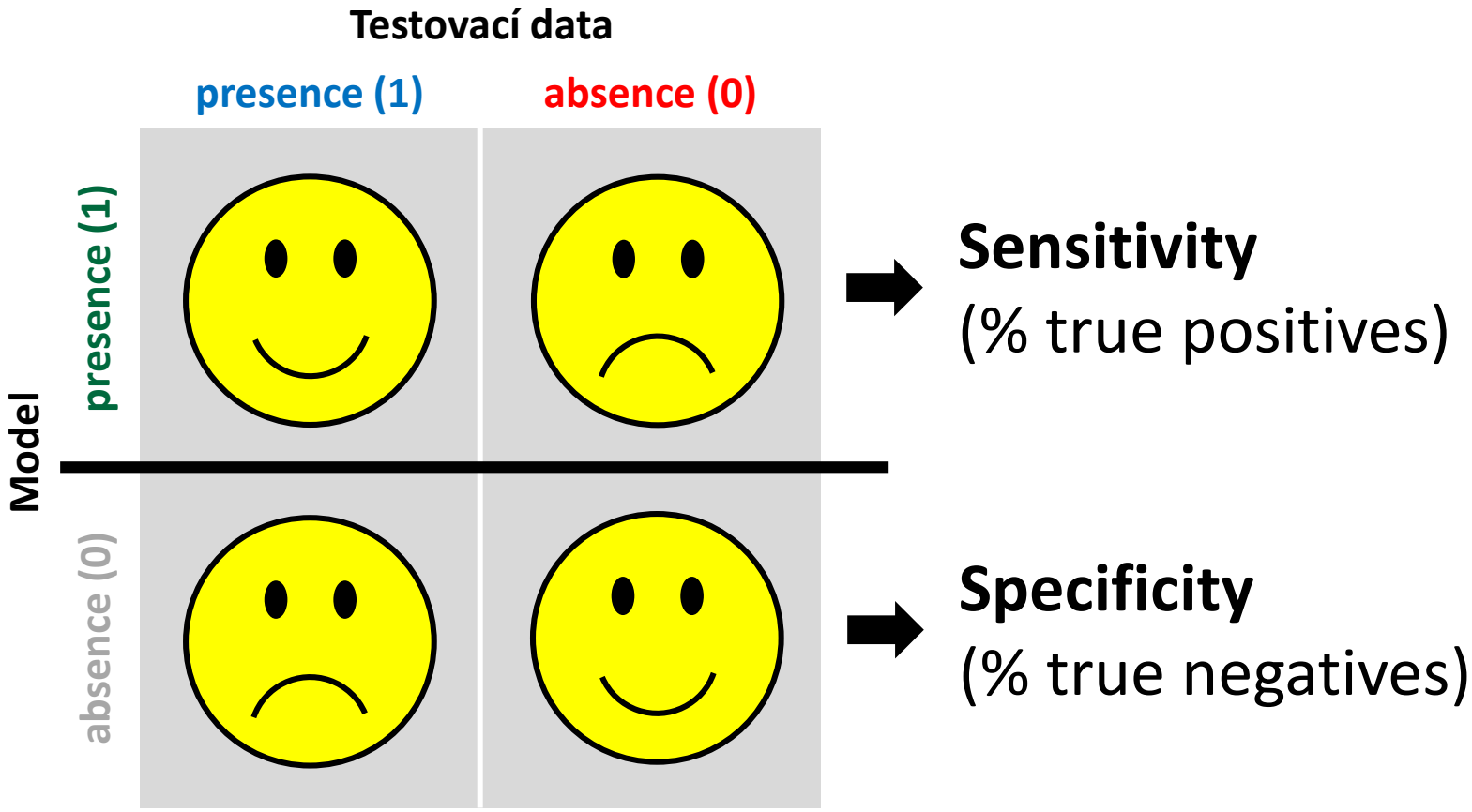
- The confusion (error) matrix





Evaluace kategoriálního modelu

- The confusion (error) matrix



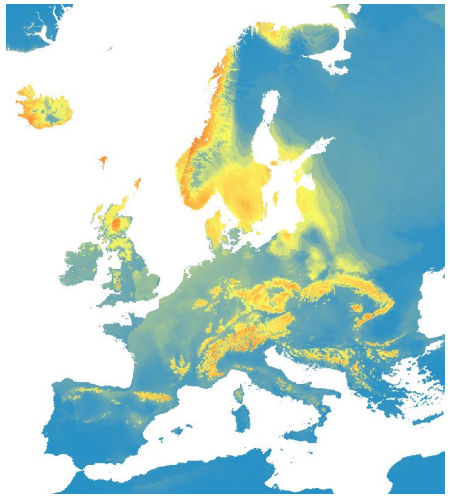
Evaluace kategoriálního modelu

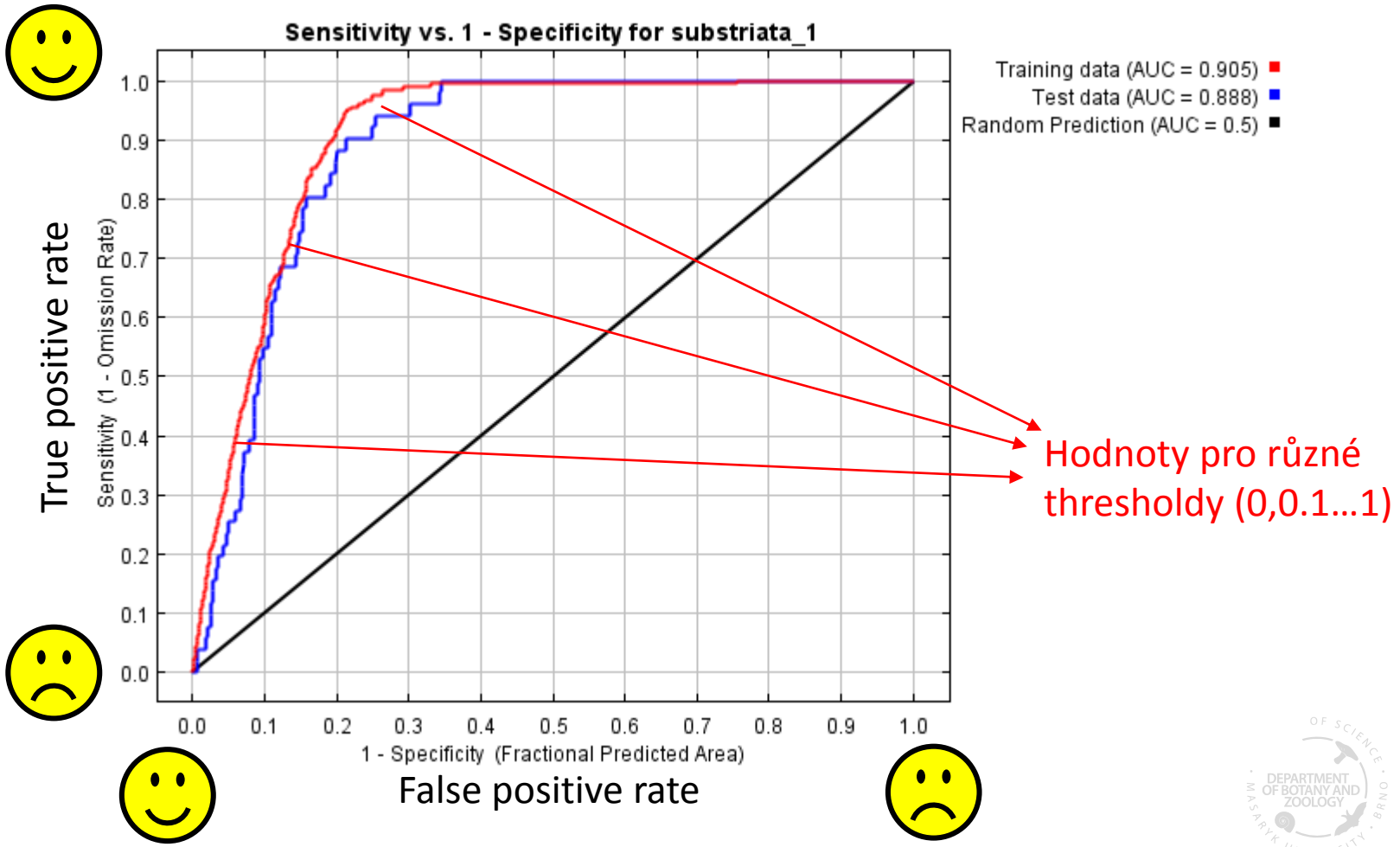
Table 9.3. *Threshold-dependent accuracy measures for species presence–absence models based on the error matrix, where n is the total number of observations used for validation; n = TP + TN + FP + FN (Table 9.2). These accuracy measures can be calculated for any probability threshold used to define categorical predictions, except for the true skill statistic which, by definition, is based on the probability threshold for which the sum of sensitivity and specificity is maximized*

Measure	Calculation
Sensitivity	$TP/(TP+FN)$
False negative rate	1 – Sensitivity
Specificity	$TN/(TN+FP)$
False positive rate	1 – Specificity
Percent correct classification	$(TP + TN)/n$
Positive predictive power	$TP/(TP+FP)$
Odds ratio	$(TP \times TN)/(FP \times FN)$
Kappa	$\frac{[(TP+TN)-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}{[n-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}$
True skill statistic	1 – maximum (Sensitivity + Specificity)

Evaluace pravděpodobnostního modelu

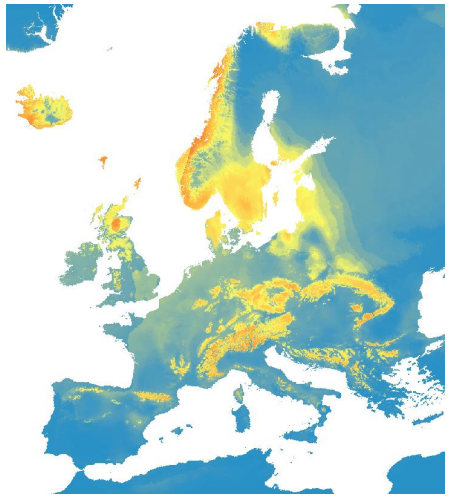
- Receiver Operating Characteristic (ROC curve)

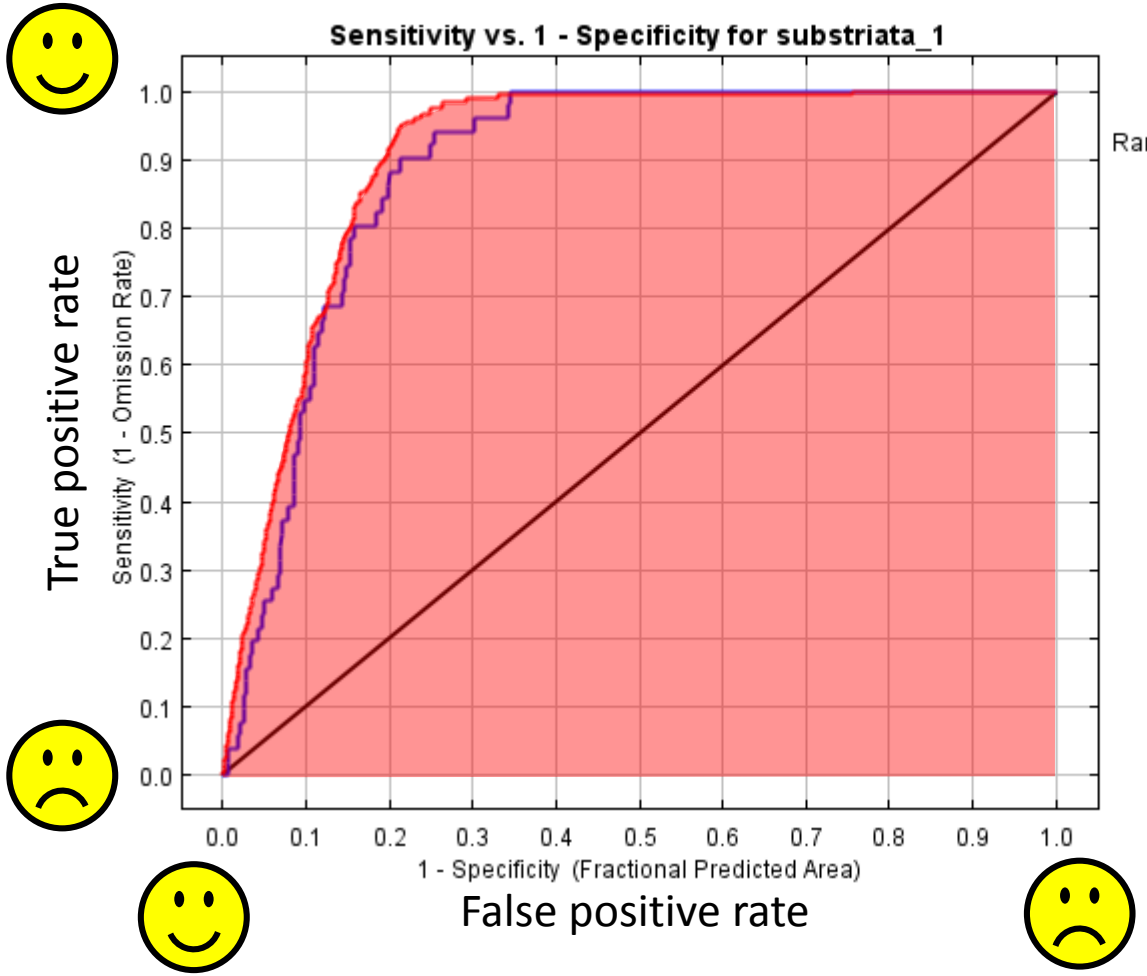




Evaluace pravděpodobnostního modelu

- Area Under ROC Curve (AUC)





Model performance (AUC values)

0.9 - 1.0: very good
0.8 - 0.9: good
0.7 - 0.8: moderate
0.6 - 0.7: low
0.5 - 0.6: very low

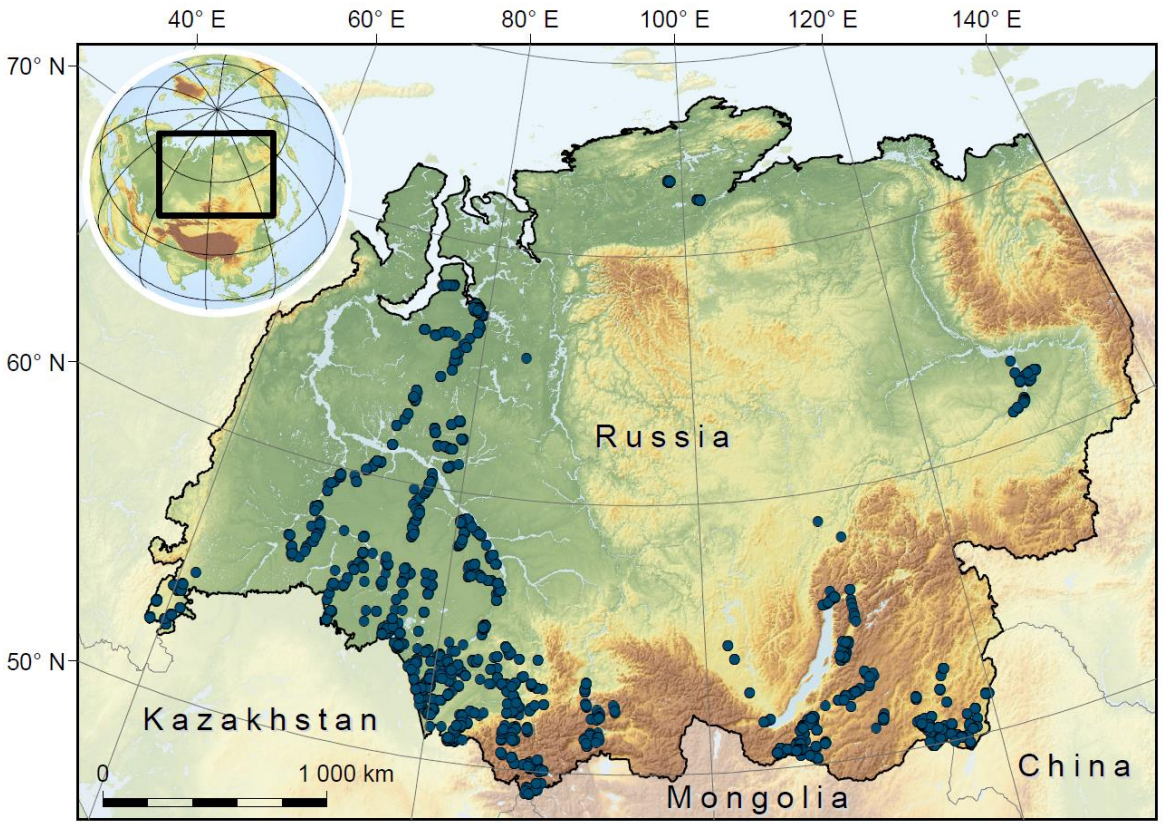
Contents lists available at ScienceDirect

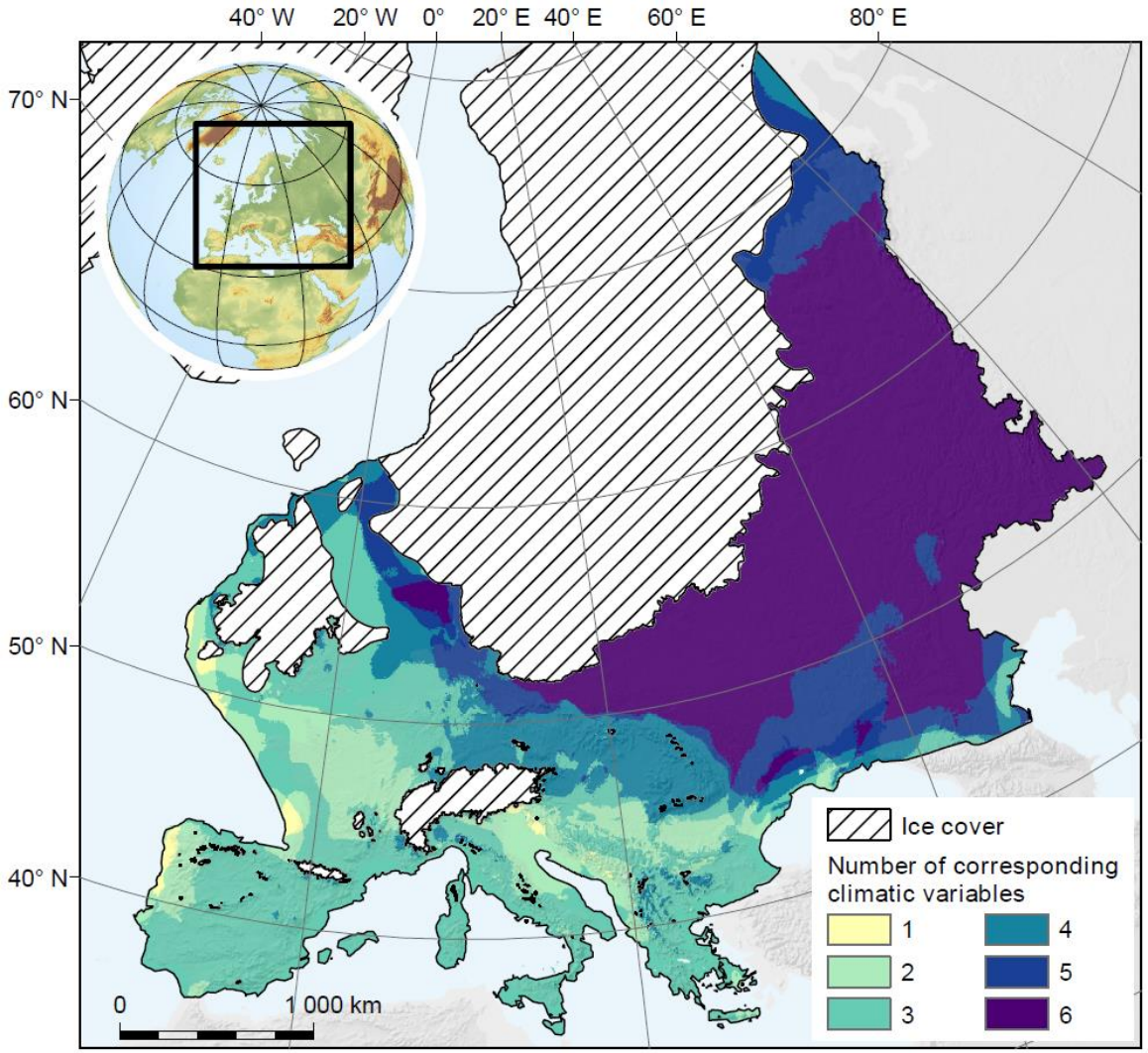
Quaternary Science Reviews

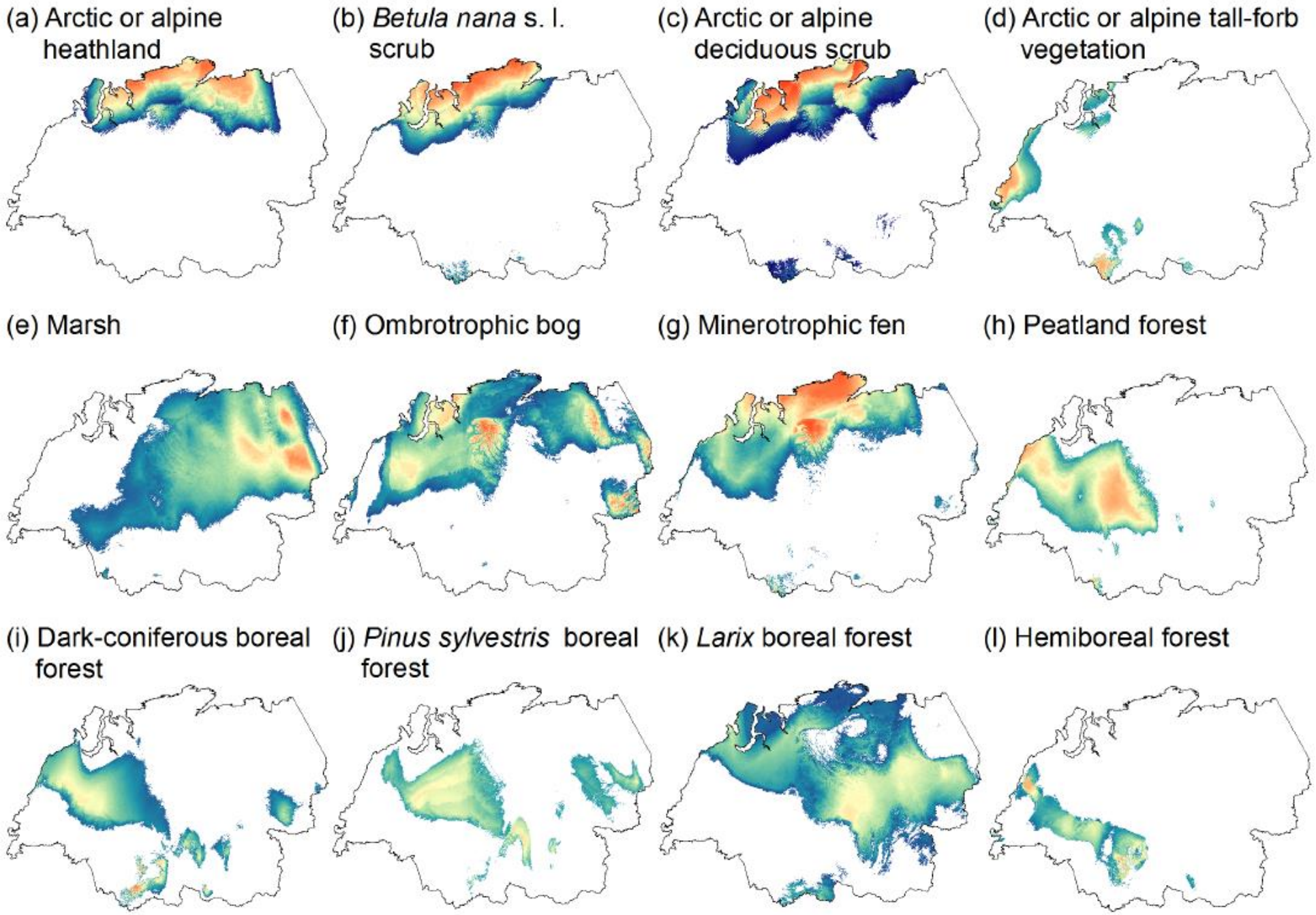
journal homepage: www.elsevier.com/locate/quascirev

Palaeodistribution modelling of European vegetation types at the Last Glacial Maximum using modern analogues from Siberia: Prospects and limitations

Veronika Janská [a], Borja Jiménez-Alfaro [b, c, d, *], Milan Chytrý [b], Jan Divíšek [a, b], Oleg Anenkhonov [e], Andrey Korolyuk [f], Nikolai Lashchinskyi [f], Martin Culek [a]






(a) Arctic or alpine heathland
(b) *Betula nana* s. l. scrub
(c) Arctic or alpine deciduous scrub
(d) Arctic or alpine tall-forb vegetation
(e) Marsh
(f) Ombrotrophic bog
(g) Minerotrophic fen
(h) Peatland forest
(i) Dark-coniferous boreal forest
(j) *Pinus sylvestris* boreal forest
(k) *Larix* boreal forest
(l) Hemiboreal forest

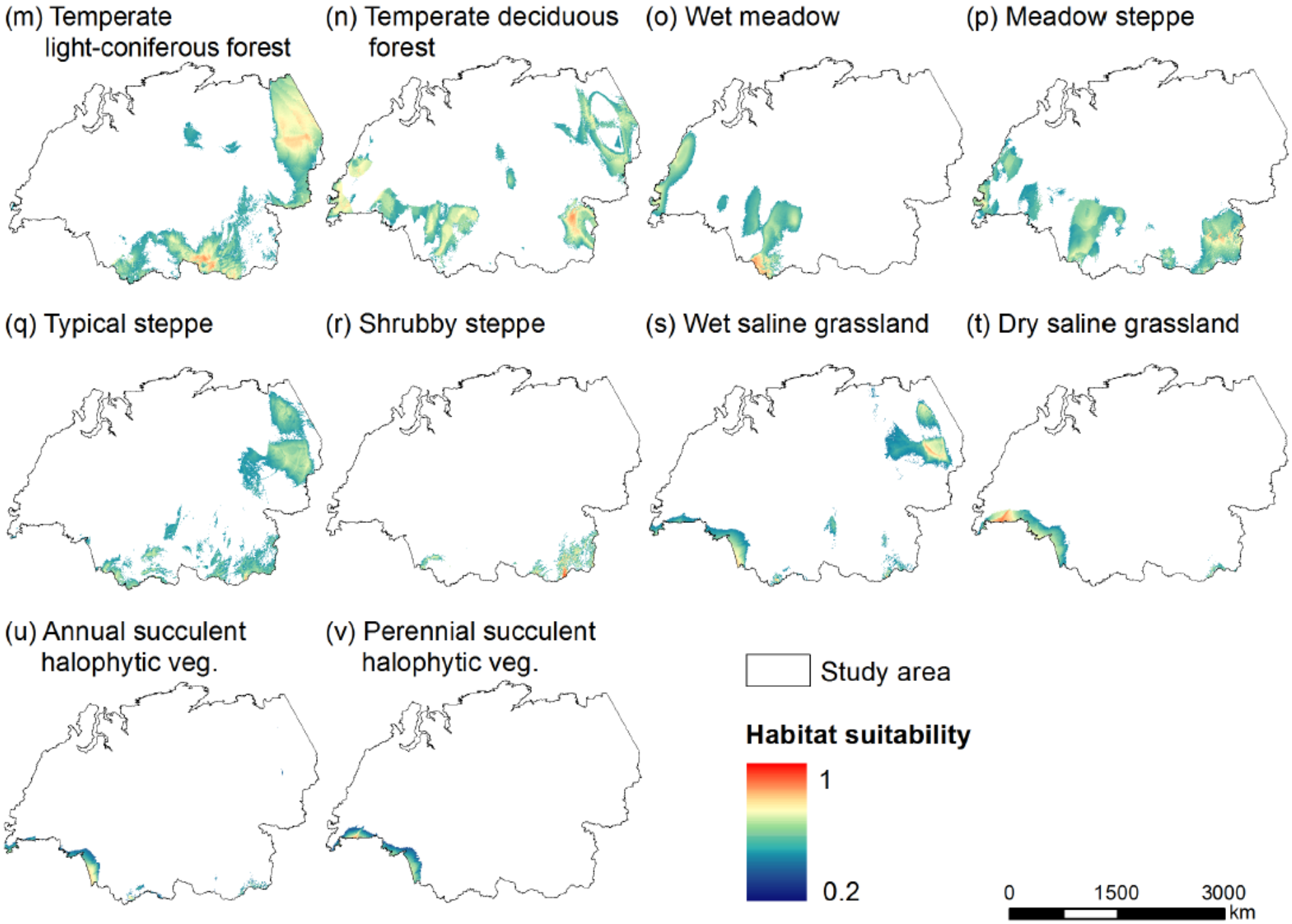
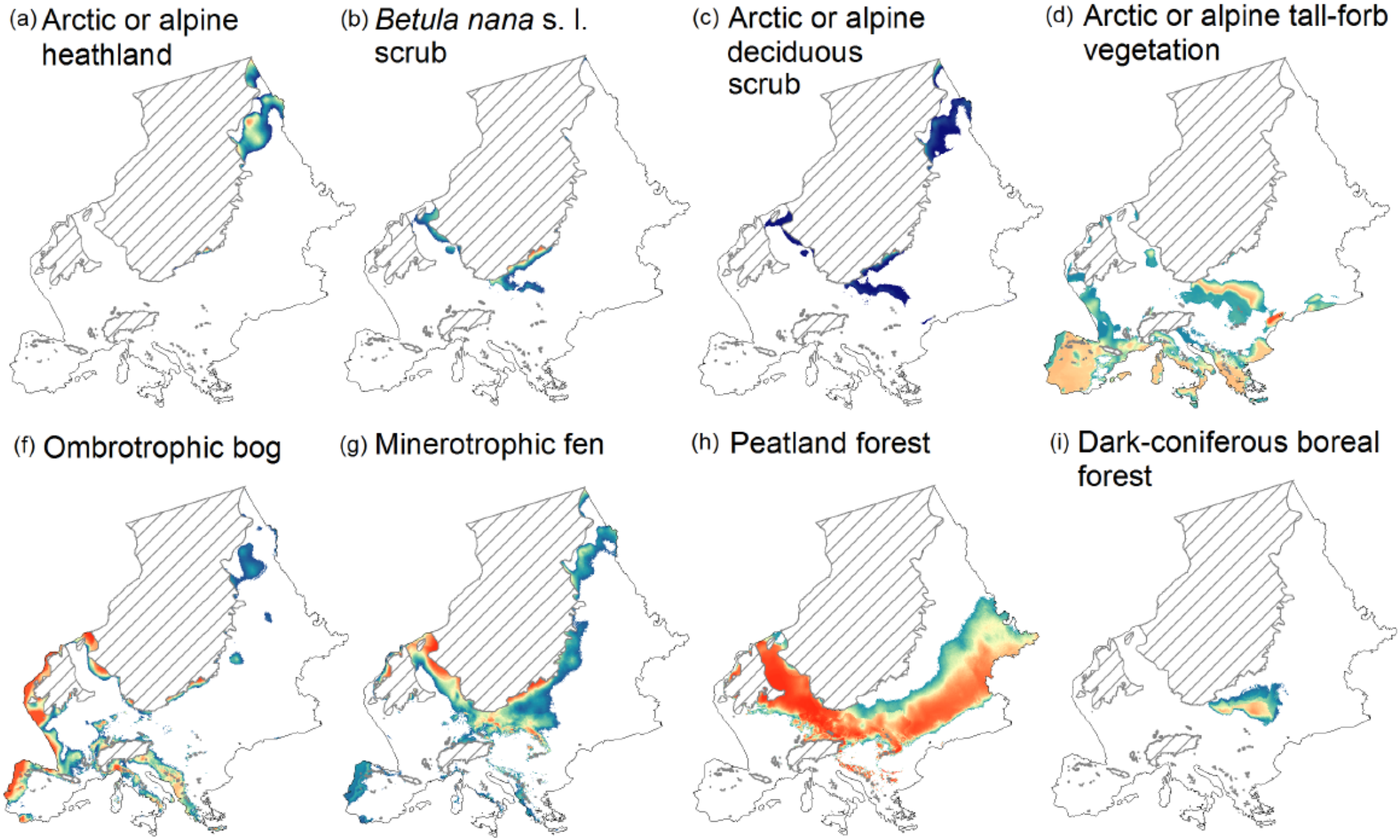
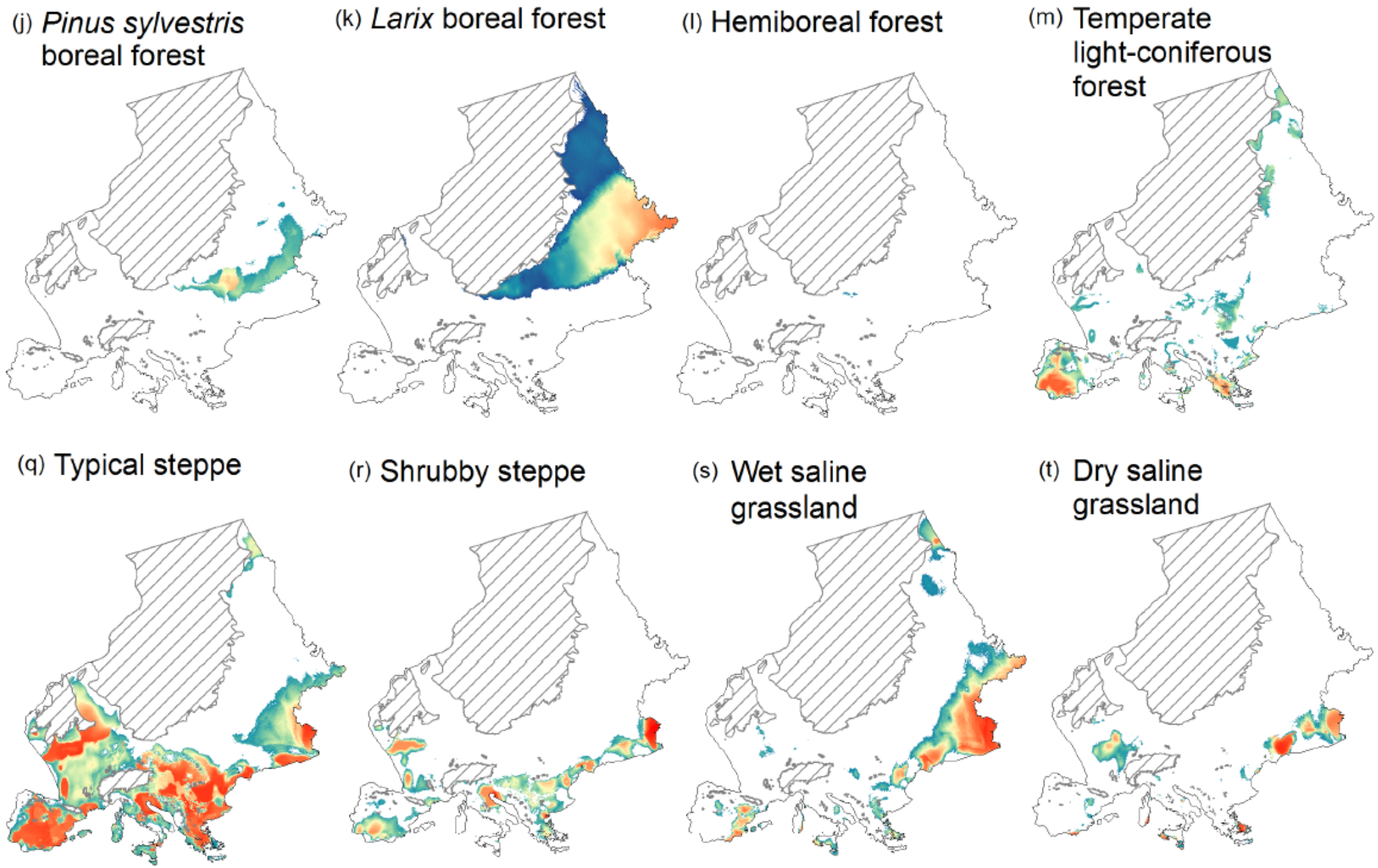


Table 3
Evaluation of distribution models computed for 22 vegetation types in present-day Siberia. The Area Under the Receiver Operating Characteristic Curve (AUC), overfitting and omission rates were calculated from MaxEnt using cross-validation. Reliability reflects the final expert evaluation of the models.

Vegetation type	AUC	Overfitting	Omission rate	Reliability
(a) Arctic or alpine heathland	0.984 ± 0.002	0.003 ± 0.003	0.040 ± 0.089	good
(b) *Betula nana* s. l. scrub	0.937 ± 0.024	0.017 ± 0.030	0.143 ± 0.143	good
(c) Arctic or alpine deciduous scrub	0.929 ± 0.063	0.026 ± 0.068	0.120 ± 0.179	good
(d) Arctic or alpine tall-forb vegetation	0.912 ± 0.031	0.043 ± 0.037	0.240 ± 0.167	moderate
(e) Marsh	0.808 ± 0.084	0.057 ± 0.117	0.367 ± 0.182	bad
(f) Ombrotrophic bog	0.893 ± 0.057	0.036 ± 0.070	0.210 ± 0.198	moderate
(g) Minerotrophic fen	0.878 ± 0.033	0.030 ± 0.047	0.300 ± 0.199	moderate
(h) Peatland forest	0.843 ± 0.038	0.035 ± 0.050	0.321 ± 0.151	moderate
(i) Dark-coniferous boreal forest	0.893 ± 0.020	0.025 ± 0.026	0.238 ± 0.079	good
(j) *Pinus sylvestris* boreal forest	0.901 ± 0.043	0.025 ± 0.056	0.180 ± 0.192	good
(k) *Larix* boreal forest	0.923 ± 0.015	0.016 ± 0.020	0.211 ± 0.139	good
(l) Hemiboreal forest	0.912 ± 0.021	0.018 ± 0.027	0.198 ± 0.085	good
(m) Temperate light-coniferous forest	0.801 ± 0.008	0.051 ± 0.015	0.284 ± 0.102	moderate
(n) Temperate deciduous forest	0.803 ± 0.036	0.075 ± 0.047	0.428 ± 0.092	bad
(o) Wet meadow	0.768 ± 0.087	0.091 ± 0.108	0.287 ± 0.256	bad
(p) Meadow steppe	0.710 ± 0.059	0.060 ± 0.080	0.412 ± 0.121	bad
(q) Typical steppe	0.752 ± 0.018	0.043 ± 0.026	0.345 ± 0.083	moderate
(r) Shrubby steppe	0.889 ± 0.030	0.032 ± 0.028	0.300 ± 0.143	moderate
(s) Wet saline grassland	0.814 ± 0.039	0.028 ± 0.054	0.262 ± 0.042	good
(t) Dry saline grassland	0.893 ± 0.025	0.019 ± 0.032	0.203 ± 0.122	good
(u) Annual succulent halophytic vegetation	0.874 ± 0.022	0.021 ± 0.031	0.318 ± 0.112	good
(v) Perennial succulent halophytic vegetation	0.939 ± 0.016	0.005 ± 0.021	0.135 ± 0.126	good


(a) Arctic or alpine heathland
(b) *Betula nana* s. l. scrub
(c) Arctic or alpine deciduous scrub
(d) Arctic or alpine tall-forb vegetation
(f) Ombrotrophic bog
(g) Minerotrophic fen
(h) Peatland forest
(i) Dark-coniferous boreal forest


(j) *Pinus sylvestris* boreal forest
(k) *Larix* boreal forest
(l) Hemiboreal forest
(m) Temperate light-coniferous forest
(q) Typical steppe
(r) Shrubby steppe
(s) Wet saline grassland
(t) Dry saline grassland

Literatura

- Legendre, P. & Legendre, L. (2012): Numerical ecology. Third Edition. Elsevier, Amsterdam.
- Borcard, D., Gillet, F. & Legendre, P. (2011): Numerical ecology with R. Springer, New York.
- Haruštiaková, D., Jarkovský, J., Littnerová, S. & Dušek, L. (2012): Vícerozměrné statistické metody v biologii. Akademické nakladatelství CERM, s.r.o., Brno.
- Komprdová, K. (2012): Rozhodovací stromy a lesy. Akademické nakladatelství CERM, s.r.o., Brno.
- http://rspatial.org/index.html
- Elith, J., S.J. Phillips, T. Hastie, M. Dudik, Y.E. Chee, C.J. Yates, 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43-57.