Izmantoto saīsinājumu saraksts:

AUC – Laukums zem ROC līknes (Area Under the ROC Curve);

CAUC – Kalibrētais AUC (Calibrated AUC);

EGV – Ekoģeogrāfiskais mainīgais (Ecogeographical Variable);

FN – Nepatiesi negatīvi (…) (False Negatives);

FP – Nepatiesi pozitīvi (…) (False Positives);

FPR – Nepatiesi pozitīvo rādītājs (False Positive Rate);

OR – Izlaiduma kļūdu īpatsvars (Omission Error Rate);

ROC – ROC līkne (Receiver Operating Characteristic);

SDM – Sugu izplatības modelēšana (Species Distribution Modeling);

TN – Patiesi negatīvi (…) (True Negatives);

TP – Patiesi pozitīvi (…) (True Positives);

TPR – Patiesi pozitīvo rādītājs / jutīgums (True Positive Rate).

NULLES MODEĻI SUGU IZPLATĪBAS MODELĒŠANĀ

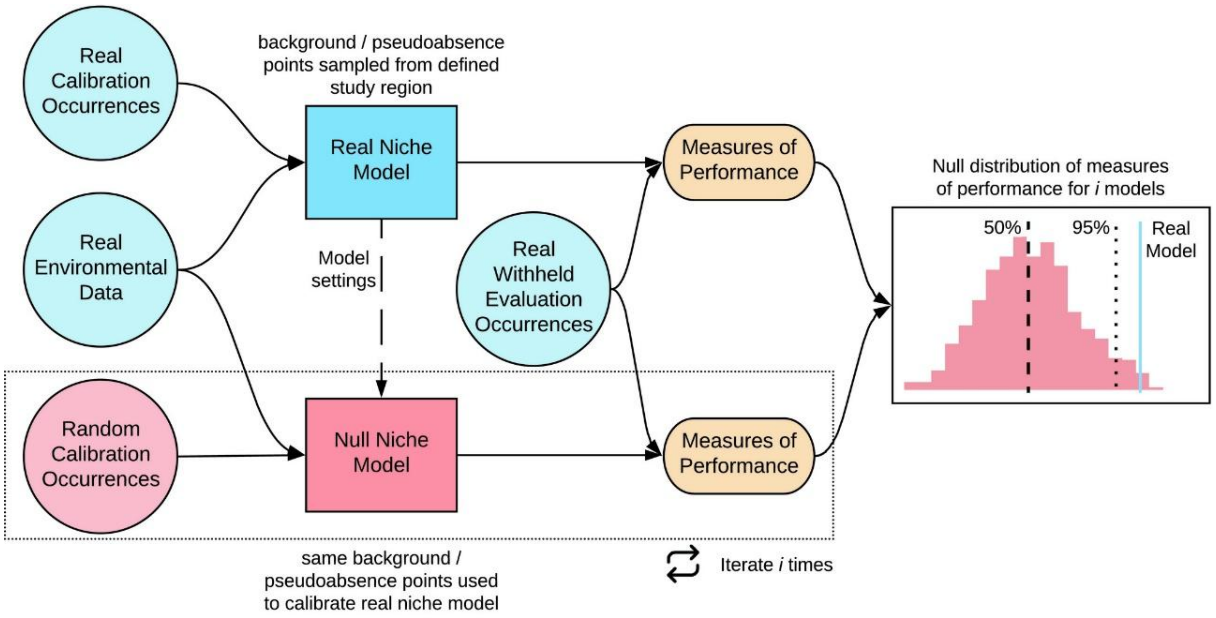
Sugu izplatības modelēšana (SDM, Species Distribution Modeling) ir plaši pielietota metode ekoloģijā, bioģeogrāfijā un dabas aizsardzībā. Šīs metodes mērķis ir kvantificēt sakarības starp vides faktoriem un sugu izplatību, interpolējot zināmās attiecības starp sugas klātbūtni un neesamību (vai tikai klātbūtni) un vides radītājiem (Raes and ter Steege, 2007). Rezultātā iegūto sugu ekoloģisko aprakstu var izmantot, lai raksturotu konkrētu faktoru nozīmīgumu un prognozētu sugu izplatību ārpus pētītajām teritorijām, padarot šo pieeju par spēcīgu rīku klimata pārmaiņu seku izpētē un nozīmīgāko teritoriju prioritizēšanā aizsardzībai (Miller, 2010).

1.1. Nulles modeļi sugu izplatības modelēšanā

Nulles modeļi statistikā ir uz nejaušību balstīti modeļi, kas tiek veidoti no tiem pašiem datiem kā pārbaudāmie procesu aprakstošie modeļi un saglabā noteiktas datu pamatīpašības, kamēr citi elementi tiek ģenerēti nejauši. Nulles modelis rada nejaušo sadalījumu, ko izmanto salīdzinājumam ar procesu aprakstošo modeli, lai noteiktu, vai novērotā datu struktūra pārsniedz to, ko varētu sagaidīt, ja tā būtu radusies tikai nejaušības un sistēmas ierobežojumu (elementi, kas nulles modelī tiek saglabāti nemainīgi, piemēram sugas sastopamību (kopējo ierakstu skaitu) vai vietu bagātību (sugu skaits katrā paraugošanas vietā) (constraints of the system)) dēļ (Gotelli and McGill, 2006; Gotelli and Ulrich, 2012; Veech, 2012).

Sugu izplatības modelēšanā nulles modeļus izmanto, lai aprakstītu sagaidāmo sugas izplatību, kas rastos, pilnībā izslēdzot kādu ekoloģisko mehānismu (piemēram, sugu mijiedarbību, nišu atšķirības vai konkurenci starp sugām) (Harvey et al., 1983). Nulles modeļa mērķis ir nodrošināt pamatu, pret kuru salīdzināt sugu izplatības modeļa rezultātus, lai pārbaudītu, vai novērotās likumsakarības sugu izplatībā ir patiesi bioloģisku procesu rezultāts, vai tās varētu izskaidrot ar nejaušību vai zināmām datu īpašībām (Bohl et al., 2019; Gotelli and McGill, 2006; Harvey et al., 1983).

Vispārīga pieeja nulles modeļu analīzei sugu izplatības modeļa pārbaudei ietver vairākus soļus (1. attēls):



1. attēls. Nulles modeļu analīzes shēma (Bohl et al., 2019)

1. solis – datu sadalīšana:

Patiesu sugas novērojumu kopa tiek sadalīta divās apakškopās:

- Kalibrācijas kopa (Real Calibration Occurrences) (jeb treniņa kopa (Training Dataset)) – dati, kas tiek izmantoti gan nišu aprakstoša modeļa, gan nulles modeļa apmācībai.
- Evaluācijas kopa (Real Withheld Evaluation Occurrences) (jeb testa kopa (Test Dataset)) – dati, kas netiek izmantoti modeļa apmācībā, bet kalpo modeļa pārbaudei un novērtēšanai.

2. solis – nišu (vai citu ekoloģisko procesu) aprakstošā modeļa izveide:

Nišas modelis (Niche Model) tiek izveidots, izmantojot kalibrācijas datu kopu un sagatavotos (patiesos) ekoģeogrāfiskos mainīgos (EGVs), kā arī izvēloties attiecīgos modeļa iestatījumus.

3. solis – nulles modeļa izveide:

Nulles modelis tiek izveidots, balstoties uz to pašu kalibrācijas novērojumu kopu kā nišu aprakstošais modelis. Tomēr, patieso novērojumu punktu vietā tas, ņemot vērā zināmo punktu izvietojumu, ģenerē jaunus nejaušus punktus tajās pašās telpiskajās proporcijās, kādas bija treniņu kopā. Vienīgā atšķirība starp reālo un nulles modeli ir ievades punktu izvietojums. Modeļa izveidē tiek pielietoti tie paši EGVs un iestatījumi, kas tika izmantoti pārbaudāmajam nišas modelim.

4. solis – modeļu veiktspējas mērīšana:

Tālāk abi modeļi tiek novērtēti, izmantojot vienu un to pašu evaluācijas novērojumu kopu. Atšķirībā no agrākajām metodēm, kur nulles modelis tika pārbaudīts pret tiem pašiem nejaušajiem punktiem (Raes and ter Steege, 2007) (pieeja, kas koncentrējas uz matemātisko atbilstību, nevis konkrētas sugas nišas apraksta kvalitāti), mūsdienās pārbaudei tiek izmantota reālo novērojumu (evaluācijas) kopa. Šī pieeja nomaina testa fokusu no matemātiskas atbilstības uz bioloģisko prognozējamību un ļauj atbildēt uz jautājumu: vai izveidots nišas modelis sniedz precīzāku sugas ekoloģiskās nišas aprakstu nekā nejaušība (Bohl et al., 2019).

5. solis – nulles sadalījuma izveide:

Atkārtojam nulles modeļa veiktspējas mērīšanu i reizes. Katrā iterācijā tiek izveidots nulles modelis, balstoties uz nejauši izvēlētiem punktiem, un tas tiek pārbaudīts pret evaluācijas kopu. Katras iterācijas rezultātā iegūstam vienu skaitlisku vērtību – izvēlētās veiktspējas metrikas rezultātu. Rezultātā tiek veidota frekvenču histogramma, kas attēlo veiktspējas diapazonu, kas sasniedzams bez jebkādām bioloģiskām zināšanām.

Pēc tam tiek ņemta vēl viena vērtība – metrikas rezultāts, kas iegūts, pārbaudot reālo modeli, – un novietota tajā pašā histogrammā (shēmā šī vērtība attēlota ar zilo līniju). Tas ļauj novērtēt, vai iegūtais rezultāts nav nejaušība. Ja procesa aprakstošā modeļa metrikas vērtība atrodas ārpus 95% ticamības intervāla (parasti izmanto vienpusējo intervālu, jo interesē tikai, vai ar procesa aprakstošo modeli iegūtā metrikas vērtība ir būtiski lielāka – vai, gadījumos ar kļūdu metriku, mazāka – par metrikām, kas iegūtas ar nulles modeli) (Raes and ter Steege, 2007).

1.2. Sugu izplatības modeļu precizitātes novērtēšanas metodes

Sugu izplatības modeļu validāciju var veikt, izmantojot dažādas modeļa precizitātes metrikas. Visplašāk izmantotās ir: jutīgums (sensitivity), specifiskums (specificity), Kohena kappa (Cohen's kappa) un laukums zem ROC līknes (AUC, Area Under the Curve). Lielākā daļa sugu izplatības modeļu precizitātes metriku ir tieši vai netieši atvasinātas no kļūdu matricas (confusion matrix) – matricas, kas parāda, cik labi modelis klasificē datus:

- Jutīgums (sensitivity) – kvantificē proporciju no novēroto klātbūtnes vietu (observed presences), kas tika pareizi prognozētas kā klātbūtne, t.i., patiesi pozitīvo frakciju:

$$\text{Jutīgums} = \frac{TP}{TP + FN};$$

- Specifiskums (specificity) – kvantificē patiesi negatīvo frakciju, t.i., novēroto neesamību vietu proporciju, kas tika pareizi prognozēta kā neesamība.

$$\text{Specifiskums} = \frac{TN}{TN + FP};$$

- Kohena kappa (Cohen's kappa) – raksturo kopējo saskaņotību starp prognozēm un novērojumiem, ņemot vērā nejaušas sakritības iespējamību.
- Laukums zem ROC līknes (AUC) – iegūstams, vizualizējot jutīgumu (sensitivity) kā funkciju no nepareizi prognozētās pozitīvās frakcijas $(1 - \text{specifiskums})$ visiem iespējamajiem notikuma varbūtības prognozes sliekšņiem.

Pirmās trīs sugu izplatības modeļu precizitātes metrikas prasa, lai modelī aprēķinātās sastopamības varbūtības (probabilities of occurrence) tiktu diskretizētas līdz binārajām kategorijām – "klātbūtne" un "iztrūkums". Šai diskretizācijai nepieciešams noteikt atdalošo robežu, parasti izmantojot slieksni 0,5. Savukārt laukums zem ROC līknes (AUC) neprasa bināras vērtības, kas padara to neatkarīgu no izvēlētā sliekšņa un nodrošina vienotu kopējā modeļa precizitātes novērtējumu (Raes and ter Steege, 2007).

1.2.1. Nulles modeļiem izmantotās precizitātes metrikas

Nulles modeļi nav piesaistīti konkrētai metrikai, bet kalpo kā statistiskā bāze, lai novērtētu dažādu rādītāju nozīmīgumu un efekta lielumu.

Visbiežāk, lai novērtētu nulles modeļus, tiek izmantota laukuma zem ROC līknes (AUC) vērtība. No visiem iepriekš minētajiem sugu izplatības modeļa precizitātes rādītājiem AUC izrādījās vienīgais, kas nav atkarīgs no datu proporcijas, kas atspoguļo sugu klātbūtni jeb prevalenci (prevalence) (Raes and ter Steege, 2007). Tomēr atsevišķos gadījumos AUC var sniegt pārlieku optimistisku novērtējumu, jo tā tieši atvasināta no optimālā ROC sliekšņa, kurš tendēts pārvērtēt reto sugu sastopamību. Tas saistīts ar to, ka AUC ignorē prevalenci, optimizējot slieksni, lai uzlabotu TPR un FPR, nevis faktiskās pareizās prognozes. Tādējādi izmantojot proporcijas, nevis absolūtos skaitļus, retām sugām, kur patiesi pozitīvo un patiesi negatīvo vietu skaits ir neliels, nelielas izmaiņas datos tiek pārvērtētas (Manel et al., 2001). Vēl viena problēma – šī metrika kļūst ļoti neprecīza, ja tiek izmantoti tikai klātbūtnes dati. ROC līknes aprēķināšanai nepieciešamas divas komponentes – jutīgums un specifiskums –, kam nepieciešama informācija par sugas neesamību. Ja nav zināmas reālas sugas neesamības vietas, tās tiek aizstātas ar fona punktiem (pseudo-absences). Problēma ir, ka fona punkti tiek izvēlēti nejauši no reģiona, kur suga varētu būt sastopama, bet nav reģistrēta. Rezultātā maksimāli sasniedzamais AUC vairs nav 1,0, bet gan $1 - \frac{\alpha}{2}$, kur α ir sugas patiesais izplatības areāla īpatsvars (kas parasti nav zināms) (Raes and ter Steege, 2007). Tā kā maksimālais AUC vairs nav 1,0, iepriekš pieņemti sliekšņi zaudē jēgu, un, izmantojot pseudo-neesamības, AUC vērtības nav tieši salīdzināmas starp dažādām sugām (Bohl et al., 2019).

Kā alternatīva metrika var tikt izmantots izlaiduma kļūdu īpatsvars (Omission Error Rate, OR). Tas mēra, cik bieži modelis kļūdaini prognozē neesamību klātbūtnes punktā (t.i., nepiemērotību) un atspoguļo nepareizi prognozēto klātbūtnes novērojumu proporciju. Turpmāk modeļa OR tiek salīdzināta ar nulles modeļu OR sadalījumu. Tā kā OR ir kļūdas rādītājs, reālā modeļa vērtībai jābūt statistiski zemākai, nevis lielākai, nekā parādīts shēmā, salīdzinot ar to, ko varētu sagaidīt nejauši, kas liecina par labu modeļa veiktspēju (Bohl et al., 2019).

1.2.2. Sarežģītākie nulles modeļi telpiskās novirzes kontekstā

Vēl viena problēma, kas rodas, novērtējot SDM modeļus ar parastajām metrikām, piemēram, AUC, ir tā, ka šo vērtību bieži vien mākslīgi paaugstina telpiskas korelācijas.

Pirmkārt, novērojumi parasti nav izvietoti nejauši – cilvēki dod priekšroku viegli pieejamām vietām, piemēram, ceļu, pilsētu vai dabas rezervātu tuvumam, radot būtiskas paraugošanas novirzes. Otrkārt, vides apstākļi (piemēram, klimats, topogrāfija un veģetācija) ģeogrāfiski tuvākajos punktos parasti ir līdzīgāki nekā tālos. Rezultātā, validācijai izmantotie klātbūtnes punkti gandrīz vienmēr atrodas ģeogrāfiski tuvāk apmācības klātbūtnes punktiem nekā validācijas trūkuma (vai fona) punkti, un reģistrētie punkti aptver tikai daļu no faktiskā vides gradienta (Raes and ter Steege, 2007). Tādējādi "naīvais" modelis saņem augstu AUC vērtējumu nevis tāpēc, ka tas precīzi identificējis sugas ekoloģisko nišu, bet gan tāpēc, ka tas prognozē augstu piemērotību vietām, kas atrodas ģeogrāfiski tuvu apmācībā izmantotajiem punktiem, un zemu piemērotību vietām, kas atrodas tālu no tiem (Hijmans, 2012).

Risinājums var būt noteiktas paraugošanas piepuļu ievirzes korekcijas iekļaušana modelī. Viens no veidiem, kā samazināt šo ievirzi, ir nulles modeļa punktus izvēlēties tikai no tām vietām, kur vispār ir veikti kādi novērojumi. Šīs metodes loģika balstās uz vienādas ievirzes ieviešanu gan nulles, gan procesu aprakstošajā modelī, lai abiem modeļiem būtu vienādi sākuma nosacījumi (Raes and ter Steege, 2007).

Šo pieeju var tālāk attīstīt, veidojot ģeogrāfisko nulles modeli. Piemēram, Hijmans, 2012 izmanto nulles modeli, kas balstīts tikai uz ģeogrāfisko attālumu, ignorējot vides faktorus. Tas kalpo kā rādītājs, kas kvantificē, cik lielu daļu no modeļa precizitātes (AUC) nosaka vienīgi telpiskā tuvība (spatial sorting bias). Šī pieeja ļauj "noņemt" paraugošanas procesu radītās novirzes no gala rezultāta. Šādā kontekstā modeļa veiktspējas novērtēšanai tiek izmantota jaunā metrika – CAUC (Calibrated AUC), kas tiek aprēķināta pēc formulas:

$$C\ AUC = AUC_{reālais} - (AUC_{nulles} - 0.5),$$

kas atskaita no modeļa novērtējuma daļu, ko tas ieguvis tikai telpiskās neobjektivitātes dēļ.

1.2.3. Izaicinājumi nulles modeļu pielietojumā SDM

Literatūrā ir uzsvērts, ka ir izaicinoši izveidot patieso nulles modeli, kas balstās vienīgi uz statistisko procesu – pat ja tiek izmantoti kāda noteiktā veidā pārkārtoti dati,

nevar izslēgt iespējamību, ka procesi, kas veidoja datus, nav netieši iekļauti modelī (Harvey et al., 1983).

Parasti nulles modeli veido, izmantojot randomizētus novērojumu datus, lai emulētu nejaušu izplatību ārpus konkrēta interesējoša ekoloģiska procesa ietekmes. Tajā nav nepieciešams norādīt visus ekoloģiskos procesus; tā vietā saglabā atsevišķus sistēmas ierobežojumus, piemēram, sugas sastopamības biežumu vai sugu skaitu katrā vietā, savukārt visiem pārējiem nosacījumiem, kas saistīti ar interesējošo ekoloģisko procesu, ļauj nejauši variēties (Bohl et al., 2019). Tomēr, mēģinot radīt nulles modeli, kas maksimāli līdzinās reālajiem datiem, izņemot tieši to procesu, kas tiek pārbaudīts, pastāv risks netīšām iekļaut pašu ekoloģisko signālu nulles modelī. Sarežģītu ekoloģisko datu gadījumā ir gandrīz neiespējami garantēt, ka ierobežojumi patiešām ir neatkarīgi no pētāmā procesa, kas palielina II. tipa kļūdas varbūtību (Harvey et al., 1983; Gotelli and Ulrich, 2012).

No otras puses, iekļaujot modelī pārāk maz ierobežojumu, rodas pārāk "naivs" nulles modelis. Tas padara reālo modeli viegli atšķiramu no nulles, mākslīgi palielinot statistisko nozīmīgumu un radot I. tipa kļūdu (Raes and ter Steege, 2007; Gotelli and Ulrich, 2012).

IZMANTOTĀ LITERATŪRA

Bohl, C. L., J. M. Kass, and R. P. Anderson, 2019, A new null model approach to quantify performance and significance for ecological niche models of species distributions: Journal of Biogeography, v. 46, no. 6, p. 1101–1111, doi:10.1111/jbi.13573.

Gotelli, N. J., and B. J. McGill, 2006, Null Versus Neutral Models: What's The Difference? Ecography, v. 29, no. 5, p. 793–800, doi:10.1111/j.2006.0906-7590.04714.x.

Gotelli, N. J., and W. Ulrich, 2012, Statistical challenges in null model analysis: Oikos, v. 121, no. 2, p. 171–180, doi:10.1111/j.1600-0706.2011.20301.x.

Harvey, P. H., R. K. Colwell, J. W. Silvertown, and R. M. May, 1983, Null Models in Ecology: Annual Review of Ecology and Systematics, v. 14, p. 189–211.

Hijmans, R. J., 2012, Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model: Ecology, v. 93, no. 3, p. 679–688, doi:10.1890/11-0826.1.

Manel, S., H. C. Williams, and S. j. Ormerod, 2001, Evaluating presence–absence models in ecology: the need to account for prevalence: Journal of Applied Ecology, v. 38, no. 5, p. 921–931, doi:10.1046/j.1365-2664.2001.00647.x.

Miller, J., 2010, Species Distribution Modeling: Geography Compass, v. 4, no. 6, p. 490–509, doi:10.1111/j.1749-8198.2010.00351.x.

Raes, N., and H. ter Steege, 2007, A Null-Model for Significance Testing of Presence-Only Species Distribution Models: Ecography, v. 30, no. 5, p. 727–736.

Veech, J. A., 2012, Significance testing in ecological null models: Theoretical Ecology, v. 5, no. 4, p. 611–616, doi:10.1007/s12080-012-0159-z.