Latvijas Universitāte

Medicīnas un Dzīvības Zinātņu fakultāte

M. biol. Rūta Starka

19.11.2025.

Referāts

APMĀCĪBAS, VALIDĀCIJAS UN NEATKARĪGĀS TESTĒŠANAS KOPAS

Sugu izplatības modeļus izstrādā, lai noteiktu saistību starp sugas novērojumiem un vides īpašībām (Elith et al. 2011), un šis vairāku posmu process citu starpā ietver novērojumu sadalīšanu apmācības, validācijas un neatkarīgas testēšanas kopās. Šis referāts sagatavots VPP projekta "Augstas izšķirtspējas bioloģiskās daudzveidības kvantificēšana dabas saglabāšanai un apsaimniekošanai: HiQBioDiv" ietvaros, kurā sugu izplatības modelēšanai tiks izmantota maksimālās entropijas analīze (*Maximum entropy analysis,* MaxEnt), tāpēc šis referāts ir galvenokārt koncentrēts uz informāciju, kas saistoša tieši MaxEnt kontekstā.

Šī referāta uzdevums ir sniegt pārskatu par sekojošiem jautājumiem:

- Kas ir apmācības, validācijas un neatkarīgās testēšanas kopas un, kāda ir to loma sugu izplatības modelēšanā?
- Vai un kādas ir atšķirības pieejās attiecībā uz fona un klātbūtnes vietām katrā no kopām?
- Kādiem mērķiem labāk kalpo atšķirīgās validācijas kopu izveides metodes un kā tas saistās ar fona un klātbūtnes vietām?

Definīcijas:

Apmācības kopa (*model training data*) – datu daļa, ko izmanto modeļa apmācībā.

Validācijas kopa (*model validation data* vai visbiežāk ar to domāta šķērsvalidācijas kopa, *cross-validation data*) – datu daļa, ko izmanto modeļa prognozēšanas spējas pārbaudē.

Neatkarīgās testēšanas kopa (*independent test data* vai *field validation data*) – klātbūtnes dati, kas nav tikuši izmantoti modeļa apmācībā vai šķērsvalidācijā, bet tiek izmantoti modeļa prognozēšanas spējas pārbaudē.

IEVADS

Modelēšanā iekļauto klātbūtnes datu apjoms, precizitāte un ticamība

Modelēšanai pieejamās klātbūtnes datu paraugkopas lielums, telpiskā precizitāte un ticamība nosaka analīzes ierobežojumus un uz šiem datiem balstītā modeļa precizitāti un praktisko pielietojamību (Wisz et al. 2008; Erickson & Smith, 2023). Atkarībā no izmantotās modelēšanas metodes, **minimālais novērojumu skaits** sugas izplatības modelēšanai ir 5 līdz 200 klātbūtnes (Erickson & Smith, 2023). Maza paraugkopa samazina izplatības modeļu (jebkuru modeļu) precizitāti, tāpēc īpaši problemātiska ir ļoti retu sugu izplatības modelēšana (Wisz et al. 2008; Erickson & Smith, 2023), turpretī palielinoties paraugkopas izmēram, samazinās modeļa neizskaidrotā daļa (*levels of uncertanty*) (Wisz et al. 2008).

Sabiedriskās zinātnes (*citizen science*) popularitātes pieaugums ir ievērojami sekmējis datu pieejamību par sugu klātbūtni (Matutini et al. 2021). Šie dati ir nejauši, oportūnistiski, pakļauti telpiskai **piepūles novirzei** (*sampling bias*), un to **ticamība** (*reliability*) pirms izmantošanas ir pārbaudāma (Matutini et al., 2021). Ir sugas, kuru klātbūtne ir pierādāma tikai ievācot fizisku materiālu un nav nosakāmas dabā dēļ slēpta dzīvesveida vai morfoloģiskām īpašībām (Frey et el. 2013). Tāpat arī viegli nosakāmas sugas var sagādāt problēmas nepieredzējušam novērotājam (Frey et el. 2013). Tas nozīmē, ka datu filtrēšanai pirms modelēšanas ir ļoti nozīmīga loma, lai atlasītu tikai ticamus novērojumus (Aubry et al., 2017; Matutini et al., 2021). Tāpat ir būtisks jautājums par to, vai šīs klātbūtnes reprezentē sugas patieso izplatību (Hijmans & Elith, 2023). **Pēc noklusējuma, MaxEnt analīze paredz, ka informācija par klātbūtnes datiem ir nejauša**, to novietojumam proporcionāli reprezentējot populācijas blīvumu (Phillips et al. 2006; Elith et al. 2011; Merow et al. 2013). Tomēr sabiedriskās zinātnes dati nav nejauši – tie koncentrējas ap cilvēku apdzīvotām vietām, infrastruktūru (Matutini et al. 2021), tātad, tiem piemīt telpiska paraugošanas piepūles novirze (Merow et al. 2013).

Svarīgs aspekts, kas nosaka modelēšanā iekļauto datu ietekmi uz tā prognozēšanas spēju, ir arī telpiskā precizitāte (Graham et al. 2008; Aubry et al. 2017). Sabiedriskās zinātnes radīti klātbūtnes dati ir ar nezināmu punkta **novietojuma kļūdu** (*locational error*), kas var rasties no izmantotās GPS iekārtas, datu ievades neprecizitātēm un koordinātu noapaļošanas (autores piebilde). Šī telpiskā novietojuma kļūda ir mazāk ietekmīga, ja novietojums ir līdzīgā vidē kā patiesā klātbūtne, jeb klātbūtnes vietu raksturojošo vides mainīgo **telpiskajai autokorelācijai**

ir pozitīva ietekme uz modeļa prognozēšanas spēju telpiski neprecīzu klātbūtnes datu gadījumā (Naimi et al., 2011). Šī atziņa tiek izmantota lēmumu pieņemšanā par fona vietu atlasi (skat. zemāk).

Maksimālās entropijas analīze ir robusta metode, kas ir mazāk jutīga uz nelielu paraugkopu kā citas izplatības modelēšanas metodes (Phillips et al. 2006, Wisz et al. 2008), samērā labi pielietojama arī telpiski neprecīzu datu pieejamības gadījumā (Graham et al., 2008; Naimi et al., 2011; Aubry et al. 2017), un gadījumos, ja novērojumu ticamība ir samērā zema (Frey et al., 2013). Tajā pašā laikā ir būtiski ņemt vērā MaxEnt metodes pieņēmumus attiecībā uz klātbūtnes un fona datu īpašībām (cik labi tās reprezentē piemēroto vidi un fonu) un pieņemt informētus lēmumus analīzes parametru noteikšanā un modeļa rezultātu interpretācijā (Merow et al. 2013; Aubry et al. 2017).

Iztrūkumi un pseido-iztrūkumi

MaxEnt metode ir izmantojama situācijās, kad ir pieejama informācija par klātbūtni, bet trūkst informācija par tās iztrūkumu, jeb tā paredz tikai klātbūtnes (*presence-only*) datu izmantošanu (Phillips et al. 2006). Citas izplatības modelēšanas metodes (piemēram, vispārinātie lineārie modeļi, GLM, vispārinātie aditīvie modeļi, GAM) paredz arī informācijas par iztrūkumiem (*absences*) pielietošanu. Tomēr īstus, ticamus **klātbūtnes iztrūkuma datus** ir grūti iegūt, jo jebkura organisma novērošana dabā ir pakļauta nepilnīgai konstatēšanas varbūtībai (Gu & Swihart, 2004). Ja modelēšanas metode paredz iztrūkumu pielietošanu, bet šāda informācija nav pieejama, klātbūtnes iztrūkumu var pieņemt, izmantojot **pseido-iztrūkumu** (*pseudo-absences*) pieeju – iztrūkumu aizstāšanu ar fona vietām (Phillips et al. 2006). Ar pseido-iztrūkumiem mēģina raksturot to vides daļu, ko pētnieks paredz kā sugai nepiemērotu (tātad – ierobežotāka vides daļa kā fona vietas, skat. zemāk), pamatojoties uz zināmo sugas ekoloģiju vai uz šīs nepiemērotās vides paraugošanu dabā (Hijmans & Elith, 2023). Tomēr šāda pieeja ir saistīta ar daudz pieņēmumiem, līdz ar to paļaujoties tikai uz klātbūtnēm, un neizdarot pieņēmumus par fonu, var izvairīties no liekas nenoteiktības. MaxEnt modelēšanā klātbūtnes iztrūkums netiek lietots (Phillips et al. 2006; Elith et al. 2011).

Klātbūtnes un fona vietas

Klātbūtne (*presence*) it kā šķiet pašsaprotams termins – tā ir vieta, kurā noticis sugas novērojums. Tomēr vai vienmēr novērojums nozīmē klātbūtni? Neskaitot jau iepriekš aprakstīto precizitāti, ticamību un piepūles novirzi, šo klātbūtni interpretējot, ir jāatceras par sugas aktivitātes rādiusu, izplatīšanās spējām, uzvedības īpašībām, nepilnīgu konstatēšanu, dažādiem populāciju pastāvēšanas modeļiem, par novērojuma vietas ilgtermiņa piemērotību (vides pārmaiņām un piemērotības sezonalitāti) u.c., proti, lai veidotu ekoloģiski jēgpilnu modeli, ir svarīgi pārzināt sugas īpašības (autores piebilde). Svarīgi piebilst, ka MaxEnt metode paredz tieši klātbūtnes, nevis sastopamības datu izmantošanu (Phillips et al. 2006). Tas nozīmē, ka ja kādā vietā sugas populācija ir daudzskaitlīgāka, tad iespējams, šī vieta ir piemērotāka, nekā vieta, kurā nejauši novērots viens indivīds, tomēr šīs atšķirības tiešā veidā netiek ņemta vērā – informācija tiek pazaudēta, samazinot datu izšķirtspēju no sastopamības (*abundance*) uz klātbūtni (*presence*) (autores piebilde). Katrā ziņā, klātbūtnes datiem ir jāraksturo vides nosacījumus, pie kuriem ir lielāka varbūtība sugai būt sastopamai, salīdzinājumā ar pārējo aptverto vidi (Hijmans & Elith, 2023).

Fona vietas (*background data*) pēc būtības ir vides paraugs (Elith et al., 2011). Tās tiek izvēlētas, lai raksturotu vidi (kā to raksturojošo ekoģeogrāfisko mainīgo gradientu) pētījuma teritorijā, un to pretiestatītu vidi raksturojošo mainīgo vērtībām klātbūtnes vietās (Phillips et al. 2006; Merow et al. 2013). Fona vietās sugas klātbūtne ir nezināma (Merow et al. 2013). Ņemot vērā augstāk minētās nenoteiktības klātbūtnes datiem, un to ka fons netiek izmantots kā klātbūtnes iztrūkuma aizstājējs, MaxEnt modeļa rezultāts ir drīzāk **prognozētā relatīvā vides piemērotība sugai** (Phillips et al. 2006). Katra pikseļa piemērotības noteikšanā tiek ņemta vērā katra vides mainīgā vidējā vērtība šajā pikselī attiecībā pret vidējām empīriskajām vides mainīgo vērtībām klātbūtnes vietās (Phillips et al. 2006). Tādējādi lēmumi fona vietu atlasē būtiski ietekmē modeļa prognozēšanas spēju (Merow et al. 2013).

Metodes fona vietu atlasei

Fona vietu atlasei var izmantot R pakotnes “dismo” (Hijmans et al., 2024) funkciju *randomPoints*() vai arī “terra” (Hijmans 2025) funkciju *spatSample()*, tajās norādot nepieciešamo fona punktu skaitu. Fona pikseļu skaitam jābūt vismaz vienādam ar pikseļu skaitu, kuros ir klātbūtnes (Phillips et al. 2006). Pēc noklusējuma MaxEnt analīze pēc nejaušības

principa atlasa 10000 fona vietu (Elith et al. 2011). Atkarībā no vides heterogenitātes, labāka pieeja ir iekļaut pietiekami lielu **fona vietu skaitu**, lai aptvertu vidi raksturojošo mainīgo variāciju, bet ne visu vidi, tādējādi optimizējot modelēšanas laika izmaksas (Phillips & Dudik, 2008).

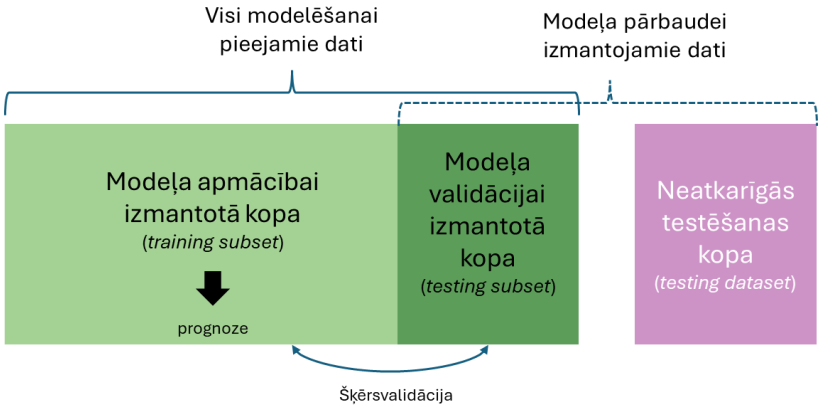
Fona vietu novietojumu var veidot pilnīgi randomizētu, nosakot tikai to atrašanās vietu pētījuma vidē, vai pakļaut dažādiem nosacījumiem (Hijmans & Elith, 2023). Nekontrolēts apmācības vai šķērsvalidācijas kopai piederīgo klātbūtnes un fona datu savstarpējais ģeogrāfiskais attālums ietekmē ar MaxEnt modeļa rezultātus, jeb rodas piederības kopai telpiskā novirze (*spatial sorting bias*, SSB) (Hijmans, 2012). Tāpēc piederības noteikšanā ir būtiski ņemt vērā pāru "klātbūne validācijas kopā – klātbūtne apmācības kopā" un "fona vieta validācijas kopā – klātbūtne apmācības kopā" savstarpējo attālumu (*point-wise distance sampling*) (Hijmans, 2012; praktisku piemēru ar R kodu skatīt Hijmans & Elith, 2023, 39. lpp). Šī piederības kopai telpiskā novirze (SSB) ir balstīta uz telpiskās autokorelācijas (*spatial autocorrelation*) pastāvēšanu vides mainīgajos (Naimi et al. 2011). Vispārīgā gadījumā, jo tuvāk divi modeļa apmācībā un validācijā izmantoti klātbūtnes punkti atrodas, jo lielāka varbūtība tiem tikt klasificētiem kā piemērotai videi, un pretēji, jo tālāk validācijā iekļauts fona punkts atrodas no apmācībā iekļauta klātbūtnes punkta, jo mazāka varbūtība tam būt klasificētam kā sugai piemērotam (Hijmans, 2012). Protams, to ietekmē arī vides heterogenitāte.

Fona vietu atlasē var arī ņemt vērā **sugas izplatīšanās spējas** (*dispersal distance*). Šī pieeja balstās idejā, ka interesējošā fona vide, kurai pretiestatīt sugas klātbūtni, ir tā, kas ir sugai sasniedzama, proti, novietojot fona punktus izplatīšanās spēju attālumā, tie sugai ir teorētiski vienlīdz pieejami (Merow et al. 2013). Papildus var ņemt vērā arī **izplatīšanās barjeras** (Elith et al. 2011). Pastāv arī pieeja, kas paredz no fona vietām izslēgt tās vietas, kas sugai vairs nav nepiemērotas (Elith et al. 2011), proti, cerams pareizi interpretējot, t.i., ņemot vērā **vides pārmaiņas** (autores piebilde). Pašas vides pārmaiņas ir iespējams modelēt, izmantojot MaxEnt metodi (Amici et al. 2017), tomēr plaša informācija par vides pārmaiņu ietekmi uz fona vietu izvēli, šķiet, nav pieejama. Cita interesanta pieeja ir veidot **mērķgrupas fona vietu atlasi** (*target-group background sampling*) (Phillips & Dudik, 2008). Šī pieeja būtībā balstās uz ideju par novērošanas piepūles novirzi, proti, par fona vietām tiek izmantotas visu pieejamo sugu klātbūtnes, kuras iespējams novērot ar līdzīgām metodēm, un tā būtiski palielina modeļa

veiktspēju (AUC vērtību) (Phillips & Dudik, 2008). Alternatīvi, pirms fona vietu izvēles var izmantot pirms tam speciāli izveidotu piepūles novirzes rastra slāni attiecīgajai mērķsugu grupai (Elith et al. 2011).Papildus, varētu veidot arī citus atlases nosacījumus, ja tie ir ekoloģiski pamatoti.

APMĀCĪBAS, VALIDĀCIJAS UN NEATKARĪGĀS TESTĒŠANAS KOPAS

Sugas izplatības (vai tai piemērotās vides) modeļa izveide, citu starpā, iekļauj pieejamo datu sadalīšanu apmācības un validācijas kopās, un lēmumu pieņemšanu par neatkarīgās testēšanas kopas avotu (1. attēls).

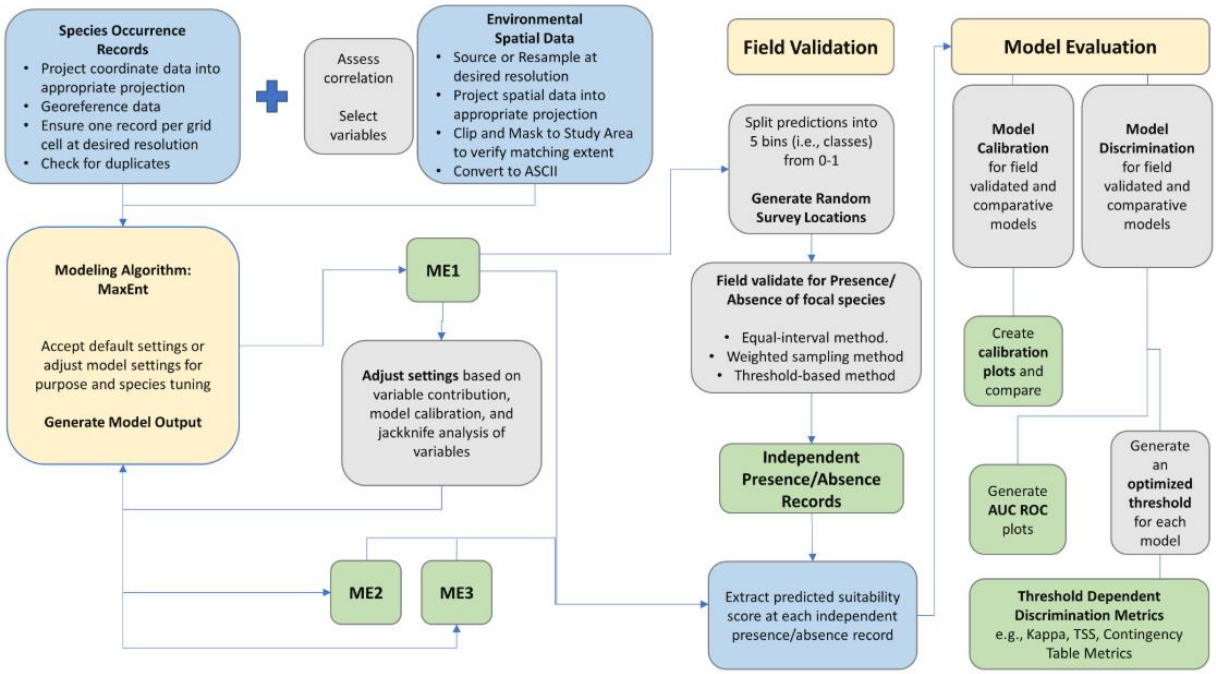


1. attēls. Apmācības, validācijas un neatkarīgās testēšanas kopu savstarpējā saistība sugu izplatības modeļu kontekstā (autores interpretācija).

Modeļa **apmācības kopa** (*training dataset* vai *training subset*) ir datu daļa, ko izmanto modeļa apmācībā (Hijmans 2012, Hijmans & Elith, 2023). Tā parasti ir lielākā daļa (70% – 90%) no visiem modelēšanai pieejamiem datiem. Tāpēc ir būtiski, lai modeļa apmācības daļā būtu tik daudz fona vietu, lai aptvertu tās variāciju, un tā izvēlētas, lai reprezentētu sugai pieejamo vidi (Phillips & Dudik, 2008).

Validācijas kopa (*testing dataset*) plašā izpratnē ir dati, kas nav izmantoti modeļa apmācībai, bet ko izmanto modeļa prognožu ticamības pārbaudei (Phillips et al. 2006; Hijmans, 2012). Validācijas dati var tikt iegūti divos veidos – kā **šķērsvalidācijas** dati (*cross-validation*), vai kā neatkarīgas testēšanas dati. Šķērsvalidācijas dati ir tā modelēšanai pieejamo datu daļa, kas netika izmantota modeļa apmācībā, respektīvi, parasto atlikušie 10 – 30% (1. attēls).Tā ir populārākā

pieeja modeļa pārbaudei, jo neparedz papildus klātbūtnes datu iegūšanu. **Neatkarīgas testēšanas kopa** ir klātbūtnes dati, kas teorētiski ir iegūti neatkarīgi no modeļa apmācībā izmantotajiem klātbūtnes datiem, un kas tiek izmantoti modeļa validācijā (Hijmans 2012). Proti, ideālā gadījumā, neatkarīgā testēšanā ietilpst klātbūtnes pārbaudīšana dabā, vietās, kur modelis ir prognozējis sugai piemērotu vidi (Johanson et al. 2023, 2. attēls).

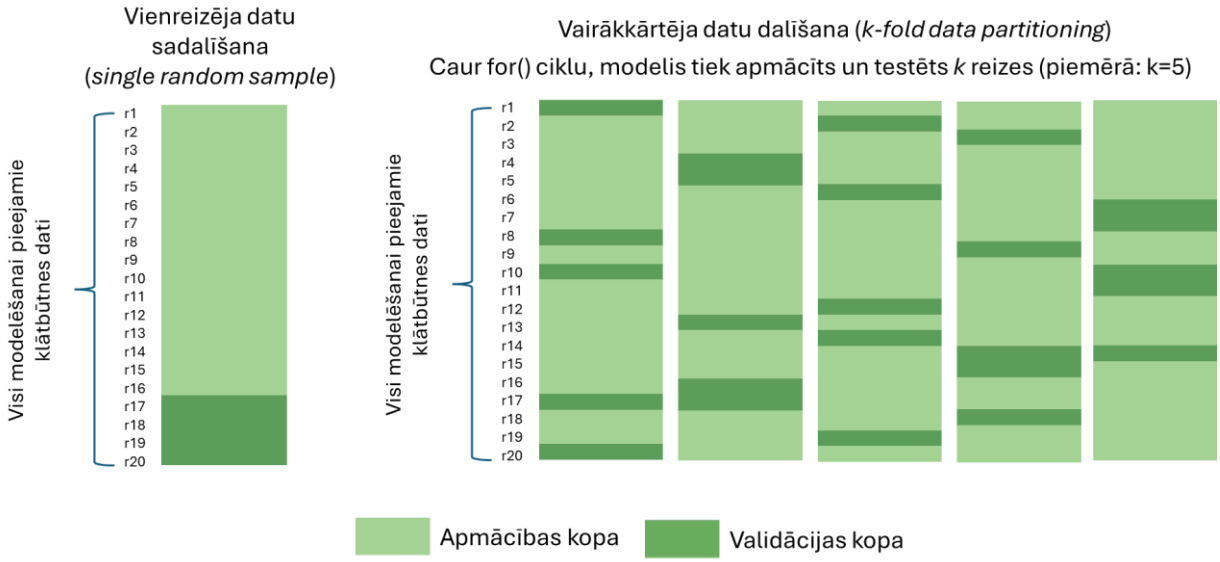


2. attēls – Darba plūsmas piemērs MaxEnt analīzes izmantošanai sugas izplatības modelēšanā, kas iekļauj gan šķērsvalidāciju, gan neatkarīgas testēšanas kopu (Johanson et al. 2023).

Visbiežāk, ja klātbūtnes dati pirms modelēšanas pakļauti stingrai to kvalitātes filtrēšanai, tad neatkarīga testēšana dabā nav nepieciešama (Matutini et al. 2021). Tomēr lai arī kuru pieeju izvēlētos (šķērsvalidāciju vai neatkarīgu testēšanu, vai abas) – validācija kā tāda ir obligāta, lai izvairītos no pārlieku pozitīvām, jeb datiem pielāgotām modeļa piemērotības metrikām (*model accuracy estimates*) (Newbold et al. 2010, Matutini et al. 2021). Ņemot vērā, ka MaxEnt ir tendence pielāgoties ievades datiem (*over-fitting*), svarīga ir randomizācija apmācības un validācijas kopu izveidē (Phillips et al. 2006), šīm kopām piederošo vietu savstarpējais telpiskais novietojums (Hijmans, 2012), kā arī modeļa parametru regularizācija (Phillips & Dudik 2008).

Metodes apmācības un validācijas kopu izveidei

Visvienkāršākais veids ir datu sadalīšanai apmācības un validācijas kopās ir vienreizēja nejauša sadalīšana (*single random sample*) (Hijmans & Elith, 2023), piemēram, 100 klātbūtnes punktu gadījumā pirmās 70 rindiņas izmantojot modeļa apmācībai, bet pēdējās 30 rindiņas – validācijai. Tomēr skaidrs, ka šādas vienas atlases veidošana jau pēc būtības paredz nestabilu rezultātu atkarībā no tā, tieši kuri klātbūtnes dati (kontekstā ar ievadā minētajiem apsvērumiem tam, cik labi klātbūtnes dati reprezentē piemēroto vidi) tiek izmantoti modeļa apmācībai un kura – modeļa šķērsvalidēšanai (autores piebilde). Tāpēc biežāk pielietotā pieeja ir vairākkārtēja datu dalīšana (*k-fold data partitioning*), kas paredz klātbūtnes datu sadalīšanu apmācības un validācijas kopās vairākas (*k*) reizes (Hijmans, 2012; Hijmans & Elith, 2023). Sākumā tiek izveidots vektors randomizētai rindas piederībai apmācības vai validācijas kopai, un tad attiecīgi *k* reizes dati tiek atlasīti modeļa izveidei un pārbaudei (3. attēls). Palielinot apmācības:testēšanas kopu izveides randomizācijas notikumu skaitu, tiek iegūta precīzāka informācija par uz konkrētajiem klātbūtnes datiem balstīto modeļu performances variāciju (Phillips et al. 2006).



3. attēls. Pamata atšķirības starp vienreizēju klātbūtnes datu sadalīšanu modeļa apmācības un validācijas kopā (pa kreisi) un vairākkārtēju randomizētu datu dalīšanu (pa labi). Ar r1–r20 apzīmētas teorētiskas datu rindas, kurās glabājas sugas klātbūtnes dati. Kopējais klātbūtņu skaits ir 20, no kuriem abos gadījumos 75% datu izmantoti modeļa trenēšanai, bet 25% - validācijai (autores interpretācija).

Fona un klātbūtnes vietu izvēles nozīme validācijas kopā

Izvērtējot to, cik korekti modelis prognozē vides piemērotību sugas klātbūtnei, ir jānosaka validācijas kopai piederošo klātbūtnes datu daļu, kas ietilpst vidē, ko modelis prognozē kā sugai nepiemērotu (*extrinsic omission rate*), jeb **izslēgšanas kļūdu** (*omission error*) (Phillips et al. 2006). Tāpat jāņem vērā **piemērotības kļūda** (*comission error*), jeb piemērotās vides daļa, ko modelis prognozējis kā nepiemērotu (Phillips et al. 2006). Bez modeļa apmācībā izmantoto datu apjoma šīs kļūdas ietekmē arī vairāki citi faktori, tajā skaitā pašu klātbūtnes un fona vietu novietojums un piederība apmācības vai validācijas kopai.

Lai gan sugu izplatības modeļu fundamentālais mērķis ir pati prognoze, nevis hipotēžu pārbaude (Hijmans, 2012), kā jebkurai statistiskai metodei, tai ir pieņemta nulles hipotēze. MaxEnt metodes **nulles hipotēze** paredz, ka sugas sastopamība vidē ir nejauša, un modelis prognozē vides piemērotību sugai tāpat kā nejaušība (Phillips et al. 2006; Hijmans 2012). Tātad, ja informācija par sugas klātbūtni jau sākotnēji nav nejauša (ir pakļauta novirzei), tad **pilnīgi randomizēta fona vietu atlase** nav obligāti labākā pieeja (Phillips & Dudik, 2008).

Lai līdz galam spriestu par fona un klātbūtnes vietu izvēles ietekmi uz modeļa veiktspēju (varbūtību, ka pikseļa piemērotība ir pareizi klasificēta), ir jāapskata arī modeļu piemērotības metrikas (*area under the curve,* AUC, un *receiver operating characteristic* ROC analīze), kuru aprēķinos tiek izmantota gan informācija no klātbūtnes, gan fona vietām. Šīs metrikas ir aprakstītas citā referātā.

Izmantotie informācijas avoti

Amici, V., Marcantonio, M., la Porta, N., & Rocchini, D. (2017). A multi-temporal approach in MaxEnt modelling: A new frontier for land use/land cover change detection. Ecological Informatics, 40, 40–49. https://doi.org/10.1016/j.ecoinf.2017.04.005

Aubry K.B., Raley C.M., McKelvey K.S. (2017). The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. PLoS ONE, 12(6): e0179152 https://doi.org/10.1371/journal.pone.0179152

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

Erickson, K.D. and Smith, A.B. (2023). Modeling the rarest of the rare: a comparison between multi-species distribution models, ensembles of small models, and single-species models at extremely low sample sizes. Ecography: e06500. https://doi.org/10.1111/ecog.06500

Frey, J. K., Lewis, J. C., Guy, R. K., & Stuart, J. N. (2013). Use of anecdotal occurrence data in species distribution models: An example based on the white-nosed coati (Nasua narica) in the American southwest. Animals, 3(2), 327–348. https://doi.org/10.3390/ani3020327

Graham, C.H., Elith, J., Hijmans, R.J., Guisan, A., Townsend Peterson, A., Loiselle, B.A. and The Nceas Predicting Species Distributions Working Group (2008). The influence of spatial errors in species occurrence data used in distribution models. Journal of Applied Ecology, 45: 239-247. https://doi.org/10.1111/j.1365-2664.2007.01408.x

Gu, W., & Swihart, R. K. (2004). Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. Biological Conservation, 116(2), 195–203. https://doi.org/10.1016/S0006-3207(03)00190-3

Hijmans, R.J., (2012). Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null-model. Ecology 93: 679-688. https://doi.org/10.1890/11-0826.1

Hijmans R.J., Elith J. (2023). Species Distribution Models. 96 lpp. https://rspatial.org/raster/sdm/raster_SDM.pdf

Hijmans R., Phillips S., Leathwick J., Elith J. (2024). dismo: Species Distribution Modeling. https://doi.org/10.32614/CRAN.package.dismo, R package version 1.3-16.

Hijmans R (2025). terra: Spatial Data Analysis. https://doi.org/10.32614/CRAN.package.terra, R package version 1.8-70.

Matutini, F., Baudry, J., Pain, G., Sineau, M., & Pithon, J. (2021). How citizen science could improve species distribution models and their independent assessment. Ecology and Evolution, 11(7), 3028–3039. https://doi.org/10.1002/ece3.7210

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. Ecography, 36(10), 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Naimi, B., Skidmore, A. K., Groen, T. A., & Hamm, N. A. S. (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. https://doi.org/10.1111/j.1365-2699.2011.02523.x

Newbold, T., Reader, T., El-Gabbas, A., Theisinger, W., Shohdi, W.M., Zalat, S., Din, S., Gilbert, F. (2010). Testing the accuracy of species distribution models using species records from a new field survey. Oikos, 119: 1326–1334. https://doi.org/10.1111/j.1600-0706.2009.18295.x

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190, 231–259.

Phillips, S.J. and Dudík, M. (2008), Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31: 161-175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., Elith, J., Dudík, M., Ferrier, S., Huettmann, F., Leathwick, J. R., Lehmann, A., Lohmann, L., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M. C., … Zimmermann, N. E. (2008). Effects of sample size on the performance of species distribution models. Diversity and Distributions, 14(5), 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x