MaxEnt parametri un to izvēle

Marks Arnolds Župerka

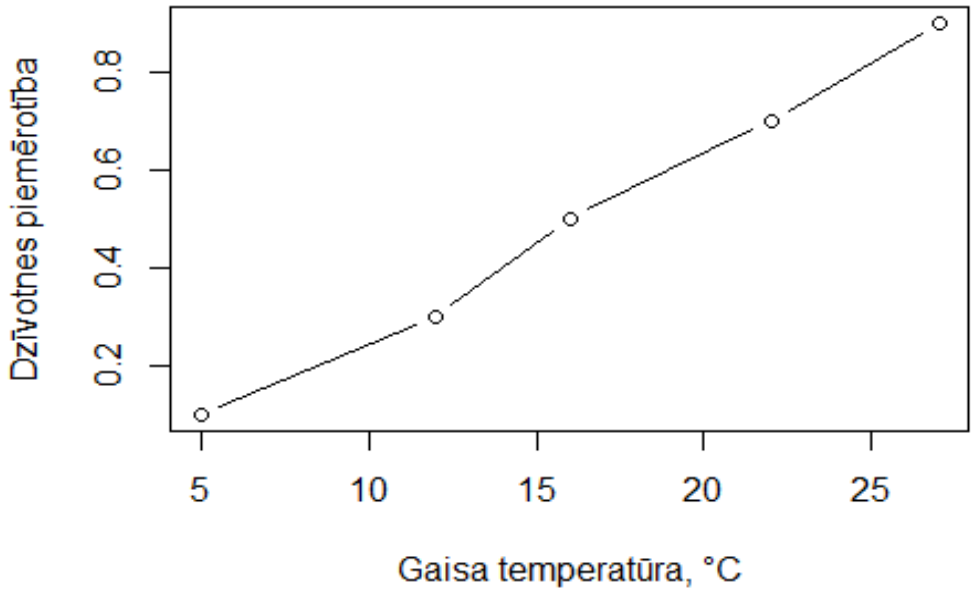
IEVADS

MaxEnt ir divi galvenie hiperparametri, kurus nepieciešams izvēlēties - pazīmju sakarības klase (angl. feature classes) un regularizācijas reizinātājs (angl. regularization multiplier, Morales et al., 2017). Pazīmju klases maiņa nodrošina iespēju definēt dažādas vienkāršas un arī kompleksākas sakarības starp vidi raksturojošajiem mainīgajiem un modeļa atbildes reakciju jeb sugas sastopamības iespejamības prognozi, kā arī starp pašiem mainīgajiem. Pēc noklusējuma MaxEnt sakarību klašu skaits tiks definēts pēc sugas novērojumu skaita - jo lielāks sugas novērojumu skaits, jo lielāks sakarību klašu skaits, savukārt ja ir virs 80 novērojumu, tiks izmantotas visas iespējamās sakarību klases (Merow et al., 2013). Jaunākie pētījumi tomēr akcentē nepieciešamību veikt ekoloģijā balstītu sakarības klašu izvēli un izvērtējumu, un nepaļauties uz MaxEnt noklusējuma iestatījumiem (Morales et al., 2017).

SAKARĪBU PAZĪMJU KLASES

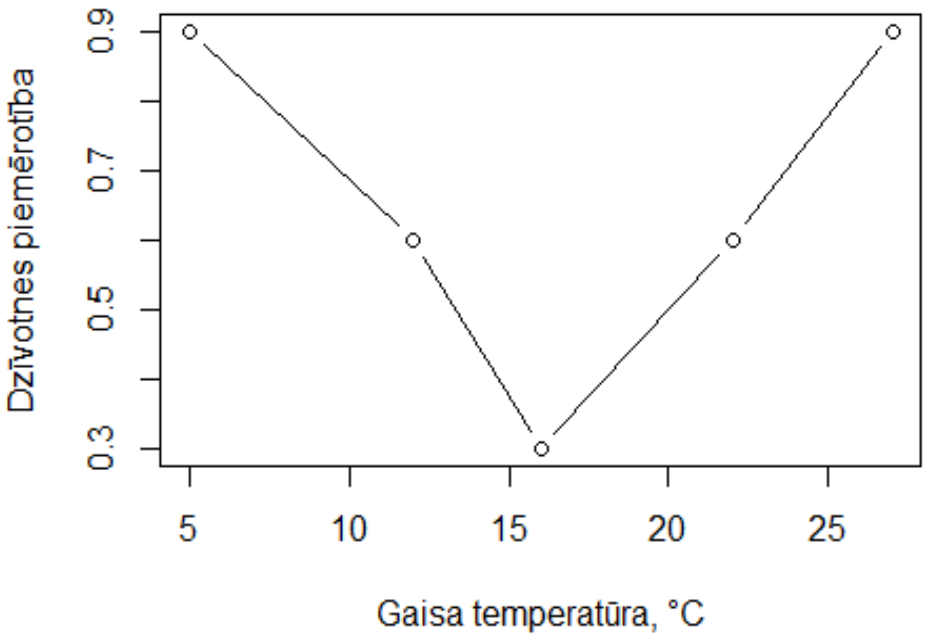
Pamatā ir sešas pazīmju sakarību klases (Phillips & Dudík, 2008, Low et al., 2020):

- Lineārā (angl. linear , L) sakarības klase ir vienkāršākā - attiecība starp vides mainīgo un prognozi ir lineāra visā vērtību diapazonā (sk. 1. att.). Tipiski šāda sakarība būtu pieņemama sugām ar samērā plašu ekoloģisko valenci - piem. zālājiem - jo lielāka ir gaisa temperatūra šūnā, jo lielāka ir sugas sastapšanas iespejamība tajā.



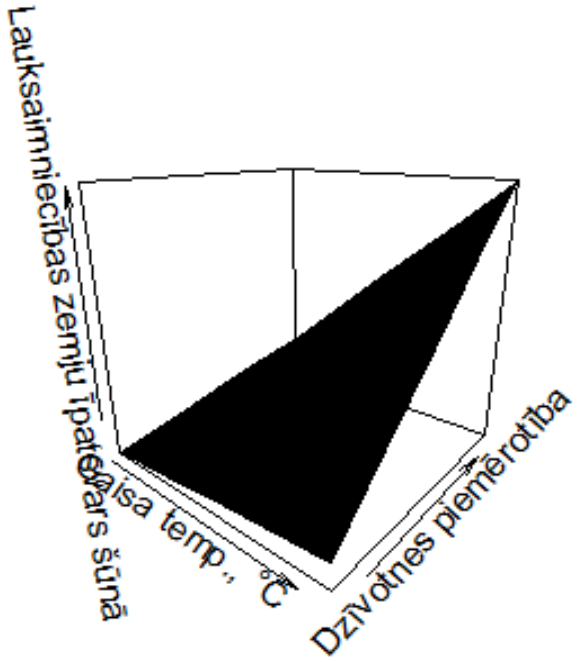
1. att. Lineāra sakarība starp dzīvotnes piemērotību un vides mainīgo (gaisa temperatūra)

- Kvadrātiskā (angl. quadratic , Q) sakarības klase norāda, ka attiecība starp prognozi un vides mainīgo būs raksturojama ar kvadrātfunkciju (sk. 2. att.). Šādas sakarības būtu ļoti šauras ekoloģiskas valances sugām - piem. temperatūrai - līdz noteiktam temperatūras slieksnim sugas dzīvotspēja palielinās, sasniedzot maksimumu pēc kura tā atkal samazinās.



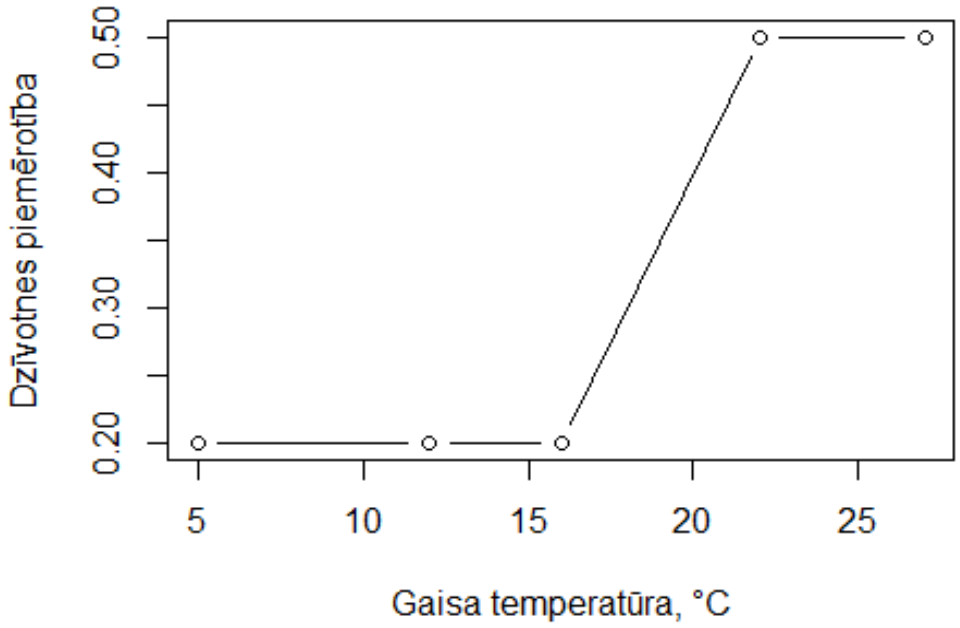
2. att. Kvadrātiskā sakarība starp dzīvotnes piemērotību un vides mainīgo (gaisa temperatūra)

- Kombinētās (angl. product , P) sakarības klases iekļaušana pieļauj, ka viena vides mainīgā ietekme uz sugu veidojas tikai kombinācijā ar kādu citu mainīgo pat, ja starp tiem nav cieša savstarpējā korelācija (sk. 3. att.). Piemērā būtu minams tas pats lineāras sakarības gadījums - jo lielāka ir gaisa temperatūra, jo augstāka ir sugas sastopamība, bet tikai, ja ap šūnu esošo lauksaimniecības zemju platība ir liela.



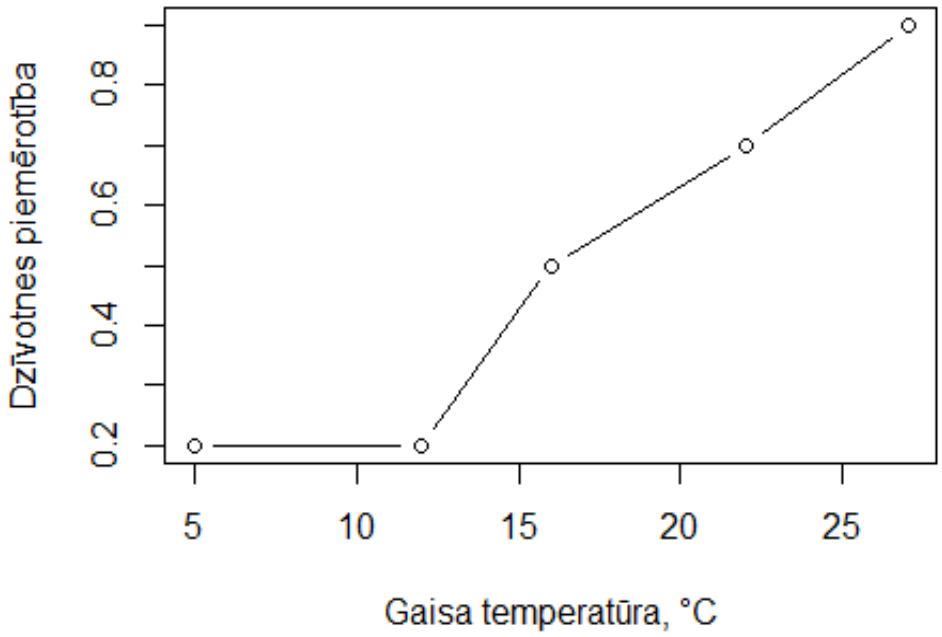
3. att. Vairāku faktoru mijiedarbības sakarība starp dzīvotnes piemērotību un vides mainīgajiem (gaisa temperatūra un lauksaimniecības zemju īpatsvars).

- Sliekšņa (angl. threshold , T) sakarības klase definē, ka līdz noteiktai vides mainīgā sliekšņa vērtībai dzīvotnes piemērotība ir konstants lielums un, pārsniedzot to, piemērotības vērtība mainās, bet joprojām saglabājas konstanta (sk. 4. att.). Piem. līdz 10°C temperatūra nav atbilstoša sugas eksistencei neatkarīga no tās absolūtās vērtības un piemērotība ir maza, bet pēc 10°C suga ir indiferenta attiecībā uz temperatūras pieaugumu.



4. att. Sliekšņa funkcijas sakarība starp dzīvotnes piemērotīu un vides mainīgo (gaisa temperatūra)

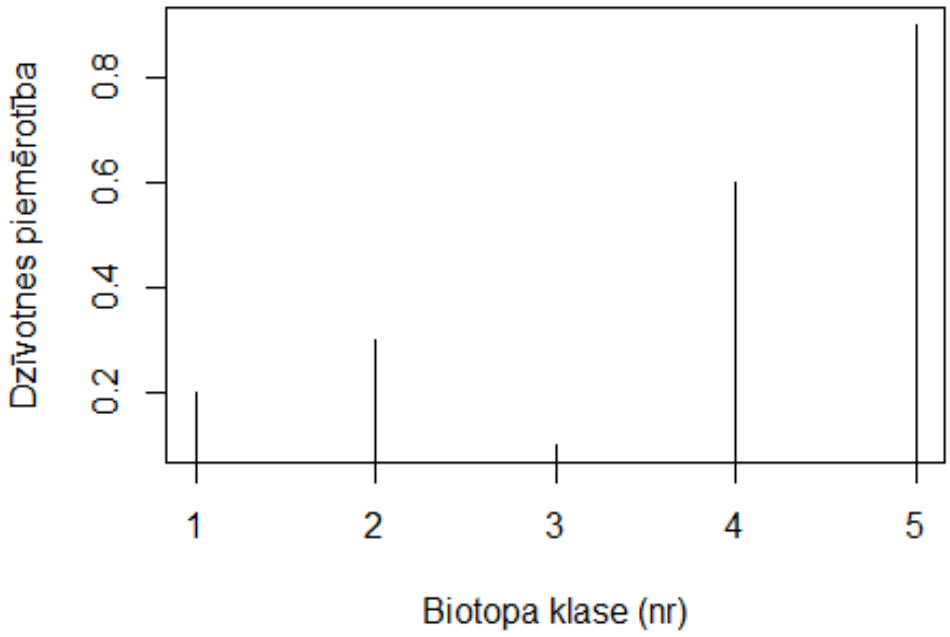
- Hindža (angl. hinge , H) sakarības klase ir līdzīga sliekšņa sakarības klasei, tikai atbildes reakcija pēc "sliekšņa vērtības" sasniegšanas kļūst lineāra (sk. 5. att.). Līdzīgi kā iepriekšējā piemērā - līdz 10°C temperatūra nav atbilstoša sugas eksistencei neatkarīgi no tās absolūtās vērtības un piemērotība ir maza, bet pēc 10°C turpinās lineāra sakarība jo lielāka ir temperatūra, jo lielāka ir vides piemērotība sugai.



5. att. Hindža funkcijas sakarība starp dzīvotnes piemērotību un vides mainīgo (gaisa temperatūra)

- Kategorijas indikatora (angl. category indicator , C) sakarības klases iekļaušana MaxEnt pieļauj kategorijās iedalīto mainīgo izmantošanu (sk. 6. att.). Piem. sugas piemērotība ir augstāka zālājos, mazāka mežos, pavisam zema purvos, u.t.t. Ja kategoriju dati kā vides mainīgie netiek izmantoti, tad nav nepieciešamības

iekļaut.



6. att. Kategorizētas vides mainīgā vērtības (biotopa klase, piem. nr) sakarība ar dzīvotnes piemērotību.

Sakarību klašu izvēle lielā mērā ir atkarīga no katras individuālās sugas un tās ekoloģiskās valences (Merow et al., 2013). Visbiežāk izmantotās kombinācijas ir LC, LQC, HC, HQC, TC un HQPTC (Phillips & Dudík, 2008), taču jāņem vērā, ka, izmantojot sarežģītākas sakarību klašu kombinācijas, visticamāk būs jāizmanto lielāks regularizācijas reizinātājs. Lineāro un hindža klašu izmantošana vienā modelī nav pamatota, jo pirmā ir otrās izņēmuma gadījums (Phillips & Dudík, 2008).

REGULARIZĀCIJAS REIZINĀTĀJS

Regularizācijas reizinātājs jeb multiplikators (Rm) ir pozitīvs skaitlis, ko izmanto lai kontrolētu kāda būs modeļa tendence vērtību prognozēšanā starp tiem datiem, kas ir bijuši apmācību stadijā un tie, kurus tas nav redzējis (Phillips, 2021). Rm kontrolē t.s. "modeļa pielipšanas" problēmu (angl. overfitting). Piem. ja kādas EGV vērtības robežu diapazons nav novērots sugas novērojumu punktos, tad pārāk "pielipušam" modelim būs vājas spējas pietiekami ticami prognozēt vērtības šajā intervālā - reāli dabā nepiemērotā vide modelis var divās blakus šūnās prognozēt pilnīgi pretējas vērtības, līdzīgi arī pārāk "plašam" modelim - modeļa nezināmā vide būs pārāk vienāda un nebūs iespējas labi izdalīt reāli piemēroto vidi no nepiemērotas. Līdz ar to būtiski ir atrast viduscelu starp abiem. Noklusējuma vērtība ir 1 - vērtības zemāk par 1 labāk "pielips" apmācību datiem, bet vērtības virs 1 veidos tendenci labāk ģeneralizēties visā mainīgo vērtību amplitūdā, ne tikai tajā, kas ir novērotas apmācību datos (Phillips, 2021). Rm izvēle līdz ar to varētu būt atkarīga no tā kāds ir paraugošanas piepūles raksturs - būtu vērtīgi apskatīties vienkāršāku aprakstošo statistiku ar vidējām vērtībām rastros sugas novērojumu punktos, minimālām un maksimālām vērtībām, u.t.t. Attiecīgi ja sugas novērojumi ir ļoti lokalizēti kādā noteiktā vidē, tad rm būtu jāliek virs 1, bet ja novērojumi ir izkliedēti pa ļoti plašu vides mainīgo vērtību amplitūdu, tad ir vērts aizdomāties par Rm vērtību starp 0 un 1. Labāko sakarības klasi un regularizācijas reizinātāju var izvēlēties arī daļēji automātiski - R valodā ir funkcija ENMeval::ENMevaluate (Kass et al., 2021), kas aprēķina statistiskos rādītājus dažādām sakarību klašu un regularizācijas reizinātāju kombinācijām, pēc kuriem attiecīgi var izvēlēties "labāko".

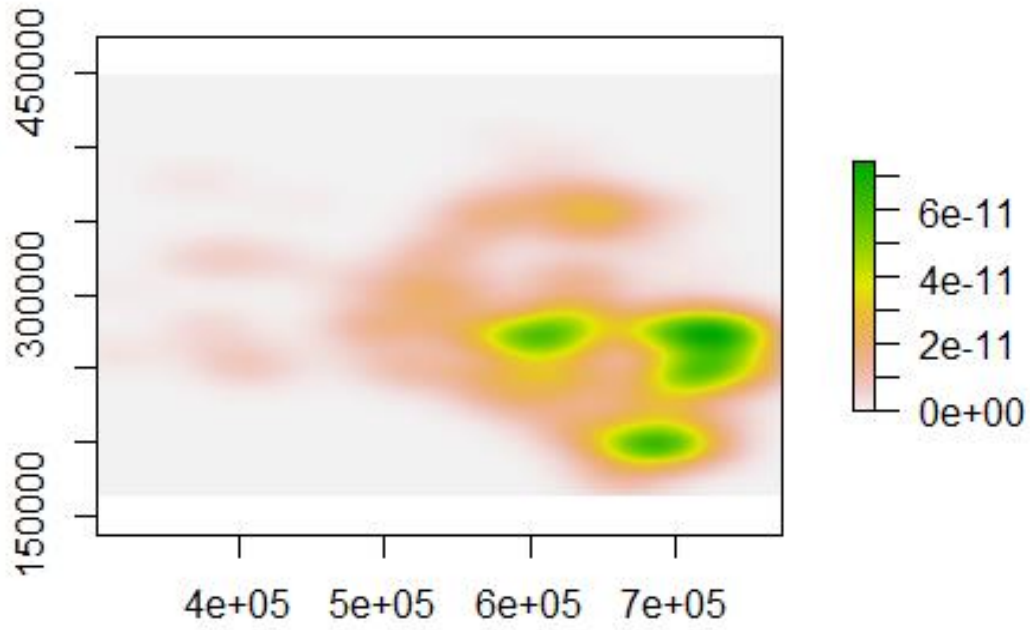
ENMEVALUATE – SIMULĀCIJA

Veikšu nelielu simulāciju lai labāk izprastu visu praksē - "izmēģinājuma" modelēšana zālāju sugai ko labi pazīstu - lielajam mārsilam (Thymus ovatus). Izmantoju "workflow" ko izstrādāja Joshua Banta (Banta, 2019). Konkrēti - "How to decide which settings to use when running Maxent, as well as how to make a bias file". Datora resursu un laika taupīšanas nolūkos izmantošu nevis 100 metru rastrus, bet 500 metru rastrus.

Piedāvāti divi ceļi pirms ENMevaluate izpildes, kuri atšķiras tikai ar to vai tiek veikta novērojumu telpiskā filtrēšana vai nē. Filtrēšana jau ir darīta novērojumu sagatavošanas stadijā, bet tā kā šoreiz ir 500 metru rastrs bet tad kad taisīju novērojumus zālājiem filtrēju 100 metru šūnās, izmantošu pieeju ar bias slāņa izveidi, kurā filtrēšanu nav jāveic. Papildus tam lai ENMevaluate tiktu izpildīts būs nepieciešama arī rJava pakotne. Manā gadījumā vajadzēja uzinstalēt 64-bitu Java versiju uz datora pirms varēja palaist library(rJava) un nebūtu kļūdas ar ceļiem reģistrā. MASS pakotne nepieciešama kernel-density estimation funkcijai, kas piemērā izmantota lai izveidotu paraugošanas piepūles tendenču (angl. bias failu).

Sekojot pieejai kas bija dota piemērā - jāizveido RasterStack. Tālāk izmēģinot ENMevaluate funkciju kļūst skaidrs ka pat 500 metru izšķirtspējā izveidot modeļus ar 131 mainīgajiem ir ļoti izaicinoši, tādēļ konservatīvi izvēlos 30 nejaušus mainīgos.

Tālāk veidots fona piepūli (angl. bias file) raksturojošs fails, izmantojot two-dimensional kernel density estimation (sk. 7. att.). Rezultātā būs tāds kā “heatmap”, kurā būs redzams vietas kur novērojumu ir bijis vairāk. Lielajam mārsilam izskatās ka tā ir Latgale.



7. att. Fona piepūli raksturojošs rastra slānis lielā mārsila novērojumiem.

Pēc visu datu sagatavošanas tiek izpildīta ENMeval:ENMevaluate funkcija (sk. 8. att.) - pie argumenta "tune.args" ir iespēja norādīt kādas sakarību klases tiek izmantotas un kāds ir regularizācijas reizinātājs. Iestatījumus nemainu un atstāju tādus kādi ir - izmantoju visas klašu kombinācijas izņemot C, jo kategoriju mainīgo mums kā EGV nav. Regularizācijas reizinātājus arī atstāju pēc noklusējuma piemērā - no 1 līdz 5 - neesmu pārbaudījis cik plašs ir vides mainīgo vērtību diapazons novērojumu punktos, bet tomēr lielajam mārsilam pieļauju ka vides mainīgo vērtību amplitūda būs ļoti, ļoti maza (suga ir ļoti izteikta) Piemērā gan nav sekots tam kas minēts vienā no manis pieminētajām publikācijām iepriekš un L un H ir izmantoti kopā. Rezultāts izveidots pēc nedaudz mazāk nekā 2h (laika starpību parāda pati funkcija izpildes beigās).

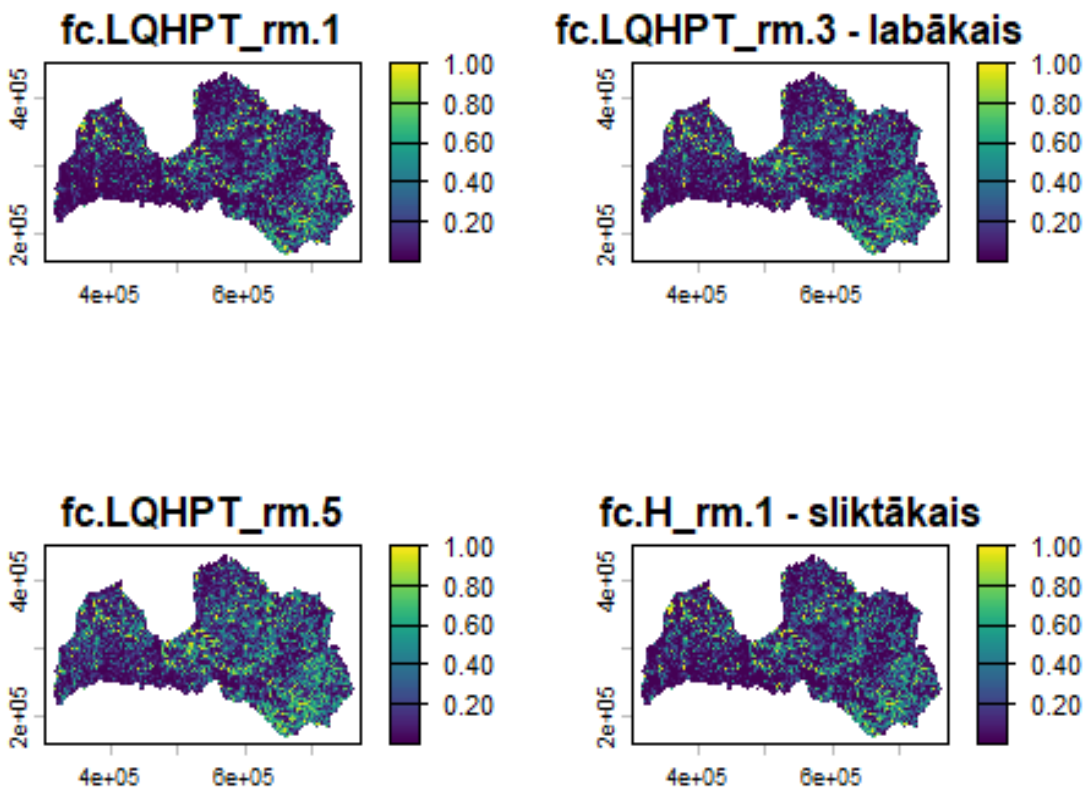
```
enmeval_results <- ENMevaluate(
                        sp,
                        env,
                        bg,
                        tune.args = list(fc =
                                           c("L","LQ","H", "LQH", "LQHP", "LQHPT"),
                                         rm = 1:5),
                        partitions = "randomkfold",
                        partition.settings = list(kfolds = 4),
                        algorithm = "maxent.jar",
                        parallel = FALSE,
                        updateProgress = TRUE)
```

8. att. Hiperparametru salīdzināšanas funkcijas ENMevaluate argumenti un piemēri. sp - sugu novērojumu fails (x un y koordinātas); env – vides mainīgie (SpatRaster formātā); bg – fona punkti (angl. *background points*); tune.args – sakarības klašu un regularizācijas reizinātāju parametru ievade; partitions – krosvalidācijas metode; partition.settings – krosvalidācijas metodes iestatījumi; algorithm – metode pēc kuras tiks veikta izplatības modelēšana; parallel – paralēlās procesēšanas nodrošināšana; updateProgress – vizuālas funkcijas izpildes loga attēlošana.

Continuous Boyce Index netika noteikts jo nebija instalēta ecospat pakotne. Zemākā "delta.AICc" vērtība ir modelim kurā ir visas 5 sakarību klases (LQHPT) un regularizācijas reizinātājs - 2. Konsekventi arī pats koriģēts Akaike informācijas kritērijs (AICc) ir vismazākais šai parametru kombinācijai. AICc un arī AIC pats par sevi būtiski neko nepasaka, taču tas ir izmantojams lai salīdzinātu vienādas saimes modeļus savā starpā un izvēlēties "labāko". Parasti tiek pieņemts, ka labākais modelis ir tas, kuram ir zemākā AIC vērtība starp salīdzināmajiem (Portet, 2020), kas ir minēts arī piemērā.

Apskatot atšķirības vizuāli skaidri redzams (sk. 9. att.), ka piem. Rietumkursas augstienē vismazākās dzīvotnes piemērotības prognoze ir ar Rm = 1. Ar Rm = 3 prognoze jau ir lielāka, pet pie Rm = 5 - vidējā vērtība reģionā varētu būt vislielākā, turklāt izskatās ka arī citās teritorijās piemērotības vērtības ir izteikti lielākas nekā pārējos reģionos. Tātad sugas gadījumā tiešām rm = 3 varētu būt labākais (neskatoties uz citiem modeļu veiktspējas rādītājiem). "Sliktākajam" modelim ar rm = 1 izskatās ka vislielākā

līdzība arī ir ar “fc.LQHPT_rm.1” tāda paša regularizācijas reizinātāja modeli. Attiecīgi Rm palielināšana var padarīt modeli pārāk ģeneralizētu (sugas gadījumā).



9. att. Dzīvotnes piemērotības modeļu prognožu rezultāti dažādām sakarības klašu un regularizācijas reizinātāju kombinācijām.

Gan sakarības klase, gan regularizācijas reizinātājs būtu jāmaina katrai sugai atsevišķi. Skaidrs, ka ENMeval:ENMevaluate var izveidot arī modeļus visām 120 iespējamajām klašu kombinācijām (5! = 120, bez kategoriju klases) un arī vairākiem multiplikatoriem, bet katras kombinācijas aprēķināšana un krosvalidēšana “X” reizes prasa gan daudz skaitļošanas resursus, gan arī laika. Līdz ar to abus iestatījumus būtu vērts mainīt no sugas uz sugu tikai ja ievērojami atšķiras tās vides mainīgo “komplekts” vai arī ievērojami mainās sugas ekoloģiskā valence.

ATSAUCES

Banta, J. 2019. Tutorials - How to decide which settings to use when running Maxent, as well as how to make a bias file. The Banta Lab. Sk. 19.11.2025. Pieejams https://sites.google.com/site/thebantalab/tutorials#h.c1h6hsh4lsn

Kass J, Muscarella R, Galante P, Bohl C, Buitrago-Pinilla G, Boria R, Soley-Guardia M, Anderson R (2021). "ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions." Methods in Ecology and Evolution, 12(9), 1602-1608.https://doi.org/10.1111/2041-210X.13628.

Low, B. W., Zeng, Y., Tan, H. H., & Yeo, D. C. (2020). Predictor complexity and feature selection affect Maxent model transferability: Evidence from global freshwater invasive species. Diversity and Distributions, 27(3), 497–511. https://doi.org/10.1111/ddi.13211

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

Morales, N. S., Fernández, I. C., & Baca-González, V. (2017). MaxEnt's parameter configuration and small samples: are we paying attention to recommendations? A systematic review. PeerJ, 5, e3093. https://doi.org/10.7717/peerj.3093

Phillips, S. J. 2021. A Brief Tutorial on Maxent. AT&T Research. Sk. 19.11.2025. Pieejams https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial_2021.pdf

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

Portet, S. (2020). A primer on model selection using the Akaike Information Criterion. Infectious Disease Modelling, 5, 111–128. https://doi.org/10.1016/j.idm.2019.12.010