Akadémiai Doktori Értekezés

Statisztikai modellezési módszerek az agrártudományi kutatásokban

Ladányi Márta

Magyar Agrár- és Élettudományi Egyetem
Matematika és Természettudományi Alapok Intézet
Alkalmazott Statisztika Tanszék

Budapest, 2024.

“… Mikor az *út*-ról szólal,
a szó ízetlen, sótlan.
Aki ránéz, nem látja,
aki hallgatja, nem hallja,
de nem-fogyó kincs annak, ki érti.”
(Lao Ce, Weöres Sándor fordítása)

Tartalomjegyzék

Az értekezésben használt rövidítések jegyzéke

Rövidítés	Angol elnevezés	Magyar elnevezés
ANN	Artificial Neural Networks	Mesterséges neurális hálózatok
AUC	Area Under the ROC Curve	A ROC görbe alatti terület
BB	Bagging and Boosting	Zsákolás és gyorsítás
CART	Classification and Regression Trees	Klasszifikációs és regressziós fák
DA	Discriminant Analysis	Diszkriminanciaelemzés
DNN	Deep Neural Networks	Mélytanuló neurális hálózatok
DT	Decision Trees	Döntési fák
FWER	Familywise Error Rate	Az elsőfajú hibavalószínűség többszöri hipotézisvizsgálati döntés után
GAM	General Additive Model	Általános additív modell
GBM	Gradient Boosting Method	Lépésenként javító módszer
GLM	General Linear Model	Általános lineáris modell
GLM	General Linear Mixed Model	Általános lineáris kevert modell
GLzM	Generalized Linear Model	Általánosított lineáris modell
GLzMM	Generalized Linear Mixed Model	Általánosított lineáris kevert modell
HC	Hierarchical Clustering	Hierarchikus klaszterezés
KMC	K-Means Clustering	K-közép klaszterezés
KNN	K-Nearest Neighbour	K-legközelebbi szomszéd
LDA	Linear Discriminant Analysis	Lineáris diszkriminanciaelemzés
LLM	Large Language Models	Nagy nyelvi modellek
LogR	Logistic Regression	Logisztikus regresszió
LR	Linear Regression	Lineáris regresszió
LRT	Likelihood Ratio Test	Likelihoodhányados-próba
MDA	Mean Decrease Accuracy	Átlagos pontosságcsökkenés
ML	Maximum Likelihood	Maximum likelihood
MLR	Multiple Linear Regression	Többszörös lineáris regresszió
NBM	Naive Bayes Method	Naiv Bayes-módszer
NLR	Nonlinear Regression	Nemlineáris regresszió
OLS	Ordinary Least Squares	Közönséges legkisebb négyzetek
OOB(E)	Out-of-Bag (Error)	A kiválasztott halmazon kívüli (hiba)
PCA	Principal Component Analysis	Főkomponens-elemzés
PCR	Principal Component Regression	Főkomponens-regresszió
PLSR	Partial Least Squares Regression	Parciális legkisebb négyzetek regresszió
QDA	Quadratic Discriminant Analysis	Másodfokú diszkriminanciaelemzés
RBF	Radial Basis Function	Sugárirányú bázisfüggvény
RF	Random Forest	Véletlen erdő
RMSE	Root Mean Square Error	Négyzetes hiba átlagának négyzetgyöke
ROC	Receiver Operating Characteristic (Curve)	ROC (görbe)
RT	Regression Trees	Regressziós fák
RSM	Response Surface Method	Válaszfelületek-módszer
SLR	Simple Linear Regression	Egyszerű lineáris regresszió
SVM	Support Vector Machines	Támogatóvektor-gépek
UPGMA	Unweighted Pair-Group Mean Average	Átlagos klasztertávolság
VIF	Variance Inflation Factor	Variancianövekedési faktor

A táblázatok jegyzéke

Az ábrák jegyzéke

A saját (társszerzős) publikációkból vett példák jegyzéke

1 BEVEZETÉS

1.1 Az alkalmazott statisztika helye az agrártudományi kutatásokban

A statisztikai elemzések az agrártudományi kutatásokban kiemelkedő jelentőséggel bírnak. Az agrártudományokban, illetve élettudományokban gyakran végeznek kísérleteket, megfigyeléseket. Az adatnyerés statisztikai elemzésre alkalmas tervezése a megbízható eredmények alapfeltétele. A tudományosan megalapozott statisztikai elemzések alapján tudjuk meghatározni, hogy a kapott eredmények mit jelentenek, mennyire pontosak, illetve általánosíthatók más körülmények között.

A sikeres publikáció elengedhetetlen feltétele a szakszerű statisztikai elemzésen túl a kapott eredmények körültekintő interpretálása a tudományos kutatási kérdésre adott válasz formájában, amelynek helyessége közvetlenül befolyásolja a következtetések hitelességét, valamint a kutatás eredményességét és tudományos hozzájárulását.

1.2 Az alkalmazott statisztikus sajátos helye az agrártudományi kutatásokban

Az alkalmazott statisztikus egy sajátos, önálló szakterület képviselője, amely speciális szakértelmet igényel. A vezető tudományos folyóiratok szigorú statisztikai elemzési követelményeket támasztanak. A publikáció sikeressége érdekében a kutatóknak meg kell felelniük ezeknek a követelményeknek, ideértve a statisztikai módszerek helyes alkalmazását, valamint az eredmények részletes bemutatását és értelmezését. Az alkalmazott statisztikus közreműködése biztosítja, hogy az elemzések megfeleljenek a legmagasabb tudományos normáknak, növelve a publikáció esélyét és színvonalát. Az agrártudományi kutatók és a statisztikusok közötti együttműködés tehát lehetővé teszi, hogy mindkét fél a saját szakterületén kiemelkedően teljesítsen. A statisztikusok hozzájárulása növeli a kutatások alaposságát és érvényességét, ami mélyebb, átfogóbb eredményekhez vezethet.

1.3 A statisztikus, az alkalmazott statisztikus és a biostatisztikus

Az agrárterületen dolgozó alkalmazott statisztikusok, a biostatisztikusok és az alapkutatást végző statisztikusok között jelentős különbségek vannak mind a szakterületük, mind pedig az alkalmazott módszereik és céljaik tekintetében. Ezek a különbségek meghatározzák az egyes statisztikai szakemberek munkáját és hozzájárulását a különböző tudományterületekhez.

Az agrárstatisztikus az agrártudományi kutatásokra és alkalmazásokra vonatkozó speciális tudással rendelkezik. Ismeri az agrárkutatások specifikus kihívásait, az ott felmerülő adatstruktúrákat, az egyes agrártudományi területek sajátos adatképzéseit, az adatok kinyerésével kapcsolatos problémákat, bizonytalanságokat, megbízhatóságokat, a releváns statisztikai módszereket, azok feltételeit és korlátait. A kutatás tervezésekor, annak kivitelezésekor, az adatok feldolgozása, elemzése, az eredmények értékelése, értelmezése, a következmények megfogalmazása folyamán a szakterületükön dolgozó agrárkutatókkal együttműködve releváns kérdéseket és javaslatokat fogalmaz meg. Közös céljuk többek között a mezőgazdasági termelés hatékonyságának növelése, a természeti erőforrások, a környezet fenntartható használata, a mezőgazdasági termelők, illetve döntéshozók támogatása a döntéshozatalban, a biztonságos és fenntartható élelmiszertermelés szolgálata, a nemesítés, a növényvédelem, az élhető emberi környezet előállítása, illetve fenntartása.

Az agrárstatisztikustól eltérően a biostatisztikus főként olyan élettudományok területén dolgozik, mint például az orvostudomány, az epidemiológia, a gyógyszerkutatás

és a biológia. Feladatai közé tartozik például a klinikai vizsgálatok tervezése és elemzése, valamint a betegségek előfordulásának és terjedésének modellezése. Az e tudományágakban fellépő adatstruktúrák, a tervezéssel, a mintavétellel való kihívások, a módszerek és azok alkalmazhatósága sok tekintetben nagyon eltér az agrártudományi alkalmazásoktól. Célja az orvosi és biológiai kutatások támogatása, amelyek hozzájárulnak az egészségügyi eredmények javításához és a betegségek megértéséhez, segítenek az új gyógyszerek hatékonyságának és biztonságosságának értékelésében, valamint a népegészségügyi programok hatékonyságának növelésében.

Az alapkutatást végző statisztikusok az elméleti statisztika fejlesztésére és új módszertani eljárások kidolgozására összpontosítanak. Munkájuk gyakran absztrakt és matematikai jellegű, és céljuk a statisztikai elméletek és módszerek általánosíthatóságának és alkalmazhatóságának növelése. Kevésbé kapcsolódnak közvetlenül egy adott alkalmazási területhez, inkább új statisztikai technikákat fejlesztenek, amelyek széleskörűen alkalmazhatók különböző területeken. Hozzájárulásuk alapvető fontosságú a statisztikai tudomány fejlődése szempontjából, és munkájuk gyakran valamely szakterület statisztikai kihívására ad választ, eredményeik ott kerülnek alkalmazásra.

1.4 Az alkalmazott statisztikus munkája és eredménye a kutatásokban

Egy közös agrárkutatási projektben, bár az alkalmazott statisztikus egyéni munkája egyértelműen elválasztható az agrártudományi szakterületi kutatók munkájától, mégis fontos megfogalmazni, hogy mit nevezhetünk az alkalmazott statisztikus saját eredményeinek. Az alkalmazott statisztikus saját eredményei olyan specifikus hozzájárulások, amelyeket főként, de nem kizárólagosan az adatok előkészítésével, elemzésével, a releváns statisztikai módszerek helyes alkalmazásával, a megfelelő modellek fejlesztésével, validálásával, értékelésével, az eredmények értelmezésével, a publikációban való közlés és vizualizáció kivitelezésével ér el. Ezek az eredmények közvetlenül tükrözik az alkalmazott statisztikus szakértelmét és szerepét a kutatásban. Ilyen értelemben, ennek a munkának a bemutatására készült ez a dolgozat a szerző társszerzős publikációiból.

Az alkalmazott statisztikus feladata először is megérteni a tudományos kutatás problémáját, a megválaszolandó kérdéseit, illetve a lehetséges, vagy a már megtörtént adatnyerés körülményeit, módszereit, lehetőségeit. Nem szerencsés, de sokszor előfordul, hogy a statisztikussal való közös munka már az adatok rögzítése után kezdődik, azaz a statisztikus nem vesz részt a kísérletek, megfigyelések tervezésében. Ennek részben az is oka – és ez a szerencsésebb eset, – hogy a kutatók a kísérletük, megfigyelésük tervezésekor a szakterületük protokolljait követik.

Az alkalmazott statisztikussal való munka intenzív konzultációkkal kezdődik. A statisztikus – a legtöbbször több ülésben – a legapróbb részletekig kikérdezi a szakembert a problémáról, az adatnyerés körülményeiről, hogy képes legyen felismerni az adatokban lévő bizonytalanságokat, feltárni az azokban rejlő struktúrákat, megérteni a rendszer viselkedését. Ezt egészíti ki a statisztikus önálló tanulási folyamata, amikor a szakterületen folyó munkához szükséges ismereteket szerzi meg. Nyilván nem lesz szakembere a területnek, de ahhoz, hogy kellően megértse a problémát, esetenként a nulláról kezdve meg kell ismerkednie az adott (szűkebb) szakterület alapvető összefüggéseivel.

Ezt követi az adat-előkészítés láthatatlan, de sokszor igen időigényes és sok szeméttel járó folyamata, melynek eredménye a statisztikai módszerek segítségével elemezhető, illetve a modellfejlesztéshez alkalmas adat. A megfelelő statisztikai módszerek kiválasztása, a modellek fejlesztése ugyanis függ az adatok típusától,

mennyiségétől, az adatnyerés módszereitől, körülményeitől, a kutatási kérdéstől, illetve a módszerek, modellek feltételrendszerétől és a módszereknek a feltételek kisebb-nagyobb mértékű sérülésére való érzékenységétől (ld. 1. Példa és 2. Példa).

Az alkalmazott statisztikus szerepe a statisztikai eredmények értelmezésével folytatódik. Ezeket a kutatótársakkal közösen a kutatási kérdés szakterületének nyelvére fordítják le, hogy a kérdésekre választ kaphassanak. Ilyenkor gyakran előfordul, hogy a vártnál mélyebb, vagy ahhoz képest meglepő válasz születik. Az alkalmazott statisztikus munkája tehát arra is kiterjed, hogy a rendelkezésre álló adatokban rejlő minél mélyebb, akár nem is várt vagy megcélzott üzenetet megtalálja, felfogja, igazolja. Ehhez sokszor nem elég a klasszikus statisztikai módszerek szolgai alkalmazása, hanem innovatív ötletekre is szükség van a modellek megformálásához, az adatokban rejlő többletinformáció kinyeréséhez. Az eredmények helyes statisztikai értelmezése biztosítja tehát, hogy a kutatók és az alkalmazott statisztikusok közösen hiteles következtetéseket vonjanak le.

A statisztikus feladata továbbá az eredmények tudományos igényű bemutatása, azok prezentálása megfelelően informatív és kompakt táblázatok és ábrák segítségével, végül mindezeknek a publikációra való előkészítése. Az eredmények világos és érthető közlése kritikus fontosságú a kutatás sikeres publikálása szempontjából (ld. 10. Példa, 11. Példa és 12. Példa).

1.5 Agrárstatisztikus képzés

Felmerül tehát a kérdés, hol lehet ilyen összetett tudásra szert tenni? Hol képeznek olyan szakembereket, akik otthonosak a statisztikai ismeretekben és hatékonyan tudnak kommunikálni az agrártudományok szerteágazó területein kutató kollégáikkal?

Míg a világ számos kiemelkedő egyeteme kínál kifejezetten biostatisztikus mesterképzést külföldön (Harvard, New York University, Arizona State University, University of Michigan, Universität Bremen, Universität Heidelberg, Universidad Valencia, Imperial College London, University of Cambridge, University of Copenhagen, Karolinska Institute, Universität Zurich, National University of Singapore, University of Hong Kong, Fudan University), illetve Magyarországon a Debreceni Egyetem és a Szegedi Orvostudományi Egyetem, addig agrárstatisztikus képzést elvétve találunk, ezek is inkább agrárképzések kissé hangsúlyosabb statisztikai kurzusokkal, vagy biostatisztikus képzések agrárstatisztikai témaköröket is említve (Texas A&M University, West Virginia University, Iowa State University, Tashkent State Agrarian University, Stavropol State Agrarian University, Technische Universität Munich, Wageningen University, University of Alberta, University of Guelph, China Agricultural University, Bogor Agricultural University, Indian Agricultural Research Institute). Magyarországon jelenleg nincs ilyen képzés. Folyamatban van egy agrárstatisztikus mesterképzési program indítása és akkreditálása a Magyar Agrár- és Élettudományi Egyetemen az értekezés szerzőjének vezetésével.

Ezen értekezés tehát azt is célul tűzi ki, hogy érvényt szerezzen ennek a speciális diszciplínának a hazai kutatási területek színes forgatagában.

1.6 A dolgozat szerkezete a célkitűzéseimet tükrözi

A dolgozat szerkezete szokatlan módon eltér a hagyományostól, aminek oka szintén a szakterület kívülállóságából ered.

Dolgozatomban a statisztikai modellező munkának azt a részét emelem ki, hogy hogyan lesz a kapott adatból az adatgazda számára hasznos eredmény, a vizsgált élő rendszer egy üzenete, amit az adatok körültekintő kibontásával együtt értettünk meg. A

dolgozat szokatlansága, hogy erről az útról fog szólni, hiszen ez az én munkám. Arról a folyamatról, amelynek egyes problémái, kihívásai, érdekes mozzanatai nem mindig kerülnek napvilágra, amikor egy publikációban a célként elért eredményt helyezzük a fókuszba. Hogy ezzel az új megközelítéssel az én eredményeimet közelebb hozhassam az Olvasóhoz, a kulisszák mögé kell tekintetünk.

A második fejezetben bemutatom a modellező munkát, majd a harmadik fejezetben a munkám során használt legfontosabb lineáris modelleket. A negyedik fejezetben rövid bevezetést nyújtok a gépi tanulási módszerekhez, illetve a módszerek értékeléséhez.

A fejezetek tárgyalása során igyekszem sok számozott példával szolgálni a mélyebb megértést. A számozott példákat csupa olyan publikációkból választottam, amelyekben társszerző vagyok.

Az ötödik fejezetben a kutatásaink során alkalmazott gépi tanulási módszereket mutatom be.

Az eredmények fejezetben hét kutatási projektet mutatok be kissé részletesebben. A témákat úgy választottam, hogy az adatelemző-modellező munka sokszínűsége lehetőleg jól kirajzolódjon.

A felhasznált irodalom jegyzékétől elkülönítve sorolom fel a dolgozat számozott példáiban idézett saját (természetesen társszerzős) publikációimat, valamint azokat, amelyeket az eredmények fejezetben részletesen is bemutattam.

2 A MODELLEZŐ MUNKA

A modellező munka lépései:

- Adatgyűjtés és -előkészítés
- Modellépítés, a modell paramétereinek becslése, a becslések finomítása, validálás
- Modellértékelés a modell típusától függő módszerekkel
- A jelölt modellek összehasonlítása
- A modellek értékelése az összehasonlítások tapasztalatai alapján; a modell továbbfejlesztése
- A legjobb modell(ek) kiválasztása
- A modellek validálása
- Az eredmények közlése és vizualizációja

2.1 Az adat-előkészítés

Az adatgyűjtés, valamint az adatok feltárásának és előkészítésének legfontosabb elemei, az adattisztítás, a felesleges/hibás adatok eltávolítása, a hiányzó, illetve kiugró (*outlier*) adatok feltárása és kezelése (Hastie és mtsai, 2013, Wickham, és Grolemund, 2023), szükség esetén az adatok összehasonlíthatóvá tétele (Boehmke és Greenwell, 2020), normálása, transzformálása. Az elemzésre alkalmas adatstruktúra előállításával kapott bemenő adatok minősége döntően meghatározza a modellező munka sikerét (Lantz, 2015). Erre William D. Mellin már közel 70 éve (The Hammond Times, 1957. november 10.) felhívta a figyelmet, és az ő figyelmeztetésének eredeztetik a széles körben elterjedt GIGO (*garbage-in-garbage out*), vagy a RIRO (*rubbish in-rubbish out*) "szemét be – szemét ki" mozaikszavakat.

A következő három példában az adattípusok sokszínűségét szeretném illusztrálni, illetve röviden bemutatom az ezek elemzésére választott néhány nagyon egyszerű alkalmazott módszert. A további módszerekről külön fejezetek szólnak majd.

1. Példa különböző adattípusokra egyetlen kutatáson belül

Vétek és munkatársai (2021) a közelmúltban betelepült, és a szilfákat jelentősen pusztító invazív kanyargós szillevéldarázs (*Aproceros leucopoda*) négy különböző gazdanövényen (*Ulmus crassifolia*, *Hemiptelea davidii*, *Zelkova serrata*, illetve kontrollként *Ulmus pumila* taxonokon) való életképességét vizsgálták.

Az egyedeket a taxon hajtásain laboratóriumi körülmények között, illetve cserepes növényeken nevelték. A hajtásokra rakott peték számát hat ismétlésben mérték, ezek *folytonos változóként* viselkedtek, a (paraméteres) ANOVA modell (ld. 3.1. fejezet) hibatagjának eloszlása meggyőzően normálishoz hasonló volt.

A cserepes növényekre rakott peték mortalitása *arányszám*, ezek elemzésére arányok összehasonlítására vonatkozó nemparaméteres módszerre volt szükség. Az *U. pumila* és a *H. davidii* peterakását arányokra vonatkozó Z-próbával hasonlítottam össze, és mivel nem volt magas a mintaelemszám, ezért a próbát kis mintaelemszámra vonatkozó korrekcióval végeztem. (Azért csak e kettő taxonét a négy közül, mert az *U. crassifolia* és a *Z. serrata* egyike sem rakott petét a cserepes növényekre.)

A hajtásokra az *U. pumila*, az *U. crassifolia* és a *H. davidii* is rakott petét, ezek mortalitását ismét arányok összehasonlítására vonatkozó nemparaméteres módszerrel végeztem, mégpedig olyannal, amely kettőnél is több arány összehasonlításakor is kezelni tudja az elsőfajú hibavalószínűség (*familywise error rate*, FWER) növekedését. A FWER annak a valószínűsége, hogy a k összehasonlítást követően legalább egyszer hamis

szignifikáns eredményt kaptunk, ha az elsőfajú hibát az egyes összehasonlítások során α-nak választjuk:

$$FWER = 1 - (1 - \alpha)^k,$$

ami k növelésével igen gyorsan nő. Ilyen fordulhat elő például, ha az összehasonlítani kívánt csoportokat páronként hasonlítjuk össze az előbb látott, arányokra vonatkozó Z-próbával az összes lehetséges párosításban. A többszörös összehasonlításra robusztus (a FWER növekedését kezelni képes) módszer az ún. Marascuilo-eljárás (Marascuilo, 1966), amit ebben az esetben alkalmaztam. Képezzük az i és j indexű két csoportra a $k-1$ paraméterű χ^2-eloszlást követő

$$\frac{\left|\widetilde{p_i} - \widetilde{p_j}\right|}{\sqrt{\left[\frac{\widetilde{p_i}(1-\widetilde{p_i})}{n_i} + \frac{\widetilde{p_j}(1-\widetilde{p_j})}{n_j}\right]}}$$

statisztikát, ahol k az összehasonlítani kívánt csoportok száma, f_1, f_2, f_3,…, f_k, a megfigyelt gyakoriságok, n_1, n_2, n_3,…, n_k a mintaelemszámok, ahonnan $\widetilde{p_i} = \frac{f_i}{n_i}$ az egyes csoportokban a vizsgált valószínűségek becslései. A $k-1$ paraméterű χ^2-eloszlás fenti értéktől vett improprius integrálja a H_0: $p_i = p_j$ (minden i, j párra) nullhipotézis tesztjének a p szignifikanciaértékét adja. A Marascuilo-eljárás alkalmazására egy bővebben ismertetett példát mutatok a 6.5. fejezetben.

Most két példát mutatok arra, amikor magas mérési szintű (folytonos típusú) adatok diszkretizálására volt szükség valamilyen oknál fogva (ld. még a 6.2. fejezetet).

2. Példa arra, amikor a folytonos adattípust elemzés előtt diszkretizáltuk

Nagy és munkatársai (2017) növény- és madárközösségek diverzitásmutatóit (SN: fajszám, TA: abundancia, SD: Shannon-diverzitás) hasonlították össze a megfigyeléshez kijelölt olyan magyarországi tájon, ahol az intenzíven kezelt mezőgazdasági területek és a magas természeti értékű, félig természetes gyepek mozaikos mintázata jól megfigyelhető. A tájon hat fókuszterületen, melyek egyenként 8 egybefüggő hatszögből álltak (rozetták) 15 kulcsfontosságú mezőgazdasági madárfaj jelenlétét, illetve hiányát jegyezték fel. A tájhoz annak természetességét leíró tájminőségi mutatót rendeltek (Magyarország növényzeti természeti tőkéjéhez való hozzájárulás indikátora, NTT, KSH, 2008). A kutatás fő kérdése az volt, hogy madarak diverzitásmutatóiból következtethetünk-e a táj minőségére, és hogy a 15 faj között találunk-e olyat, amely bizonyos szintű tájtermészetességnek indikátora lehetne.

A diverzitásmutatók (SN, TA, SD) folytonos változók, amelyeket válaszváltozókként illesztettem a MANCOVA modellbe (ld. 3.1. fejezet), a fókuszterület (FA) és a rozetták (R) független blokkok (az R az FA-ba ágyazva), a hatszögek tájminőségi mutatója (NTT) pedig kovariáns magyarázó változó volt.

Amikor azonban a madarak között esetleges indikátorfaj(oka)t akartunk találni, el kellett gondolkodnunk azon, hogy az NTT folytonos változó értékeit (azaz a táj természetességét) a madarak hogyan érzékelik, mert döntésüket feltehetőleg nagyjából ez alapján hozzák meg arról, hogy az a táj nekik megfelelő-e (azaz ott megjelennek-e), vagy nem. Mi pedig ebből a jelenlét/nem jelenlét információból szerettünk volna következtetni egy táj természetességére (a diszkutáláshoz ismerve hozzá a madárfajok egyéb tulajdonságait, elsősorban tájérték-érzékenységüket, illetve kapcsolatukat más fajok

jelenlétével). Egy madár nyilván nem magát a folytonos tájértéket észleli, hanem inkább azt, hogy az magas vagy alacsony. Ezért az NTT-értékeket öt kategóriába soroltam (0: NTT = 0; 1: 0<NTT ≤ 0,2; 2: 0,2<NTT≤ 0,4; 3: 0,4<NTT ≤ 0,6; 4: 0,6<NTT). Ezzel egy magasabb mérésszintű változót "butítottam le" alacsonyabb mérésszintűre, mégis ez vezetett minket célra. A most már két ordinális változó kapcsolatát a Somers-féle d szimmetrikus és aszimmetrikus, rangszámon alapuló asszociációs mérőszámmal jellemeztem (Somers, 1962).

A Somers d-hez megszámoljuk, hogy az X és Y változók összes lehetséges (x_i, y_i), (x_j, y_j) párjai közül hány konkordáns (C), hány diszkordáns (D) és hány egyik sem (kapcsolt, *tied*, T). Egy ilyen pár konkordáns, ha együtt mozognak, azaz ha $x_i < x_j$ és $y_i < y_j$, vagy $x_i > x_j$ és $y_i > y_j$ teljesül, diszkordáns, ha egymással ellentétes irányban mozognak, azaz ha $x_i < x_j$ és $y_i > y_j$, vagy $x_i > x_j$ és $y_i < y_j$ teljesül. Ha valahol egyenlőség áll fenn, akkor kapcsolt.

A szimmetrikus d statisztika

$$d = (C - d) / \left(C + D + \frac{(T_X + T_Y)}{2} \right)$$

(ahol T_X az X változóban kapcsolt, de az Y-ban nem, T_Y az Y változóban kapcsolt, de az X-ben nem) -1 és +1 közötti értékeket vesz fel. Előjeléből és szignifikanciaszintjéből kimutatható a faj jelenléte és a táj minősége közötti kapcsolat erőssége anélkül, hogy a kapcsolat irányáról (melyik változó értékének ismeretében nő a másik változó előrejelezhetősége) informálna.

Az aszimmetrikus d már erre is választ ad:

$$d = (C - D)/(C + D + T_X).$$

Meghatározásához tehát rögzíteni kell, melyik ordinális változónk a magyarázó (X) és melyik a válaszváltozó (Y). Ennek értéke és szignifikanciája jelzi, hogy egy madár jelenlétének ismeretében (magyarázó változó) következtethetünk-e az NTT-re vonatkozóan (válaszváltozó). A szerepek felcserélésével kapott Somers d érték pedig azt méri, hogy a táj természetességének ismeretében mennyire nő a madarak jelenlétének előrejelezhetősége.

A Somers d asszociációs mérőszám alkalmazásával hat olyan fajt találtunk, amelyek jelenléte magas természetességű táj indikátora lehet, három olyan fajt, amelynek jelenléte inkább alacsony természetességű tájra utal. Nem meglepő, hogy a táj természetességének ismeretében a madarak egyetlen fajának jelenlétére sem következtethetünk.

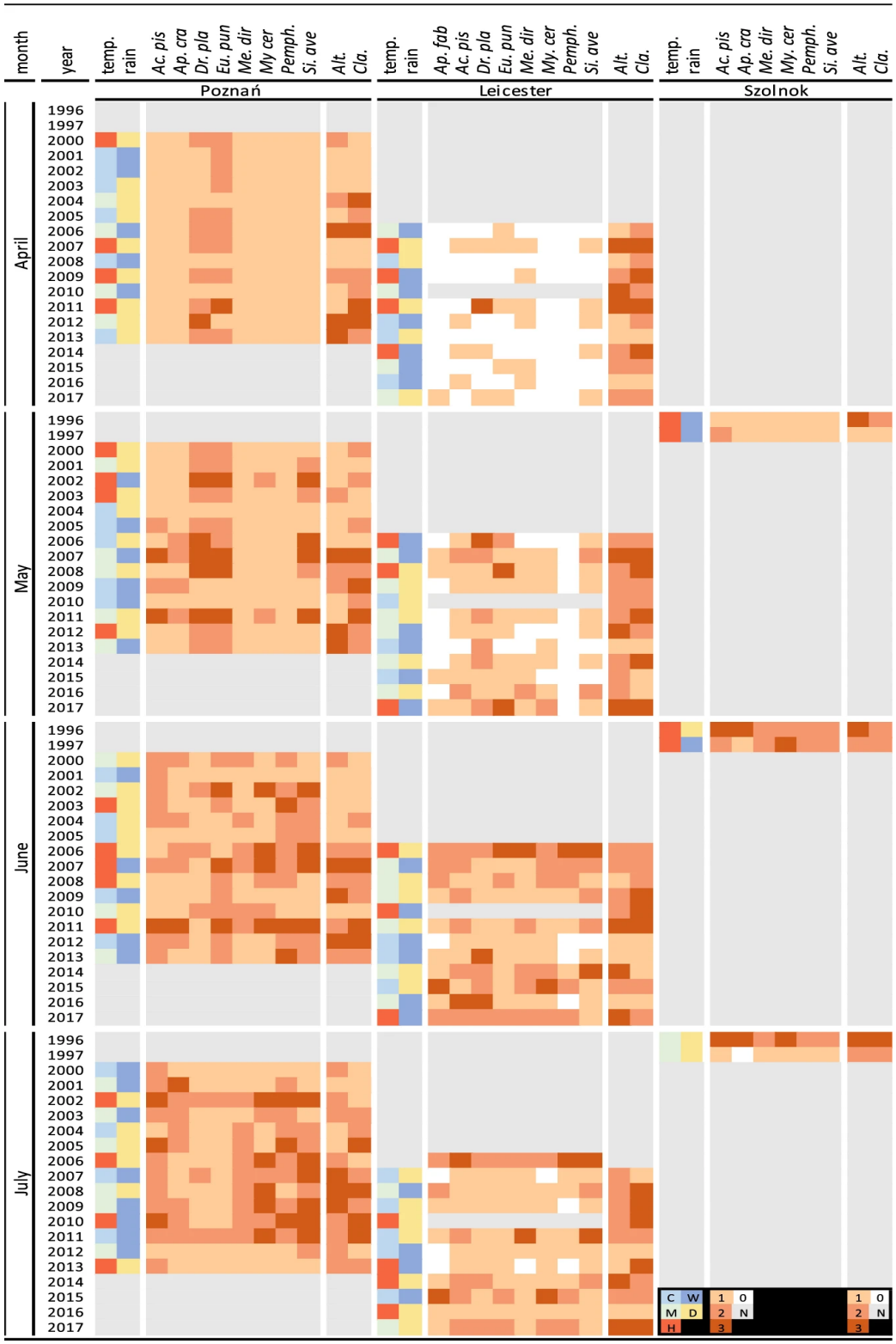
Az alkalmazott, rangszámon alapuló statisztika részünkről nagyon óvatos és konzervatív megközelítés volt, amelyet azért választottunk, mert semmit sem tudtunk a fajok gyakoriságának eloszlásáról (sem az NTT-értékekéről). A Somers-féle d semmiféle követelményt nem támaszt az ordinális változók eloszlására vonatkozóan. Azt vártuk, hogy ezzel a konzervatív megközelítéssel talált összefüggések valósak és robusztusak lesznek, amelyek a tájszintű növény- és madárközösségek közötti alapvető összefüggésekből, illetve a két közösséget alakító közös tényezőkből erednek.

3. Példa arra, amikor az adatokat az összehasonlíthatóság céljából diszkretizáltuk

Egy másik esetben, amikor magasabb mérésszintű változó diszkretizálásával jutottunk célra Magyar és munkatársait (2023) idézzük, akik azt vizsgálták hogy levéltetű-populációk méretéből lehet-e következtetni a patogén és allergén *Alternaria* és

Cladosporium spórák légköri koncentrációjára, aminek előrejelzése népegészségügyi szempontból igen fontos lehet. A levéltetvek által kinyert mézharmat szolgál ugyanis tápanyagként a korompenész fejlődéséhez, ami azt sejteti, hogy a levéltetvek populációdinamikája hatással van a mézharmaton élő gombák spóráinak szintjére a levegőben.

Az idézett tanulmányban hét levéltetű taxon populációméretének hatását elemeztük két patogén és allergén gombára, az *Alternaria*ra és a *Cladosporium*ra a spórás évszakokban összesen 20 éves adathalmazon, amelyek Magyarországról, az Egyesült Királyságból, valamint Lengyelországból származtak. Figyelembe kellett azonban venni, hogy a meteorológiai tényezők mind a levéltetű-, mind a gombapopulációk méretét közvetve és közvetlenül is jelentősen meghatározhatják, valamint azt, hogy ezért a három helyszínen mért egyedszámok, illetve spórakoncentrációk a helyszínek igen különböző meteorológiai paraméterei miatt nem összehasonlíthatóak. Ezért a helyszínekre jellemző, az egyes helyszínek havi hőmérséklet- és csapadékeloszlásától függően két csapadékosztályt (száraz, csapadékos) és három hőmérsékletosztályt (hűvös, enyhe, meleg) definiáltunk. Hasonlóan, a gyakran nagyságrendi eltéréseket is mutató egyedszámokat és spórakoncentrációkat is három-három osztályba soroltuk. Az egyes osztályok időbeli megjelenésének (együttállásának) hőtérképét az 1. ábra mutatja be.



1. Ábra Hőtérkép az időjárási, a levéltetű- és a spóraosztályokról az egyes helyszíneken (Poznan, Leicester, Szolnok). Színkódok (jobb alsó sarok): C: hideg; M: enyhe; H: meleg; W: nedves; D: száraz időjárás; 1: alacsony; 2: közepes; 3: magas levéltetű/spórakoncentráció; 0: nulla (nem kimutatható); N: nincs adat. Oszlopnevek: month: hónap; year: év; temp.: hőmérséklet, rain: csapadék, a levéltetvek neveinek rövidítései, valamint *Alt.*: *Alternaria* spp., *Cla.*: *Cladosporium* spp.

Ezután kereszttábla-elemzéssel, illetve folytatásként a módosított standardizált hibatagok abszolútértékén alapuló post hoc teszttel (Sharpe, 2015) tudtuk kimutatni a hatféle, hőmérséklet-csapadékosztály-párból álló komplex meteorológiai osztály és az egyes levéltetűtaxonok populációméret-osztályainak együttállását, és időbeli eltolódással a spóraelemszám-osztályok előfordulását.

Az alábbiakban a kiugró értékek szűrésére mutatok két példát.

4. Példa Kiugró értékek szűrése 1σ-2σ-3σ-szabály alkalmazásával

A kanyargós szillevéldarázs (*Aproceros leucopoda,* Hymenoptera: Argidae) hidegtűrési stratégiájának megismeréséhez (Vétek és mtsai, 2020, ld. 6.5. fejezet) az első lépés annak a fajra jellemző hőmérsékletnek a meghatározása, amikor a spontán jégképződés bekövetkezik, és amit szuperhűlési pontnak (*Super Cooling Point*, SCP) nevezünk. A kísérleti SCP-adatokban extrém magas kiugró értékek is voltak, ami azt a téves látszatot kelthette, hogy már olyan magas hőmérsékleten is elpusztulnak az egyedek. Ezeknek az egyedeknek a halálát azonban nagy valószínűséggel nem a hideg, hanem a minták előkészítése során az apró nimfák hámjának véletlen (és a kezelő által talán nem is észlelt) sérülése okozta. A kiugró értékek felismerése és kiszűrése tehát kiemelt fontosságú volt, nehogy a faj hidegtűrési képességét alulbecsüljük.

Ez esetben a kiugró értékeket standardizálással szűrtem. Az 1σ-2σ-3σ-szabály szerint a normális eloszlású változók 1,96-nál nagyobb, vagy -1,96-nál kisebb standardizált értékei tekinthetők kiugrónak. (Ez felel meg ugyanis a standard normális eloszlás $\pm 2\sigma$ kritikus értékeinek, az ezen kívül álló értékek valószínűsége kisebb, mint 5%.) Ennél enyhébb szabályt is alkalmazhatunk: ha csupán a 3,29-nél nagyobb abszolútértékű standardizált értékeket szűrjük ki, akkor ezek valószínűsége kisebb, mint 1% (3,29 a standard normális eloszlás $\pm 3\sigma$-hoz tartozó kritikus értékének abszolútértéke). Mi 2,2 felett választottuk le a kiugró értékeket, feltételezve, hogy az SCP-értékek normális eloszlásúak. Az így választott küszöbérték feletti standardizált értékkel rendelkező SCP-érték valószínűsége 0,014. Ily módon 312-ből 18 extrém magas SCP-értéket zártunk ki.

5. Példa Többdimenziós kiugró értékek szűrése Mahalanobis-távolsággal

Többdimenziós esetben a kiugró értékeket nem lehet kiszűrni az előbbi példában látott, normális eloszlás esetén alkalmazható 1σ-2σ-3σ-szabály alapján. Egy többdimenziós kiugró érték bizonyos esetekben egyetlen dimenzióban sem egydimenziós kiugró érték.

A Mahalanobis-távolság (Mahalanobis, 1936) az Euklideszi-távolság egy többdimenziós általánosítása: egy n dimenziós ponthalmaz x pontjának a ponthalmaz $\bar{x}$ középpontjától vett távolságát oly módon definiálja, hogy azt a Σ kovarianciamátrixszal súlyozza:

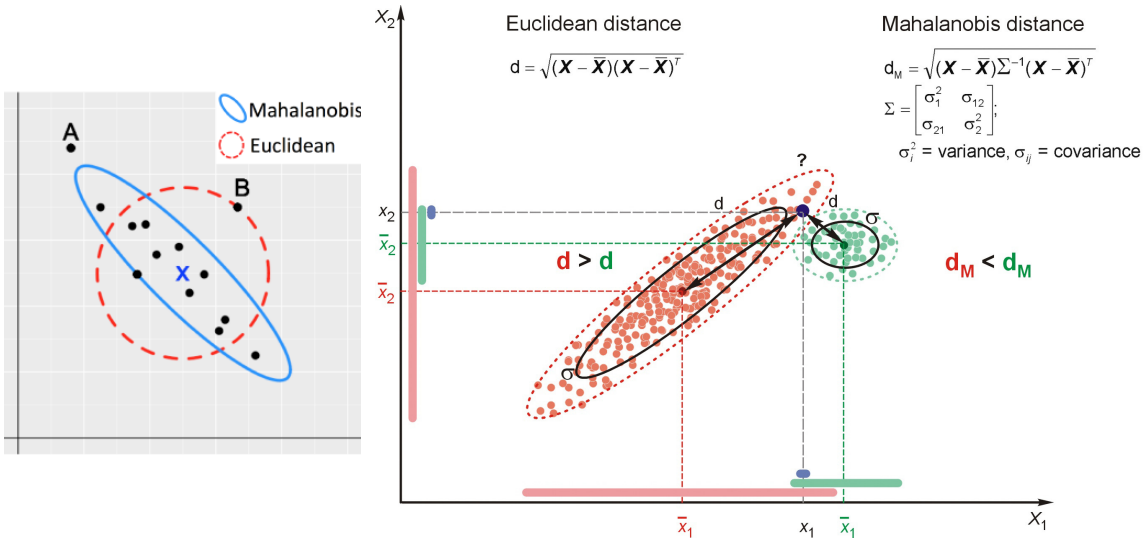
$$d^2_{Mahalanobis} = (x - \bar{x})^T \Sigma^{-1} (x - \bar{x}),$$

ezzel lényegében nemcsak a pontnak a középponttól való távolságát méri, hanem figyelembe veszi annak irányát is (2. Ábra). Minél nagyobb a Mahalanobis-távolság értéke, annál ‘szokatlanabb’ az adatpont (azaz annál valószínűbb, hogy többdimenziós kiugró értékről van szó).

A 2. Ábra bal paneljén a Mahalanobis-távolság kritériumként való alkalmazásával a kék görbén belüli pontok nem kiugró értékek. A B pont ezen kívül esik, tehát a Mahalanobis-távolság szerint kiugró érték, pedig Euklideszi-távolsága kisebb, mint a piros

körön kívüli, de a kék görbén belüli, a Mahalanobis-kritérium szerint nem kiugró pontok Euklideszi-távolsága.

A 2. Ábra jobb paneljén egy példát láthatunk. Itt a kérdőjellel jelölt fekete pontot vizsgáljuk a Mahalanobis-távolság alapján, hogy vajon kiugró érték-e egyrészt mint a piros, másrészt mint a zöld halmaz tagjaként egy szigorúbb (sötét fekete görbe), másrészt egy kevésbé szigorú (szaggatott piros, illetve zöld vonalak) feltétel szerint. (Egydimenziós értelemben sem a piros, sem a zöld halmaz egyik dimenziójában sem kiugró.) A zöld halmaz szórása mindkét dimenzióban kisebb, mint a pirosé. A kevésbé szigorú feltétel szerint a kérdéses fekete pont a piros halmazban nem kiugró, a zöldben azonban az ($d_{\mathrm{M(zöld)}} > d_{\mathrm{M(piros)}}$), holott Euklideszi távolsága a piros halmaz középpontjától messzebb van, mint a zöld halmaz középpontjától ($d_{\mathrm{zöld}} = d_{\mathrm{Euklideszi(zöld)}} < d_{\mathrm{Euklideszi(piros)}} = d_{\mathrm{piros}}$). A Mahalanobis-kritérium azért nem tekinti a fekete pontot kiugrónak a piros halmazban, mert a fekete pont irányában a piros halmaznak nagyobb a szórása, amit figyelembe vesz, azaz ott a pont helyzete nem olyan szokatlan.



2. Ábra A kétdimenziós ponthalmaz középpontjától mért Euklideszi- és Mahalanobis-távolság sematikus ábrája (bal) és egy példa (jobb panel). (Forrás: Ou Zhang url)

A Mahalanobis-távolság n szabadsági fokú χ^2-eloszlást követ, így a szignifikáns kiugró értékek egyszerűen szűrhetőek. Ökölszabályként a $p < 0{,}001$ esetben tekintünk egy pontot kiugrónak (Tabachnick és Fidell, 2019).

A Mahalanobis-távolsággal való kiugró értékek szűrését például nyolc paradicsom (*Solanum lycopersicum* L.) tájfajta ('Balatonboglár', 'Cegléd', 'Fadd', 'Máriapócs', 'Mátrafüred', 'Tarnaméra', 'Tolna megye', 'San Marzano'), két éven át (2015 és 2016) felvételezett bioaktív vegyületeinek (összes oldható szárazanyag-tartalom, összes savtartalom, antioxidáns-kapacitás változók (polifenol, FRAP, DPPH, ABTS és CUPRAC); színparaméterek (h°: színezet és C*: króma), valamint enzimaktivitási paraméterek (GST: glutation-S-transzferáz, AsPOX: aszkorbát-peroxidáz és POX: peroxidáz) elemzésénél alkalmaztam (Divéky-Ertsey és mtsai, 2022).

Tájfajtákról lévén szó nem lepődtünk meg azon, hogy az enzimaktivitási paraméterek nagyfokú változékonyságot mutattak a termőhelyektől és az évjárattól függően, de akár egyetlen termőhelyen és évjáratban is, ezért a kiugró értékek szűrése elsődleges feladat volt.

Egy másik példa a többváltozós kiugró értékek szűrésére Mahalanobis-távolsággal a 6.6. fejezetben található.

Most arra mutatok két példát, amikor az elemzés előtt az adatok transzformációjára volt szükség.

6. Példa Az adatok normálására a [0,1] zárt intervallumra

Ha bizonyos objektumokat (egyedeket) több változóval írunk le, és ezeknek az objektumoknak a karakterét szeretnénk megismerni, akkor sok esetben nem az egyes változók nominális értékeire vagyunk kíváncsiak, hanem arra, hogy annak változónak az értékei önmagukban nagyok vagy kicsik. Ha ezeknek a változóknak különböző a nagyságrendje, akkor ilyenkor célszerű először valamilyen alkalmas normálással összehasonlíthatóvá tenni azokat. Erre egy példa lehet az X változóra alkalmazott

$$(X - min(X))/(max(X) - min(X))$$

transzformáció, amelynek hatására minden változót a [0,1] zárt intervallumra képezünk úgy, hogy a változó értékei között az arányok nem változnak. Így minden változó legmagasabb értéke 1 lesz, a legalacsonyabbaké pedig zérus, azaz az összes komponens összehasonlítható válik függetlenül azok nagyságrendjétől. Erre mutatok egy példát a 6.6. fejezetben, amikor elemzésünk végén a különböző nagyságrendű változókat egy pókhálóábrán kívántuk megjeleníteni, hogy lássuk, hogy az összehasonlítani kívánt objektumok az egyes változókban kicsi vagy nagy értékeket vesznek fel (Nyitrainé és munkatársai, 2021).

7. Példa Box-Cox transzformációra

A statisztikai módszerek közül számos megkövetel valamilyen normalitási feltételt. Ha egy n dimenziós X (folytonos) változó normalitása sérül, azt könnyen észrevehetjük, ha a változó ferdeségét (*skewness*) és csúcsosságát (*kurtosis*) képezzük, melyek az eloszlás (sűrűségfüggvényének) alakjáról nyújtanak információt. A ferdeség (S), illetve a csúcsosság (K) egy valószínűségi változó harmadik, illetve negyedik momentumával definiálható (Joanes és Gill, 1998):

$$S = \frac{\sum(X-\bar{X})^3/n}{\sigma^{3/2}}\ ; \qquad K = \frac{\sum(X-\bar{X})^4/n}{\sigma^2} - 3, \qquad ahol\ \sigma = \frac{\sum(X-\bar{X})^2}{n},$$

melyeket a mintából az alábbi módon becsülhetünk:

$$\tilde{S} = \frac{\sqrt{n(n-1)}}{n-2}S; \qquad \tilde{K} = \frac{n-1}{(n-2)(n-3)}\big((n+1)K + 6\big).$$

A normális eloszlás ferdesége és csúcsossága zérus, ezzel hasonlíthatjuk össze az x változó ferdeségét és csúcsosságát. Minél nagyobb az n mintaelemszám, annál nagyobb abszolútértékű ferdeségű, illetve csúcsosságú változót fogadhatunk el még normálisnak, de ez függ attól is, hogy a módszer, amelyet használni szeretnénk, mennyire érzékeny erre a feltételre. Ökölszabályként a ferdeség, illetve csúcsosság abszolútértékére elvárhatjuk, hogy azok 1, illetve 2 alatt legyenek, de nagy mintaelemszám esetén 2, illetve 4 is elfogadható még (D'Agostino, 1971, D'Agostino és Pearson, 1973, Pearson és mtsai, 1977, West és mtsai, 1995, Field, 2024a,b, Kim, 2013, Tabachnick és Fidell, 2019).

Ha sérül a normalitás, egy megoldás lehet az adatok transzformálása egy olyan megfelelő monoton függvénnyel, amely az eloszlás ferdeségét, illetve a csúcsosságát egy elfogadható tartományba hozza. Gyakran alkalmaznak például logaritmustranszformációt, amely a jobbra ferde (pozitív ferdeségű) eloszlást javítja. Arányadatokra például, amelyek

0 és 100 közötti értékeket vesznek fel, mivel azok eloszlásának sűrűségfüggvénye jobb és bal oldalon sok esetben le van vágva, az $arcsin(\sqrt{x/100})$ függvény is megoldás lehet. Box és Cox (1964) azt javasolták, hogy képezzük az X változó megfigyelt értékeinek λ valós paramétertől függő $x_i(\lambda)$ transzformáltjait az

$$x_i(\lambda) = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda}, ha\ \lambda \neq 0 \\ ln(x_i), ha\ \lambda = 0 \end{cases} (i = 1, 2, \dots n)$$

módon, és tekintsük az

$$f(x, \lambda) = -\frac{n}{2} ln \left[\sum_{i=1}^{n} \frac{\left(x_i(\lambda) - \overline{x_i(\lambda)}\right)^2}{n} \right] + (\lambda - 1) \sum_{i=1}^{n} ln\ (x_i(\lambda))$$

függvény λ-ban vett maximumát (ahol $\overline{x(\lambda)}$ az $x_i(\lambda)$ értékek átlaga). Ezzel a λ paraméterrel lesz a fenti transzformáció optimális olyan értelemben, hogy a legsikeresebben normalizálja az X változót.

Bodor-Pesti és munkatársai (2013) 'Sauvignon blanc' és 'Syrah' szőlőkből gyűjtött levélminták ampelometriai elemzése céljából 22 biometrikus paramétert használtak. A paraméterek a szőlőlevelek egyes jól meghatározható részeinek hosszából, illetve az ezek által bezárt szögekből, valamint a szőlőlevél területéből álltak. A 22 paraméter közül 18 az OIV-nek (International Organisation of Vine and Wine) a szőlőlevelek ampelometriai leírásáról szóló hivatalos ajánlásában szerepel (O.I.V., 2009a,b). Az elemzés célja az volt, hogy ezek közül a paraméterek közül meghatározzák a fajtákra jellemző azon paramétereket, amelyek a különböző rügyterhelés mellett is stabilitást mutatnak. A 22 paraméter többségének normalitása nem sérült, ám némelyek esetében a ferdeségük vagy csúcsosságuk ellensúlyozására transzformációra volt szükség. Box és Cox fent leírt módszerével különböző optimális λ paramétert használtam az egyes változók esetén.

Jegyezzük meg, hogy nem csak az optimális λ paraméterrel való transzformáció lehet megfelelő. Általában az optimális λ közelében lévő paraméterekkel való transzformáció is elegendően módosítja az eloszlás alakját. Ez azért fontos, mert olyan változókkal is lehet dolgunk, amelyeket egymás között is össze kívánunk hasonlítani (pl. ilyen lehet egy ismételt mérésekből álló kísérlet adathalmaza, amikor a faktorhatás(ok)on kívül a különböző időpontokban mért értékeket is össze kívánjuk vetni). Ilyen esetben természetesen azonos λ paramétert kell használnunk, és ekkor a különböző optimális λ paraméterek közül konszenzusos módon választjuk ki azt, amelyik mindegyik transzformálni kívánt változóra alkalmas lehet.

A következő példa arról szól, hogy az elemzés előtt arról kell döntést hoznunk, hogy bizonyos adatok homogénnek tekinthetőek-e, illetve összevonhatóak-e.

8. Példa arra, amikor kérdéses, hogy az adatok összevonhatóak-e

Király és munkatársai (2022) hagymatripszek (*Thrips tabaci*) három változatának (L1, L2, T) párzási szokásait vizsgálták különféle változatpárok viselkedésének megfigyelésével laboratóriumi körülmények között.

Itt az adat-előkészítés során az a probléma merült fel, hogy bár mindig törekedtek arra, hogy olyan egyedeket állítsanak párba, amelyek korábban még nem találkoztak más egyeddel, mégis előfordult, hogy – megfelelő egyedek szűkében – a párosításokban olyan egyedeket is kénytelenek voltak használni, amelyekben a pár egyik vagy mindkét példányát

korábban már használták egy keresztezéses párosításban (de a korábbi találkozásokkor párzás nem történt). Ezért, mielőtt statisztikailag együtt kezeltük volna ezeket az eseteket, meg kellett vizsgálnunk, hogy azok a párosítások, amelyekben mindkét példányt először használták, viselkedésükben különböznek-e azoktól, amelyekben az egyik vagy mindkét példányt másodszor használták. E két csoport összehasonlításához három Fisher-féle egzakt tesztet futtattam le, egyenként 2 × 2 kontingenciatáblára, azon párosítások számával, amelyek (I) kölcsönhatásba léptek vagy sem; (II) sikeresen párosodtak (az összes párosítást figyelembe véve) vagy sem; és (III) sikeresen párosodtak (csak azokat a párosításokat figyelembe véve, amelyekben kölcsönhatás történt) vagy sem. Mivel szignifikáns különbségeket nem lehetett kimutatni (Fisher-féle egzakt teszt, $p > 0{,}05$, 1. Táblázat), ezért arra a következtetésre jutottunk, hogy egy másik kifejlett tripsz egyeddel való sikeres párosodás nélküli korábbi találkozás nem befolyásolja a tripszek későbbi párzási viselkedését, ezért ezeket az adatokat összevonhatjuk.

1. Táblázat A T♀ + T♂, L1♀ + L1♂, vagy L2♀ + L1♂ hagymatripsz (*Thrips tabaci*) párok esetében a párzási viselkedéssorozat egyes lépéseit teljesítő párok száma és százalékos aránya (I. kölcsönhatásba lépés; II. sikeresen párosodtak az összes párosítást figyelembe véve; (III) sikeresen párosodtak csak azokat a párosításokat figyelembe véve, amelyekben kölcsönhatás történt) két körülmény esetén. A: mindkét egyedet először párosították; B: a pár egyik, vagy minkét egyede korábban már párosításban találkozott más egyeddel. A p a Fisher-féle exact teszttel való összehasonlításkor kapott szignifikancia érték.

		Párzási viselkedéssorozat lépése					
		I.		II.		III.	
Párok		arány	n	arány	n	arány	n
T♀ + T♂	A	75%	12	67%	12	89%	9
	B	50%	30	47%	30	93%	15
	p	0,128		0,204		0,620	
L1♀ + L1♂	A	90%	10	80%	10	89%	9
	B	70%	23	65%	23	94%	16
	p	0,212		0,339		0,600	
L2♀ + L1♂	A	80%	15	53%	15	67%	12
	B	77%	26	54%	26	70%	20
	p	0,572		0,614		0,573	

A következő példával olyan esetet mutatok, amikor az adatokat laboratóriumi méréseken alapuló szimulációval állítottam elő.

9. Példa Monte Carlo szimuláció alkalmazására

Forgács és munkatársai (2017) *V. berlandieri x V. rupestris* 'Richter 110' alanyból származó embriogén kalluszkultúrákat képeztek és tartottak fenn nyolc különböző táptalajon. Négy hét elteltével a tenyészeteket különböző sűrűségűre hígították (2. Táblázat) és továbbszaporították. Két hónap elteltével az egyes lemezeken lévő embriók összsúlyát és 4 × 100 embrió súlyát rögzítették.

Az embriók várható számának becslését úgy lehet összehasonlítani, ha 1,0 g embriogén sejtaggregátumra vonatkoztatva adjuk meg az összes hígításban kapott eredményeket. Ehhez arra volt szükség, hogy szimuláljuk azt a helyzetet, mintha az 1,0 g kezdeti embriogén sejtek teljes mennyiségére elvégezték volna a kísérletet.

2. Táblázat Sejtsűrűség (mg/mL) az embrió differenciálódása során kezdetben és a negyedik heti hígítást követően

		közeg									
		M1					M2				
sejtsűrűség (mg/mL)	kezdetben	0,5	1	2	4	8	0,5	1	2	4	8
	4 hét után	0,5	-; 0,5; 1; 2; 4; 8	2	4	1; 8	0,5	1	2	4	8

Ahhoz tehát, hogy meg tudjuk becsülni az embriók várható számát 1,0 g embriogén sejtaggregátumra vonatkoztatva, Monte Carlo szimulációt végeztem (Fishman, 1996). A Monte Carlo-módszer a bemeneti adatok alapján nagyszámú véletlenszám-generálással modellezi a vizsgált rendszert. Nagyszámú futtatásának eredményeképpen becsülhetjük a modellezett rendszer paramétereit. Az embriók 12 ismétlésben megszámolt számából képzett átlagok és szórások, valamint a kezdeti és a negyedik heti hígítási sűrűség különböző arányai mint szimulált mintaelemszámok (N) alapján 100 000 alkalommal generáltam normális eloszlású véletlen mintákat. A szórásokra korrekciót alkalmaztam, amelyet az N mintaelemszámú és a 12-es mintaelemszámú minták szórásainak hányadosaként számoltam ki, mindkettőt ugyanabból az eloszlásból véve. Végül kiszámítottam a 95%-os konfidenciaintervallumot az 1,0 g embriogén sejtre vonatkozó várható embriószámra vonatkozóan a táptalaj és a sűrűség páros minden egyes esetére.

2.2 A modellparaméterek becslése (kalibráció) és a modell validációja

A modellek építése során azok paramétereit becsüljük (kalibráció) valamilyen módszerrel. Ezután a paraméterek jóságát teszteljük, illetve a modelleket validáljuk. Ezek a lépések különböző modelltípusokra vonatkozóan más-más módon történnek. A validáláció során azt vizsgáljuk, hogy a modell kimenete mennyire megbízhatóan, illetve pontosan illeszkedik az adatokra. A modellek validációja lehetőség szerint olyan adatokon történik, amelyek a modell kalibrálása folyamán nem kerültek felhasználásra, bár ez nyilván nagy mintaelemszámot igényel.

A validálás egy gyakori, a mintaelemszámmal spórolósabban bánó módszere, az ún. keresztvalidáció (*cross-validation,* CV). A k-szoros keresztvalidáció (*k-fold cross-validation*) úgy működik, hogy az adatokat k (nagyjából) egyenlő részekre osztva $k-1$ rész használunk a kalibrációhoz, egy részt a validáláshoz. Ha k a mintaelemszámmal egyenlő, akkor az ún. egyet kihagyó (*leave-one-out*) keresztvalidációról beszélünk (Hastie és mtsai, 2013, Lantz, 2015, Ghatak, 2017, James és mtsai, 2013). A másik remek módja a mintaelemszámmal való spórolásnak, az újra-mintavételezés, vagy minta-a-mintából (*resampling*). A *bootstrap*-módszer (Efron, 1979, Efron és Tibshirani, 1973, 1997, James és mtsai, 2013) lényege, hogy a mintából a mintáéval megegyező elemszámú véletlen visszatevéses mintát veszünk, és erre végezzük a kalibrálást, a maradék mintaelemeket (*out-of-bag*, OOB, Breiman, 1996b) validálásra használjuk. Az eljárást sokszor megismételve az egyes elemek torzító ereje, illetve maga a modell torzítása is csökken.

A maximum likelihood függvény egy feltételes valószínűségi függvény, ami a megfigyelt D adat valószínűségét fejezi ki az eloszlás P paraméterhalmazát, illetve az adatokra illesztett M modellt feltételezve: $L(P,M) = P(D|P,M)$. A maximum likelihood becslés lényege, hogy az a legjobb becslés a paraméterhalmazra, amely esetében ez a függvény maximális (Hastie és mtsai, 2013).

2.3 A modellek értékelése

A modellek értékelésére a legfontosabb modellek ismertetése után térünk ki (4.6. fejezet), mivel az értékelés módszere a modell típusától függ.

2.4 A modellek összehasonlítása

A modellek összehasonlítására általánosan alkalmazott módszerek, az Akaike információs kritérium (*Akaike information criterion*, AIC, Akaike, 1973) és a Bayes-i információs kritérium (*Bayesian information criterion*, BIC, Schwarz, 1978) úgy hasonlítja össze a modelleket, hogy a paraméterek k számától, illetve az N mintaelemszámtól függően az L maximum likelihood függvény segítségével büntet (*penalty term*):

AIC	Kis mintaelemszámra módosított AIC	BIC
$AIC = -2ln\,(L) + 2k$	$AIC_c = AIC + \frac{2k(k+1)}{N-k-1}$	$BIC = -2ln\,(L) + kln(N)$

Azon az elven alapszik, hogy két (közel) egyformán jó modell közül az egyszerűbb (a kevesebb paraméterű) a jobb, azaz az alacsonyabb AIC és BIC értékkel rendelkező modell a jobb. A BIC valamivel jobban bünteti a modelleket a további paraméterekért, mint az AIC (Ghatak, 2017, James és mtsai, 2013).

A modellek összehasonlításának további módszereire egyes modellek tárgyalásakor még visszatérünk (5.1.5. fejezet, 16. Példa, 6.5. fejezet).

2.5 A statisztikai eredmények közlése alkalmas vizualizációval

Egy példát már láttunk arra, hogy bizonyos bonyolult összefüggések jól szemléltethetőek hőtérképpel (3. Példa).

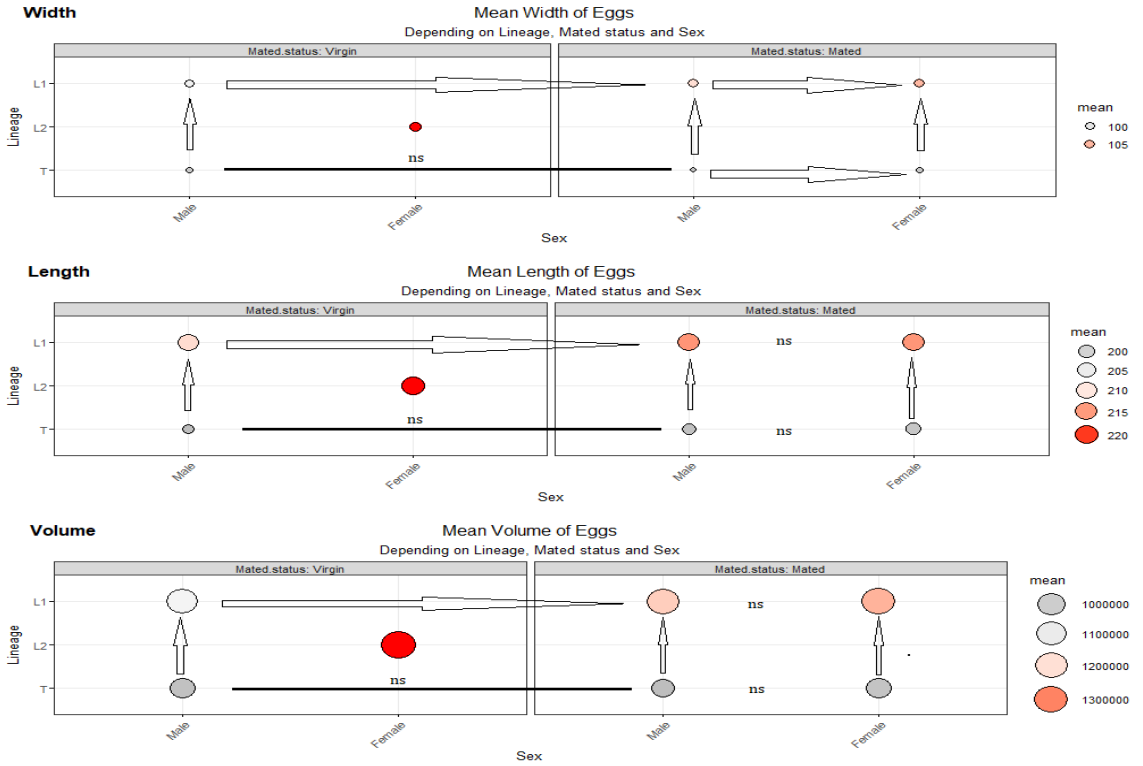
Az alábbiakban három további példát mutatok be olyan esetre, amikor a kézirat tervezésekor az eredmények áttekinthetetlennek tűntek, ezért azok illusztrációja fontos lépés volt a világos közléshez.

10. Példa egy vizualizációra speciális táblázattal

Ebben a példában Musa és munkatársai (2023) eredményeinek egy vizualizációját mutatjuk be. L1, L2 és T dohánytripsz fajtakomplex változatok (*Thrips tabaci* Lindeman; Thysanoptera: Thripidae) nőstényeinek petéit vizsgálták azok morfometriai tulajdonságai szerint. Az L1 és a T változatok arrenotokiával szaporodnak (Farkas és mtsai., 2019; Toda és Murai, 2007). A szűz arrenotok nőstények 100%-ban hím petéket raknak, a párosodottak hím és nőstény egyedeket is hoznak létre. Az L2 vonal telitok szaporodású (Kobayashi és Hasegawa, 2012): az utódok 100%-a szűzen nemzett nőstény. A peték vizsgált morfológiai tulajdonságai (amelyek azok életképességére is utalhatnak) függenek attól, hogy a peterakó nőstény melyik változatba tartozik (L1, L2 és T), szűz, vagy párosodott-e előtte, illetve, hogy hím, vagy nőstény petét rakott-e. Ahhoz, hogy a morfológiai különbségeket ezek mentén a feltételek mentén áttekinthetővé tegyük, olyan ábrát terveztem, amely összefoglalja az összehasonlítások szignifikáns és nem szignifikáns eredményeit (3. Ábra).

A 3. Ábrán a *T. tabaci* L1, L2 és T változataiba tartozó szűz és párosodott anyák hím és nőstény petéinek szélessége (µm), hosszúsága (µm) és térfogata (µm^3) (átlag ± szórás) látható. N a vizsgált egyedek száma. Az L2 anyák szignifikánsan szélesebb, hosszabb és nagyobb petét raktak az összes többi petéhez képest. Jellemzően, ahol

szignifikáns különbséget találtunk, ott az L1 anyák nagyobb petéket raktak, mint a T anyák, a párosodott anyák nagyobb hím petét raktak, mint a szűzek, illetve nagyobb nőstény petét raktak, mint hím petét.

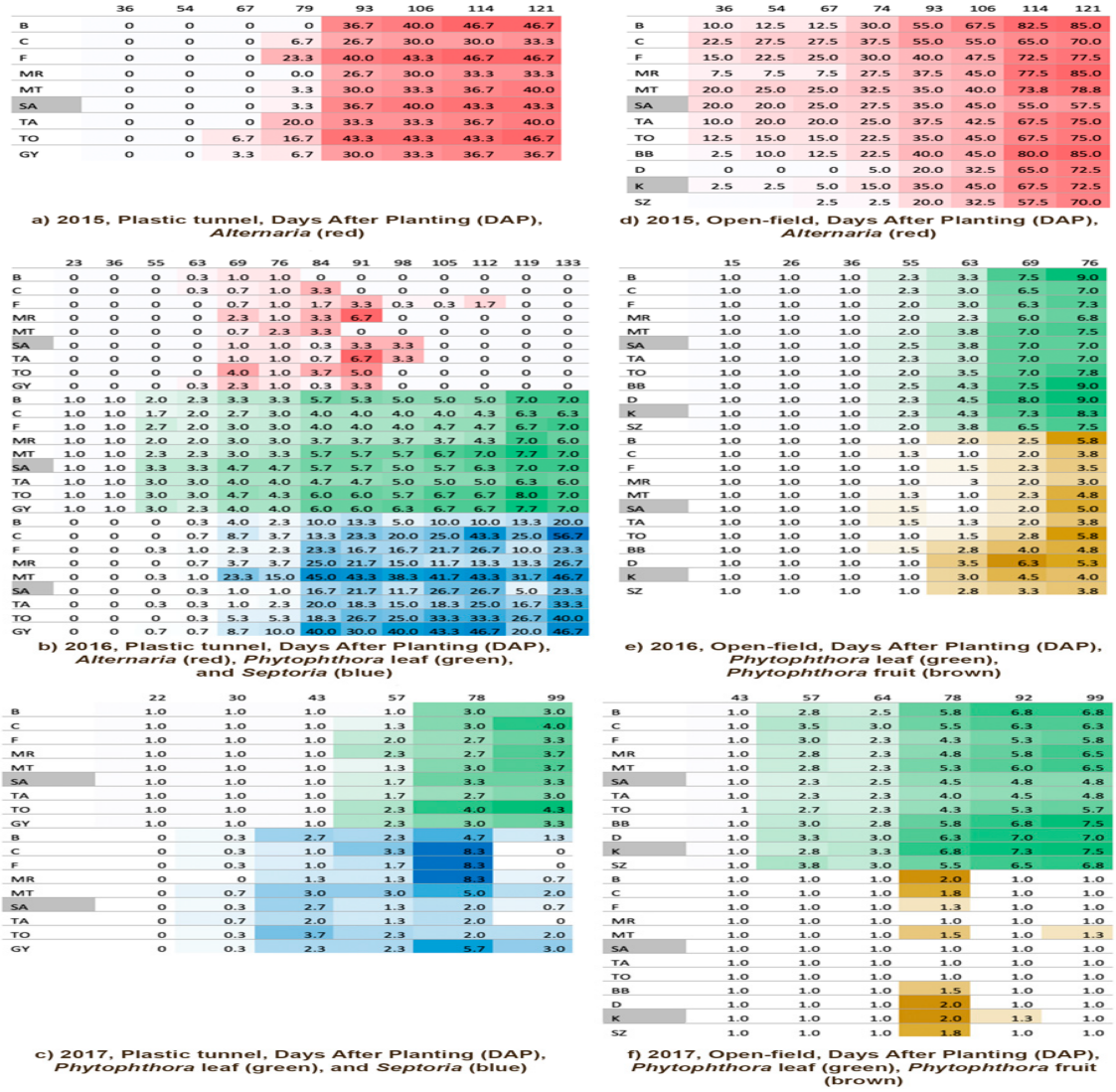


		Szűz		Párosodott	
	Nem / Változat	Hím (átlag ± szórás)	Nőstény (átlag ± szórás)	Hím (átlag ±s zórás)	Nőstény (átlag ± szórás)
Szélesség (µm)	L1	101,16±5,34 (*N*=598)		103,68±4,92 (N=91)	105,00±4,31 (N=119)
	L2		**109,26±3,97** (*N*=889)		
	T	96,85±4,08 (*N*=1558)		95,80±3,60 (N=112)	96,61±3,47 (N=325)
Hossz (µm)	L1	210,79±7,86		215,33±8,24	215,25±7,49
	L2		**221,10±7,23**		
	T	195,88±7,95		196,50±8,11	197,54±7,77
Térfogat ($10^*µm)^3$	L1	1132,80±125,92		1215,78±133,76	1245,54±120,56
	L2		**1384,58±117,83**		
	T	964,03±93,15		94,611±85,65	966,45±77,24

3. Ábra A *T. tabaci* L1, L2 és T változataiba tartozó szűz és párosodott anyák hím és nőstény petéinek szélessége (µm), hosszúsága (µm) és térfogata ($µm^3$) (átlag ± szórás). *N* a vizsgált egyedek száma. Az L2 anyák szignifikánsan szélesebb, hosszabb és nagyobb petéinek átlagai (az összes többi petéhez képest) vastagon szedve. A nyilak a szignifikánsan nagyobb értékek irányába mutatnak ($p < 0{,}05$). A vízszintes szakaszok a nem szignifikánsan eltérő értékeket jelölik ($p > 0{,}05$).

11. Példa az eredmények vizualizációjára hőtérképpel

Ebben a példában Boziné-Pullai és munkatársai (2021) eredményeinek egy vizualizációját mutatom be. Bizonyos esetekben az eredmények csupán abból adódóan nehezen áttekinthetőek, hogy szabadföldön és fóliasátorban is, több évben (2025, 2026, 2017), sok fajtán ('Balatonboglár', 'Cegléd', 'Fadd', 'Gyöngyös', 'Máriapócs', 'Mátrafüred', 'Tarnaméra', 'Tolna megye', 'San Marzano', 'Dány', 'Szentlőrinckáta', 'Kecskeméti 549') többféle fertőzést (*Alternaria solani* Sorauer, *Phytophthora infestans* M. de Bary) levélen és gyümölcsön, illetve *Septoria lycopersici* Speg.) vizsgáltak ismételt megfigyelések (6–13 alkalommal) alapján. Az idézett publikációban paradicsom (*Solanum lycopersicum* L.) tájfajtákról volt szó, és a statisztikai összehasonlítások eredményeit a 4. ábrán látható módon egy egyszerű hőtérképpel tettem követhetővé.

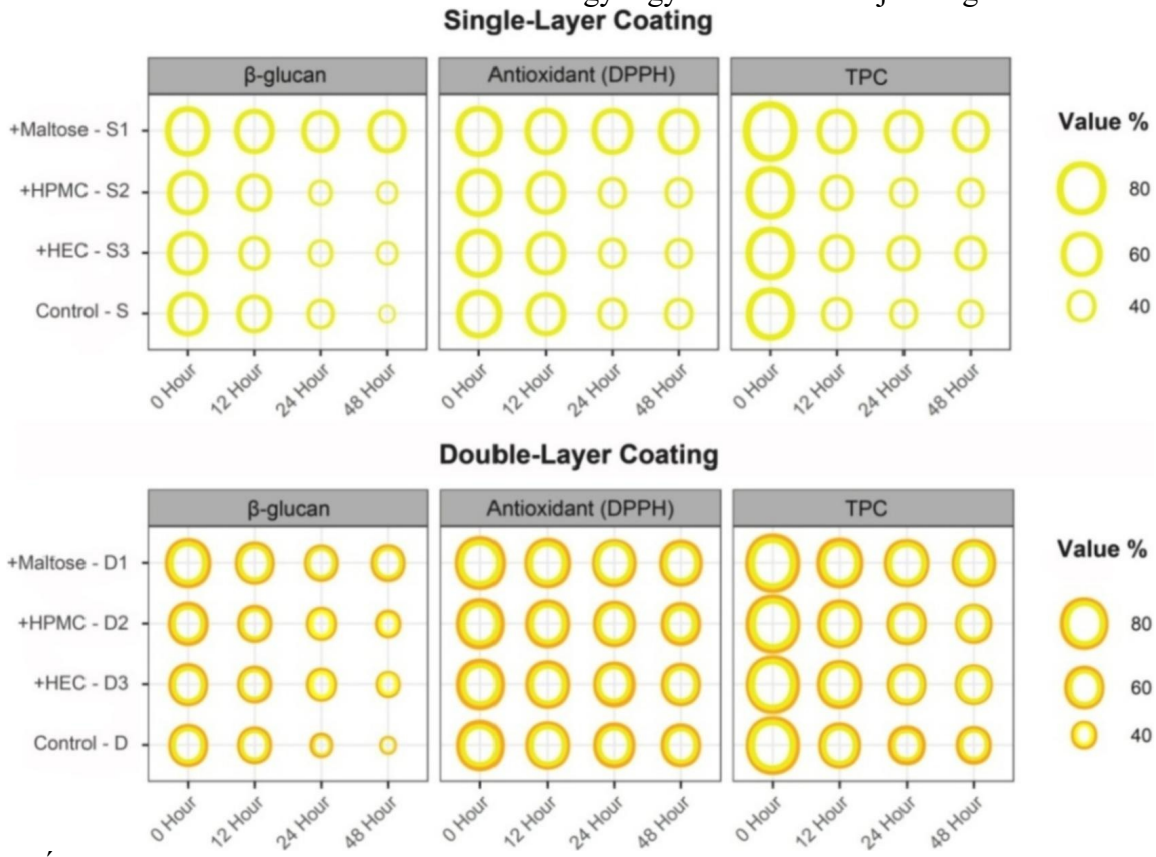


4. Ábra Paradicsom (*Solanum lycopersicum* L.) tájfajták *Alternaria* (piros), *Phytophthora* levélen (zöld), gyümölcsön (barna) és *Septoria* (kék) fertőzöttségének átlagos százaléka 2015-ben (első sor), 2016-ban (második sor), 2017-ben (harmadik sor) fóliasátorban (bal oldal) és szabadföldön (jobb oldal) hőtérképen ábrázolva. B: 'Balatonboglár'; C: 'Cegléd'; F: 'Fadd'; GY: 'Gyöngyös'; MR: 'Máriapócs'; MT: 'Mátrafüred'; TA: 'Tarnaméra'; TO: 'Tolna megye'. A kontroll fajta: SA: 'San Marzano'. Determinált génbanki tételek: BB: 'Balatonboglár'; D: 'Dány'; SZ: 'Szentlőrinckáta'. A kontroll fajta: K: 'Kecskeméti 549'. A kontrollfajták szürke háttérszínnel vannak kiemelve.

12. Példa arra, amikor az eredményeket a minták sematikus ábrázolásával mutattuk be

Ebben a példában Mirmazloum és munkatársai (2021) eredményeinek egy vizualizációját ismertetem.

A kutatás lényege az volt, hogy probiotikus *Lactobacillus acidophilus* sejteket *Ganoderma lingzhi* gombakivonattal együtt kapszulázták, ezekből egy- és kétrétegű kalcium-alginát géllel bevont gyöngyöket képeztek azzal a céllal, hogy a sejtek életképességét meghosszabbítsák a tárolási idő alatt. Különböző reagensekkel végzett kísérleteket követően a maltóz, a hidroxietil-cellulóz (HEC), valamint a hidroxipropil-metilcellulóz (HPMC) felhasználásának eredményeit kívánták egyszerűen ábrázolni a kontrollal való összehasonlításban. Azt kívánták megmutatni, hogy a reagensek beépítése hogyan csökkentette a gombafenolok, antioxidánsok és β-glükán felszabadulási sebességét a tárolási idő alatt 12, 24 és 48 órát követően egy-, illetve kétrétegű bevonat alkalmazásával. Ezt a feladatot a gyöngyök sematikus ábrázolásával (5. Ábra) oldottam meg, ahol minél kisebb a gömb, annál nagyobb a kapszulázott molekulák diffundált mennyisége. Az ábráról könnyen leolvashatóak a publikációban statisztikailag igazolt legfontosabb eredmények: (1) Az antioxidánsok és a β-glükán felszabadulási aránya jelentősen csökkent a kapszulázást követően; (2) A maltóz alkalmazása mutatta a legjobb tartósító hatást; (3) A HPMC és a HEC alkalmazása nem fokozta a gyöngyök védelmi tulajdonságát.



5. Ábra A β-glükánnak, az antioxidánsoknak és az összes fenoloknak (TPC) az egy- és kétrétegű kalcium-alginát gélbevonatú, maltóz-, hidroxietil-cellulóz- (HEC), valamint hidroxipropil-metilcellulóz- (HPMC) tartalommal kiegészített gyöngyökbe való kapszulázással megtakarított mennyisége a tárolási időszak alatt (12, 24 és 48 óra elteltével) a kezeletlen kontrollhoz képest. Minél kisebb a gömb, annál nagyobb a kapszulázott molekulák diffundált mennyisége.

3 LINEÁRIS MODELLEK

3.1 Az általános lineáris modellek

Az általános lineáris modell (*General Linear Model*, GLM) legegyszerűbb formában úgy írható fel, ahogy azt Rutherford (2011) fogalmazta: Adat = Modell + Hiba. Ezt kibontva az egyenlet bal oldalán a modellezni kívánt $Y = (Y_1, Y_2, \dots, Y_l)$ függő válaszváltozó áll, a jobb oldalán pedig egy $X = (X_1, X_2, \dots, X_k)$-szel jelölt, $k \geq 1$ számú magyarázó változóból álló vektor, amelynek $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ paramétereit becsüljük. A jobb oldalon additív tag a modell normális eloszlású, zérus várható értékű ε véletlen hibája: $Y = \beta X + \varepsilon$ (Eye és Wiedermann, 2023).

Ha az egyetlen tagból álló $Y = (Y_1)$ folytonos függő változó, valamint $X = (X_1)$ is egyetlen tagból álló diszkrét (faktor-) változó akkor egytényezős varianciaelemzésről beszélünk (ANOVA, *one-way ANOVA,* alkalmazására egy példa a 6.5. fejezetben). Ha a modellben egynél több (fix) faktor szerepel (azaz X egy faktorváltozókból álló k dimenziós vektor, $k > 1$), akkor a faktorok interakcióját is mérjük, és a modellt többtényezős (faktoriális) ANOVA-nak nevezzük. Ha a modellben egyetlen folytonos $X = (X_1)$ (magyarázó) változó van, akkor egyszerű regresszióról (*simple linear regression*), ha egynél több folytonos magyarázó változó van a modellben, akkor többszörös regresszióról (*multiple linear regression*) van szó.

Ha pedig a folytonos Y függő változót olyan modellel becsüljük, amelyben a faktor változó(k) mellett folytonos (kovariáns) változó(k) is szerepel(nek), akkor a modellt ANOCOVA-nak nevezzük. Ilyenkor a modellünk akkor ad torzítatlan becslést, ha a különböző faktorszinteken a kovariáns változó(k) meredeksége azonos, azaz a kovariáns változó és a faktorváltozó interakciója nem szignifikáns.

Ha pedig egynél több összefüggő, folytonos válaszváltozóra szeretnénk modellt illeszteni (azaz $l > 1$), akkor (kizárólag) diszkrét magyarázó változó esetén MANOVA (*Multivariate ANOVA*), kizárólag folytonos magyarázó változó esetén többváltozós regressziós modellünk lesz. Természetesen megfogalmazhatunk többtényezős MANOVA, illetve többszörös többváltozós regressziós modellt is.

Az általános lineáris modell linearitása nem feltétlenül jelenti a magyarázó változó és a függő változó lineáris kapcsolatát. Az X és/vagy az Y változók nemlineáris transzformálásával sok esetben ugyanis elérhetjük, hogy a transzformált változók között már lineáris kapcsolatot tudunk igazolni. Ezért megkülönböztetünk lineárisra visszavezethető és nem visszavezethető nemlineáris regressziót. Lineárisra nem visszavezethető esetben a függő változó és a modellparaméterek közötti kapcsolat nem lineáris.

13. Példák MANOVA alkalmazására

MANOVA modellre gyakran van szükség agrártudományi alkalmazásokban. Csak három példát mutatunk ezek közül.

Musa és munkatársai (2022) a dohánytripsz (*Thrips tabaci*, Lindeman) két változatát (L1, T) vizsgálták. A kutatás egyik kérdése az volt, hogy vajon a petesejt mérete határozza-e meg a megtermékenyülés valószínűségét, vagyis, hogy a petesejtek mérete a megtermékenyítés előtt már éppúgy különbözhet, mint utána, vagy a petesejtméret nem különbözik a szűz és a párosodott anyáknál, hanem a megtermékenyítés ténye növelheti a petesejtméretet.

Hogy erre válasz kapjunk, összehasonlítottuk a szűz és a párosodott anyák által rakott peték méreteit három paraméterük (szélesség, hosszúság, térfogat) mint folytonos

válaszváltozók tekintetében. Az összehasonlítást egytényezős MANOVA modellel végeztem, egyetlen faktornak két szintje az anyák szűz vagy párosodott állapota volt.

Woldemelak és munkatársai (2021) szintén dohánytripszek változataival foglalkoztak laboratóriumi körülmények között. A hőmérsékletnek (15, 23 és 30 °C) mint faktornak a peterakást megelőző időszak hosszára, a lerakott peték számára, illetve a peték kikelési arányára mint összefüggő válaszváltozókra való hatását egytényezős MANOVA-val elemeztem.

A MANOVA alkalmazására egy bővebben ismertetett példa a 6.1. fejezetben található.

3.2 Az általános lineáris kevert modellek

A modellbe fix faktor mellett ún. véletlen (*random*) faktort is beépíthetünk (Agresti, 2002). Amikor ugyanis kizárólag fix faktor van a modellben, akkor azt feltételezzük, hogy a különböző faktorszinteken azok hatásain kívül minden más tekintetben homogén körülmények között mértük a válaszváltozó értékeit. Ez a feltétel azonban sérülhet, ha az adatok más, nem mérhető körülménytől is függenek, például a különböző helyszínek egyéb hatásaitól, az évjárathatástól, a vizsgálatba bevont egyedek egyéni hatásától. Az ilyen faktort mint véletlen faktort illesztjük be a modellbe. Más tekintetben a kevert modellek megoldást jelenthetnek az ún. pszeudoreplikáció problémájára, amikor az adatokat függetlennek tekintjük, holott azok némely csoportja között összefüggés van (Freeberg és Lucas, 2009, Spurgeon, 2019, Arnqvist, 2020, Millar és Anderson, 2004). Azt a modellt, amelyben fix és véletlen faktor is szerepel, kevert modellnek (*General Linear Mixed Model,* GLMM) nevezzük.

A lineáris kevert modellek véletlen faktorai között megkülönböztetünk véletlen tengelymetszetet (*random intercept*) és véletlen meredekséget (*random slope*). A véletlen tengelymetszetet tartalmazó kevert modell a véletlen faktor különböző szintjeihez különböző konstans tagot, a véletlen meredekséget tartalmazó modell pedig különböző meredekséget tud becsülni. Ha véletlen tengelymetszetet és meredekséget is beteszünk a modellbe, akkor véletlenfaktor-szintenként mindkettőre egyéni becslést kapunk.

Ha a fix vagy véletlen faktorok valamilyen struktúrában egymásba ágyazódnak (*nested factors*), akkor többszintű (*multilevel/hierarchical)* modellről beszélünk.

14. Példa kevert, beágyazott modellre

Jelen példánkban Basky és munkatársai (2017) kutatásait idézzük. A parlagfű (*Ambrosia artemisiifolia* L.) nemcsak egy kellemetlen mezőgazdasági gyomnövény, amely gazdasági károkat okoz a mezőgazdasági terményekben, hanem – ami még fontosabb – súlyos allergén hatása miatt komoly egészségügyi problémákat is okoz. Az invazív parlagfű pollenje a városi környezetben különösen jelentős allergén kockázati tényezővé vált. Sok városi területen tilos a herbicidek kijuttatása, ezért a kaszálás a legszélesebb körben alkalmazott védekezési intézkedés.

A pollenszámlálás munkaigényes és az egészségre veszélyes munka, ezért a pollentermelési adatok főként becsléseken alapulnak. Ezért, hogy a becsléseknél pontosabb eredményeket kaphassanak, Basky és munkatársai (2017) által kézi számlálással előállított pollenadatokat elemeztünk kétéves szabadföldi kísérletek alapján. A vizsgálat a különböző kaszálási forgatókönyveknek (korai, kései, többszöri), illetve a növénysűrűségnek és a vágási magasságnak a közönséges parlagfű biomasszájára, pollentermelésére és magtermelésére vonatkozó együttes hatására terjedt ki.

A biomassza (g), a növénymagasság (cm) és a hím-, valamint nőivarú virágzatok száma mint összefüggő válaszváltozók elemzésére MANOVA modellt használtam, ahol a kezelés faktorváltozó szintjeit a különböző kaszálási forgatókönyvek képezték. Az év és azon belül a mintavétel dátuma egymásba ágyazott blokkok voltak. A pollenszámok elemzésére kevert modellt használtunk, ahol a véletlen faktor szintjei a növényegyedek voltak, amelyeknek virágjairól gyűjtötték és számlálták a polleneket.

15. Példa szakaszos lineáris kevert modellre

László és munkatársai (2018) a városi környezetben mérhető rendkívül alacsony frekvenciájú elektromágneses terek pulyka (*Meleagris gallopavo*) modellállatokra gyakorolt lehetséges biológiai hatásait vizsgálták, mivel stresszhelyzetben ezeknek a madaraknak az emberekhez hasonló élettani és viselkedési folyamatait írták le korábban.

A kezelés egyhetes adaptációval kezdődött (zéró kitettség), ezután háromhetes kezelés, majd öthetes regenerációs időszak következett.

A vizsgálatban a kezelt és a kontrollcsoportban megfigyelt különbségeket kívánták feltárni az idő függvényében. Az előző példához hasonlóan a pulykák egyéni hatását véletlen faktorként kellett kezelni, azaz lineáris kevert modellre volt szükség. A folyamat azonban egy jól felismerhető töréspontot tartalmazott, amikor a kezelés megszűnt, és elindult egy regenerálódási folyamat. Így egy speciális LMM modellt, az ún. szakaszos lineáris kevert modellt alkalmazták (*piecewise linear mixed model*, PLMM). A PLMM az általános lineáris kevert modell (GLMM) egy általánosítása, amelyben a fix és véletlen faktorok szegmensenként rugalmasan változhatnak. Olyankor használjuk, amikor az adatokon bizonyos töréspontok jelennek meg, például, ahogy a fenti példában, egy kezelés hatására, vagy olyankor, amikor a trendek változása bizonyos küszöbértékhez köthető (Muggeo, 2003, Naumova és mtsai, 2001).

A fenti modellek mindegyikét magába foglaló modellcsalád az általános lineáris modell (GLM), amelynek az általánossága abban áll, hogy a vizsgálatunk nem csupán a fix és/vagy véletlen faktorhatást, illetve a folytonos változó(k)tól való függést célozhatja meg. Egy összetettebb probléma jellemzően abban áll, hogy a válaszváltozót több (akár sok) nem feltétlenül egyforma típusú magyarázó változó együttesen magyarázza. Ekkor egy jól illeszkedő modellt keresünk úgy, hogy a modellben szereplő lehetséges változók közül keressük a legmegfelelőbbeket, miközben az azok közötti kapcsolatokat is figyelembe vesszük, és ezekre optimalizáljuk a modellparamétereket.

16. Példa kevert, beágyazott modellre

Palla és munkatársai (2021) a *Pinus sylvestris* L. fenyők anatómiai jellemzői (sejtfal vastagság, sejtátmérő, sejtszám) alapján vizsgálták a Keleti-Kárpátok és a Pannon-medence hat élőhelyéről származó populáció alkalmazkodását az ökológiailag szélsőséges viszonyokhoz.

Adathalmazunk tehát a sejtek anatómiai adatain kívül tartalmazta a termőhelyet, a sejt kialakulásának évét és az évgyűrűn belüli pozícióját, valamint a fa kódját (ID). Ezen kívül meteorológiai adatok is rendelkezésre álltak, amelyekből indikátorokat képeztünk: az adott év havi csapadékösszegeit és szárazságindexeit (áprilistól októberig: 2*7 változó), valamint az előző év havi csapadékösszegeinek és szárazságindexeinek négyhavi (januártól decemberig) csúsztatott átlagait (2*18 változó).

Az adathalmaz szerkezetét látva a fa a termőhelybe ágyazva (*nested*) egy olyan blokknak tekinthető, amelyet véletlen (random) hatásként kell figyelembe venni, ahogy a fákon belül (szintén beágyazva) a sejteknek az évgyűrűin belüli pozícióját is.

Ennek megoldására lineáris kevert modelleket (LMM) használtunk. A modellekben a helyszínt, a sejt kialakulásának évét és a meteorológiai indikátorokat fix faktorként, a helyszíneken belül a fákat, valamint a fákon belül a sejtek pozícióját véletlen (*random*) hatásként kezeltük.

Az időjárási változók száma túl sok volt ahhoz, hogy mindegyik bekerüljön a modellbe, ezért a másik két független változót (helyszín, év), valamint a két véletlen faktort (a sejt pozíciója, fa ID) mindig a modellben tartva a maradék független változókat úgy válogattuk be a beágyazott (*nested*) modellbe, hogy vettük azok összes lehetséges kombinációját, és likelihoodhányados-próbával (likelihood ratio test, LRT, Fisher, 1922, Neyman és Pearson, 1933) döntöttük el a legerősebb szignifikáns magyarázó erővel bíró kombinációkat.

Az LRT két modell, az egyszerűbb, ún. nullmodell M_0, illetve az M_0 paraméterein felül még több paramétert magában foglaló M_1 modellt hasonlítja össze. Először képzi a két modell feltételes valószínűségi, ún. likelihood függvényét, L_0-t és L_1-et (ld. 2.2. fejezet), majd ezekből az ún. likelihood hányadost ($LR = -2 * \ln(L_0/L_1)$, amely χ^2-eloszlást követ a két modell paramétereinek számának különbségével mint szabadsági fokkal. Minél nagyobb az LR, annál nagyobb a különbség a két modell között.

A vizsgált lineáris kombinációk (teljes modellek) minden egyes változóját akkor tekintettük szignifikánsnak, ha az adott változót nem tartalmazó redukált modell (nullmodell) szignifikánsan kevesebbet magyarázott, mint a teljes modell.

A legalacsonyabb p-értékek (azaz a legmagasabb $\chi 2$-értékek) alapján azokat a lineáris kombinációkat választottuk ki a végleges modell elemeinek, amelyeknél az összes redukált (null) modell LRT eredmény gyengébb volt.

3.3 Ismételt méréses ANOVA/MANOVA és marginális modellek

Ismételt méréses (*repeated measures*) ANOVA/MANOVA modellről beszélünk, amikor az adatnyerés több ízben ugyanarról az egyedről származik (például az egyed fejlődését folyamatában vizsgálva, vagy az egyed több szervéről véve a mintákat stb.), ilyenkor megkülönböztetünk faktorszinten belüli (*within-factor effect*) és szükség esetén a fix faktorszintek közötti (*between-subject effect*) hatásokat, illetve ez utóbbiak interakcióját is, ha több van belőlük. Fontos és erős feltétele a normalitás mellett a szfericitás, azaz, hogy az ismételt mérések páronkénti különbségeinek varianciái azonosak legyenek. A modell kovarianciastruktúrája az ún. összetevő-szimmetria (*compound symmetry*), azaz a kovarianciamátrix főátlójában azonos szórásnégyzetek, a többi részében azonos kovarianciák állnak.

A marginális modellek (*marginal models*) ennél sokkal általánosabbak és rugalmasabbak (Agresti, 2002). Szintén ismételt méréseken alapulnak, de nem követelik meg a szfericitást, és speciális korrelációt is megengednek az ismételt mérések között, így például az autoregressziót, azaz a korábbi állapot(ok)tól való függést is, így a modell a visszacsatolást is tudja kezelni. A modell kovarianciastruktúrája a faktorszintek eltérő varianciáját is megengedi, valamint a kovarianciája is lehet különböző, sőt, akár minden kovarianciája lehet más és más (*unstructured covariance structure*). Az összetevő-szimmetrikus és a strukturálatlan kovarianciastruktúra között számos közbülső eset lehetséges. Ezek közül a megfelelőt jól megválasztva a marginális modellekkel esetenként sokkal jobb illesztést lehet elérni. Hátrányuk viszont, hogy a nagyobb számú becslés miatt több szabadsági faktort emésztenek fel, így a gyakorlatban akkor alkalmazzuk, ha a kovarianciamártix dimenziója nem nagy, illetve méretéhez elegendően nagy a mintaelemszámunk.

Megjegyezzük, hogy egy adott adathalmazra esetenként többféle modellt is illeszthetünk (kevert, ismételt méréses ANOVA/MANOVA, marginális). Hogy ezek közül melyik lesz az optimális, az a modellek összehasonlítása során derül ki.

Az általános lineáris modell mindegyike azonban megköveteli, hogy a modell hibatagja normális eloszlású legyen, valamint azt, hogy a mintaelemek egymástól függetlenek legyenek.

3.4 Az általánosított lineáris modell

Az általánosított lineáris modell (*Generalized Linear Modell*) az általános lineáris modell általánosítása, amikor is nem követeljük meg a válaszváltozó normalitását, csupán azt, hogy az eloszlása az ún. exponenciális eloszláscsaládból származzon (Agresti, 2002, Bolker és mtsai, 2009). (Ebbe a családba tartozik a normális eloszláson felül például a binomiális, a Poisson-, az exponenciális, a gamma- stb. eloszlás.) Ilyenkor az Y válaszváltozó $E(Y)$ várható értékét, a modell ún. véletlen komponensét becsüljük a magyarázó változók lineáris kombinációjával (az ún. szisztematikus komponenssel) egy monoton növekedő g ún. kapcsolati függvényen (*link function*) keresztül:

$$g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_{1k}.$$

Itt már nem szükséges, hogy Y folytonos legyen. Ha a például a függő változó kétértékű, akkor logisztikus regressziós modellről beszélünk, ahol Y eloszlása az $(n = 1, p)$ paraméterű binomiális eloszlás, kapcsolati függvénye a $logit(p) = log\left(\frac{p}{1-p}\right)$ függvény (Agresti, 2002, James és mtsai, 2013). (Ezt tovább általánosíthatjuk, ha Y multinomiális eloszlású, és ekkor multinomiális regressziós modellről beszélünk.) Ha a g függvény az identitás, az általános lineáris modellt kapjuk vissza.

Az általánosított lineáris modellbe is tehetünk véletlen faktort, akár véletlen tengelymetszetet, akár véletlen meredekséget, akár mindkettőt.

R környezetben (R Core Team, 2023) ANOVA modelleket a `'stats'` csomag `'aov()'`, `'anova()'`,vagy `'oneway.test()'` függvényeivel végeztem a feltételek teljesülésétől függően (R Core Team, 2023). Ismételt méréses és ANCOVA modellek esetén az `'rstatix'` csomag `'anova_test()'` függvényével dolgoztam (Kassambra, 2023). Kevert modellekhez vagy az `'nlme'` csomag `'lme()'` függvényét (Pinheiro és Bates, 2000, Pinheiro és mtsai, 2023), vagy az `'lme4'` csomag `'lmer()'` függvényét használtam (Bates és mtsai, 2015). Lineáris regresszióhoz a `'stats'` csomag `'lm()'` függvényét alkalmaztam. MANOVA modellhez a `'stats'` csomag `'manova()'` függvényét használtam.

ANOVA, MANOVA, ismételt méréses ANOVA, lineáris és nemlineáris regressziós modellekhez az IBM SPSS statisztikai szoftvert is használtam a 16-tól a 29. verzióig (Armonk, NY, 2023).

3.5 Válaszfelületek

A statisztikában a mintaelemszám megválasztása mindig is fontos kérdés, amivel részleteiben itt nem foglalkozunk. Annyit érdemes megjegyezni, hogy bizonyos módszerek (pl. amelyek határeloszlás-tételen alapulnak) eleve csak nagy mintaelemszám esetén alkalmazhatóak, más módszerek – például a várható értékek összehasonlítására használatosak – esetén a kívánt mintaelemszám a próba kívánt erejének ($1-\beta$, ahol β a másodfajú hibavalószínűség), és a várhatóértékek különbségének (effect size, Δ)

függvényében határozhatóak meg adott csoportszám esetén, hogy csak a legegyszerűbb, egytényezős ANOVA példáját említsük. A szükséges mintaelemszám meghatározása összetett modelleknél már igen bonyolult lehet, sok esetben iterációs módon jutunk egy alkalmas becslésre.

Igen ám, de vannak olyan esetek – és ez az agrártudományban gyakran elő is fordul – amikor a kísérlet olyan időigényes és/vagy költséges, hogy nem lehetséges nagy ismétlésszámban gondolkodni a tervezésnél. Ilyen például, amikor egy élelmiszertermék előállításához egy vagy több paraméternek bizonyos értékét kell beállítani, és úgy akarjuk ezeket a paramétereket meghatározni, hogy a terméken egy tulajdonság értékét optimalizáljuk. Ilyenkor alkalmazzuk az ún. válaszfelületek-módszert (*response furface method*, RSM). Gondosan megválasztott szerkezetben tervezzük meg a kísérletet: ún. faktoriális tervet készítünk attól függően, hogy hány faktorhatást (paraméterbeállítást) és a faktorhatások hány szintjét szeretnénk (van lehetőségünk) a kísérletbe bevonni. Vegyük azt a példát, amikor két faktor, A és B két-két szintjének – alacsony (-) és magas (+) – hatását szeretnénk vizsgálni. Ez egy ún. 2^2-es faktoriális terv szerint történhet (6. Ábra, bal felső panel), amit egy négyzet négy csúcsával azonosíthatunk. Három, egyenként két szintű faktor esetén minimálisan 2^3 sorból áll a tervünk, ami egy kocka csúcsainak felel meg (6. Ábra, jobb felső panel).

Ha a négyzetek ezen pontjaiban (ezekkel a paraméterbeállításokkal) végezzük el a termék előállítását, és mérjük a termék optimalizálandó tulajdonságát, és a mért értékeket a négyzet síkjára merőleges koordinátatengelyen mérjük fel, akkor erre egy felületet simíthatunk, és e felület maximumának vagy minimumának a helyét keressük a négyzet síkjában. Ez lesz az optimális paraméterbeállítás. Ezzel a simítással persze egy erős folytonossági feltételt tettünk, ami bizonyos folyamatoknál természetes módon teljesül.

Az 6. Ábra alsó paneljén egy két, egyenként három-három szintű, ún. 3^2 kísérleti tervet látunk, ahol a középső szint az alacsony és a magas paraméterérték felezőpontjánál van, és amely egy négyzet csúcsainak, oldalfelezőpontjainak, illetve középpontjának felel meg.

A kísérlet véletlen sorrendje	A	B
2	+	+
4	-	-
3	+	-
1	-	+

A kísérlet véletlen sorrendje	A	B	C
7	+	+	+
1	-	-	+
4	+	-	+
8	-	+	+
9	+	+	+
5	-	-	+
2	+	-	+
6	-	+	+

A kísérlet véletlen sorrendje	A	B
7	+	+
1	+	0
4	+	-
8	0	+
9	0	0
3	0	-
5	-	+
2	-	0
6	-	-

6. Ábra Egy 2^2 (felső sor, bal panel), egy 2^3 (felső sor, jobb panel) és egy 3^2 (alsó sor) faktoriális terv válaszfelület-modellhez

A faktoriális beállítások ennél sokkal összetettebbek lehetnek, és szükség is van olyan ötletekre, amelyekkel költséges kísérletek szükséges számát mérsékelhetjük. Egy egyszerű faktoriális terv két faktor és három szint esetén minimum kilenc mérést igényel, három faktor és három szint esetén huszonhetet. Ezzel szemben Box és Behnken (1960)

speciális kísérleti terve huszonhét helyett tizenhárom kísérlettel is meg tudja oldani ugyanezt a feladatot.

Egy egyszerű válaszfelület-modell általános alakja:

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} X_i X_j + \sum_{i=1}^{k} \beta_{ii} X_{ii}^2 + \varepsilon,$$

ahol β_0 konstans tag, X_i $(i = 1,2, \dots k)$ faktorok, β_i és β_{ij} együtthatók, $\sum_{i=1}^{k} \beta_i X_i$ az ún. elsőfokú tag, $\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij} X_i X_j$, az interakciós tag, $\sum_{i=1}^{k} \beta_{ii} X_{ii}^2$ a másodfokú tag, ε pedig a zérus várható értékű, normális eloszlású hibatag.

17. Példa a válaszfelület-módszer alkalmazására

Varga és munkatársai (2023) világos láger sör fordított ozmózissal való alkoholmentesítését vizsgálták. A teljes faktoriális kísérleti tervben a faktorok a következők voltak: transzmembránnyomás (bar) három nyomásértéken mérve, illetve retentátum áramlási sebesség (L/h) szintén három beállítással. Az optimalizálandó válaszváltozó az etanolfluxus volt. A célfüggvény globális maximumát rácspont-módszerrel (*grid-search*) határoztuk meg. Modellünk eredménye alapján a legmagasabb etanolfluxust a legmagasabb transzmembránnyomás és a legalacsonyabb retentátum áramlási sebesség beállításával kaptuk.

4 A GÉPI TANULÁS

4.1 A gépi tanulás fogalma

A gépi tanulás (*machine learning*) olyan gépi algoritmus, amely automatikus fejlődésre képes az adatok egyre mélyebb megismerése által (Mitchell, 1996, Boehmke és Greenwell, 2020, Hastie és mtsai, 2013, Lantz, 2015, Ghatak, 2017). Három elemből áll, a feladatból (T, *task*), a tapasztalatból (ami az adatnak felelhet meg, E, *experience*), valamint a feladat elvégzésének minőségétől (P, *performance*). Az adathalmazt, amire épül, legegyszerűbben egy mátrixként képzelhetjük el, amelynek sorai egy-egy megfigyelt mintaelemre vonatkoznak, és melynek oszlopai ezen elemek egyes mért tulajdonságai, ismérvei (*features*, 3. Táblázat).

Az adathalmazt kezdetben ún. tanuló (*train data*) és tesztelő (*test data*) részre bontjuk. A tanuló adathalmazon az algoritmus modellt épít, és kapott a modellt a tesztelő adathalmazon értékeli és validálja.

3. Táblázat A gépi tanulás adathalmaza

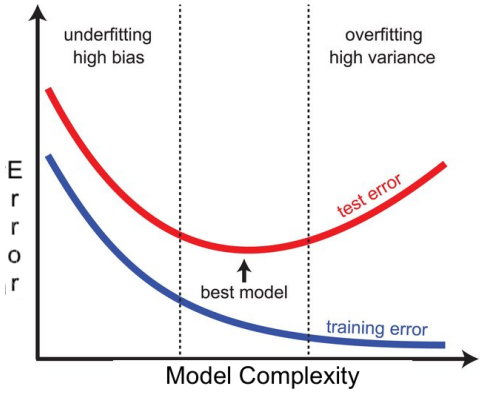
	Ismérvek (*features*)				
Mintaelem ID	Ismérv 1	Ismérv 2	…	Ismérv *k*	Címke (pl. tünet)
ID 1	Y	1,76	…	fehér	nincs
ID 2	Y	2,33	…	piros	nincs
⋮	⋮	⋮	⋮	⋮	⋮
ID N	N	4,14	…	kék	van

4.2 A gépi tanulási módszerek típusai

A gépi tanulás három legfontosabb típusa a felügyelt (*supervised*), a felügyelés nélküli (*unsupervised*), illetve a környezetével kommunikáló megerősítő (*reinforcement*) gépi tanulás (Ghatak, 2017, Kang és mtsai, 2018, Boehmke és Greenwell, 2020). E három típussal átfedésben vannak, illetve ezeket kiegészítik a félig felügyelt, illetve a metatanuló algoritmusok.

A *felügyelt gépi tanulás* adathalmaza a megfigyelt elemekről nemcsak ismérvváltozókat tartalmaz, hanem egy, az ismérvektől függő válaszváltozót is (ld. 5.1. fejezet). A válaszváltozó lehet bináris, nominális, ordinális, vagy folytonos mérési szintű, mely valamilyen értelemben minősíti az egyes elemek ismérvváltozóit (pl. ’tünetes’ – ’tünetmentes’ címkékkel).

Az algoritmus a tanuló adathalmazon az ismérvváltozók hasonlóságának elemzése során olyan összetett modellt állít elő, melynek eredménye egy becslő függvény, pontosabban egy becslő függvény paraméterhalmaza (*hyperparameters*). A becslő függvény paramétereit ugyanis a modell úgy kalibrálja, hogy annak értéke a lehető legpontosabban becsülje a válaszváltozót. Ez alatt azt értjük, hogy a modell sem nem felülillesztett, sem pedig alulillesztett (Ghatak, 2017, 7. Ábra). A felülillesztettség (*overfitting*) azt jelenti, hogy a modell túl komplex, specifikus lett, azaz speciálisan az adott adatsorra jellemző, más adatsor esetén rosszul illeszkedik a válaszváltozóra, vagyis nem ragadta meg a mögöttes általános összefüggést. Az alulillesztettség (*underfitting*) ennek ellenkezője, amikor a modell túl egyszerű, kimenete rosszul illeszkedik a válaszváltozóra (Kuhn és Johnson, 2013).



7. Ábra A tanuló és tesztelő adatokból eredő hiba a modell komplexitásának függvényében. (Forrás: Beyeler és mtsai, 2019, saját átszerkesztéssel)

A tesztelő adathalmazon vizsgálhatjuk a modell jóságát, amikor is a modell az előbbiekben kalibrált paramétereivel előállítjuk a kimenetet, és azt hasonlítjuk össze a tesztadathalmaz válaszváltozójának értékeivel. Mivel a tanulás folyamán a tesztadathalmazt nem látta az algoritmus, mi lényegében azt vizsgáljuk, hogy a lényeget ragadta-e meg a modell, azaz ismeretlen adatokon is jól működik-e. Amennyiben a modellünk jónak bizonyul, újabb megfigyelt elemek ismérvváltozóinak ismeretében a válaszváltozó értéke immár a modellel becsülhetővé válik.

Bár a modellezés során több lépést automatizálhatunk, a modellező személy szerepe továbbra is fontos a felügyelt tanulásban. Nemcsak a tanuló adathalmaz kimenetét (a válaszváltozót) jelöli meg, hanem az ismérvváltozók közül is az ő döntésén is múlik a szükséges releváns magyarázó változók (ismérvek) kiválasztása (*feature selection,* Kuhn és Johnson, 2013, Boehmke és Greenwell, 2020). A változók helyes megválasztása biztosíthatja a túlillesztés elkerülését, ami a modell általánosíthatóságát fokozza. A modellező dönt az egyes algoritmusok használatáról, illetve adott esetben azok paramétereit is ő határozza meg, mindezeket többek között az algoritmusok feltételrendszere, illetve azok egyes adathalmazon való működésének lehetősége és sikere alapján.

18. Példa arra, amikor az ismérvváltozók kiválasztása egyedi volt

Szűgyi-Reiczigel és munkatársai (2022) egy tudományos és történeti szempontból jelentős adathalmazt elemeztek. Németh (1966) 97 szőlőfajtát (*Vitis vinifera*, L.) 125 morfológiai bélyeg alapján írt le. A 125 morfológiai bélyeg ebben az adatsorban nyilvánvalóan összefüggő, tehát redundáns információt is tartalmaz. A szerzők azt vizsgálták, hogy hogyan határozható meg a morfológiai bélyegek azon részhalmaza, amely a legsikeresebben magyarázza a szőlőfajták Németh által besorolt convarietas osztályozását (*pontica, occidentalis* és *orientalis*), mégpedig kettős célt fogalmazva meg: (1) definiáljunk olyan információveszteség-mentes szelekciós algoritmust, amely nagyszámú függő, ordinális és nominális típusú ismérvet is tartalmazó adatokra alkalmazható, és amely a felhasználó mint szakértő által szabályozható; (2) mutassuk meg, hogy a kiválasztott tulajdonságok milyen sikerrel képesek a három kategóriába sorolást elvégezni.

Ehhez az aszimmetrikus Goodman-Kruskal-féle (Goodman és Kruskal, 1954, 1979) λ mutatót bizonyult alkalmasnak, amely a faktortípusú változók (bélyegek) arányos hibacsökkentésének mérőszáma. A 0-tól 1-ig terjedő értékeket felvevő λ azt méri, hogy mennyire nő az Y függő változónak az előrrelejezhetősége, ha ismerjük az X magyarázó

változó értékét ahhoz képest, mintha nem ismernénk: $\lambda(Y|X) = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1}$, ahol ε_1 és ε_2 az Y előrejelzésének hibája, ha ismerjük (ε_2), vagy ha nem ismerjük (ε_1) az X-et. Az $\lambda(Y|X) = 1$ azt jelenti, hogy az Y pontosan meghatározható X ismeretében, $\lambda(Y|X) = 0$ pedig azt, hogy Y-t nem tudjuk pontosabban becsülni, hiába ismerjük az X-et. A Goodman-Kruskal-féle λ aszimmetrikus mérték: előfordulhat, hogy Y nem jósolható meg X alapján, de X megjósolható Y alapján, ahogy ilyenre volt is példa az adatsorban. Ha X és Y független, akkor $\lambda(Y|X) = \lambda(X|Y) = 0$, de $\lambda(Y|X) = 0$-ból nem következik a függetlenség.

Először azok az esetek kerültek leválogatásra, amikor két bélyegre $\lambda(Y|X) = 1$ fennállt, ilyenkor az Y bélyeg kikerült az adathalmazból. Érdekes kérdés volt, hogy ha $\lambda(Y|X) = 1$ és $\lambda(X|Y) = 1$, tehát amikor X és Y kölcsönösen meghatározzák egymást, melyik változó kerüljön ki a modellből. Ilyenkor a szerzőtárs ampelométer szakemberrel közösen született döntés.

Később a két, három, illetve négy változó együttes ismeretében maradéktalanul meghatározható bélyegek kerültek leválogatásra. Így összesen 66 bélyeget került eltávolításra információveszteség nélkül, a megmaradt 59 bélyeggel pedig véletlen erdő (RF, ld. 5.3.2. fejezet) módszerrel végeztük el a szőlőfajták klasszifikációját.

Természetesen az ismérvek összefüggése erre a specifikus adatsorra jellemző, de leválogatásuknak ez a módszere bármely más adatsorra alkalmazható a redundáns változók kiszűrésére.

A módszer nemcsak önmagában volt hasznos, hanem a szőlész szakemberek számára is érdekes összefüggéseket tárt fel " ha ilyen és ilyen volt, akkor következésképpen az is jellemző volt rá" formában.

A *felügyelés nélküli gépi tanulás* a felügyelt tanulással ellentétben nem igényel előzetes válaszváltozót. A felügyelés nélküli tanulást általában akkor használjuk, ha a bemeneti változók közötti kapcsolatok feltárása a cél. Ahelyett, hogy a felügyelt tanulás eredményéhez hasonló *kimeneti érték* megadása lenne a cél, a felügyelés nélküli tanulás a bemeneti változók *mintázatát* keresi, hasonlóságon, illetve különbségeken alapuló osztályozását végzi (ld. 5.2. fejezet).

A *félig felügyelt gépi tanulás* esetében a felügyelt tanulás során a válaszváltozó által felcímkézett adatokat használjuk a modell képzésére, és ez robusztusabbá és pontosabbá teszi a modellt; a válaszváltozók előállítása azonban némelykor költséges. Számos esetben az adatoknak csak egy kis részét címkézik válaszváltozóval, míg az adatpontok többségét felirat nélkül hagyják. A félig felügyelt tanulási algoritmusok modellépítése tehát címkézett és címkézetlen adatokra egyaránt épülhet.

A *megerősítő gépi tanulás* során a tanuló gép kommunikál a környezetével, és néhány lehetséges műveletet tud végrehajtani. A gép által eleinte véletlenül választott művelet alapján „jutalmat", illetve „büntetést" (negatív/pozitív visszajelzést) kap, majd a modell ennek megfelelően frissül, és ebből tanulva a műveleteket egyre inkább a pozitív visszajelzés elérését célozva választja ki. A megerősítő tanulási típus tehát visszacsatolási mechanizmust alkalmaz a pozitív műveletek jutalmazására és a negatív cselekmények büntetésére. Ezt a megközelítést alkalmazzák például az önvezető autókban.

A *metatanuló* gépi algoritmusok képesek más algoritmusoktól, azok kimeneteitől, paramétereitől tanulni, illetve adott esetben „tanulást tanulni".

4.3 A gépi tanulási módszerek feladatai (céljai)

A gépi tanulás az alábbi leggyakoribb alapfeladatokra használható (Ghatak, 2017):

- osztályozás
- regresszió
- hasonlósági csoportok létrehozása (klaszterezés)
- adatredukció
- anomáliakeresés

A fenti feladatokat röviden az alábbiakban mutatjuk be. Megjegyezzük, hogy a felsorolt feladatok nem kizárólag gépi tanulási módszerekkel végezhetőek el: hagyományos statisztikai módszerek is alkalmasak ilyen célra. Hogy hogyan oldjuk meg ezeket a feladatokat (hagyományos statisztikai módszerekkel vagy gépi tanulási módszerekkel), az leginkább az adatok mennyiségén, minőségén és struktúráján múlik (ld. 4.7. fejezet).

A *regresszió* ún. becslő eljárás, amely a bemeneti jellemzők függvényében egy folytonos numerikus változót ad eredményül. A gépi tanulási algoritmusok az egyes független ismérvváltozók együtthatóit úgy optimalizálják, hogy a hiba valamely definíciója alapján annak minimumát érjék el az előrejelzésben (ld. 5.1.3. fejezet).

Az *osztályozás* a bemeneti jellemzőket egy diszkrét (faktortípusú) kimeneti változóra képezi. Bináris osztályozás esetében a kimeneti változó csak 0 vagy 1 lehet. Többszintű osztályozáskor a kimeneti változó több osztályból állhat (ld. 5.1.3. fejezet). Az osztályozás a hasonlóság, illetve különbözőség valamely definíciója alapján a hasonló mintaelemeket igyekszik azonos csoportba sorolni, a különbözőeket pedig szétválasztani.

Hasonlósági csoportok létrehozásakor (klaszterezés) az algoritmus a mintaelemeket (adatpontokat) hasonlósági csoportokra osztja az adatok ismérvváltozóinak értékei alapján csoportosítással, vagy az adatok különbözősége (távolsága) alapján elválasztással (ld. 5.2.1. fejezet).

Az *adatredukciós módszerek* ún. dimenziócsökkentő módszerek. Az erősen összefüggő vagy nem túl releváns változók egy része el is távolítható az adatkészletből. A megmaradt hasonló információt hordozó (összefüggő) változókat a módszer tömöríti, így kevesebb számú, de már független változókat hoz létre. Ezt a módszert sok esetben kiegészítő módszerként használják más gépi tanulási feladatokhoz, például regresszióhoz és osztályozáshoz (ld. 5.2.2. fejezet).

Az *anomáliafelismerési* feladatot elsősorban felügyelés nélküli tanulási módszerekkel kezelik. A klaszterezéshez hasonlóan az anomáliafelismerő algoritmusok hasonlóságon alapuló csoportosítást végeznek, és így megjelölik az adatkészletben előforduló kiugró (semmihez sem hasonló; *outlier*) értékeket (ld. 5.1.1. fejezet).

4.4 A gépi tanulási módszerek lépései

A gépi tanulási projekt, mint minden modellező munka, az adatok gyűjtésével és azok előkészítésével kezdődik (2.1. fejezet).

Ennek utolsó lépése a gépi tanulásra speciálisan jellemző lépés. Az adathalmazt tanuló, illetve tesztelő adathalmazra osztjuk fel (Ghatak, 2017, Boehmke és Greenwell, 2020, ld. 4.1. fejezet).

Ez a művelet külön figyelmet igényel, hiszen a tanuló és tesztelő adathalmaz arányától (is) függ, hogy a modellünk felül-, illetve alulillesztett lesz-e, avagy megfelelő (ld. 4.2. fejezet). Szintén fontos, hogy a tanuló adathalmaz (is) reprezentatív legyen. Nincs

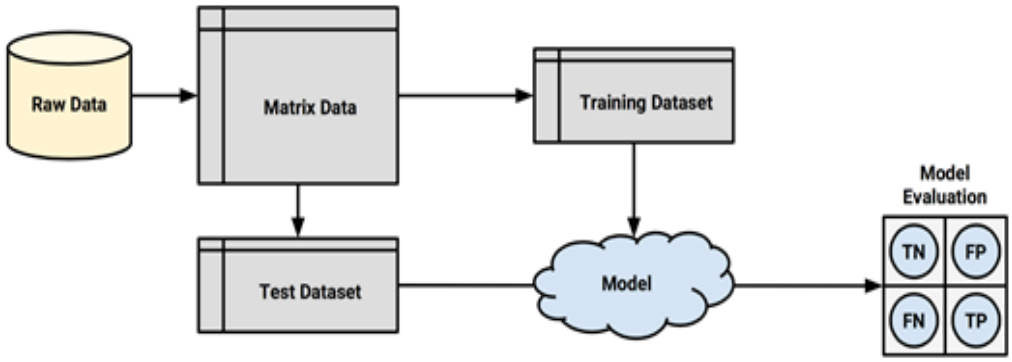
egy minden esetben alkalmas tanuló/tesztelő adathalmaz arány. A leggyakoribb választások az alábbiak: Tanuló: 80% + Tesztelő: 20%; Tanuló: 67% + Tesztelő: 33%.

Amennyiben a válaszváltozó diszkrét, azaz osztályozási feladatunk van, akkor előfordulhat, hogy a válaszváltozó egyes szintjei (csoportjai) igen eltérő mintaelemszámúak. Ilyenkor a fenti, egyszerű arányokra alapuló tördelés nem mindig sikeres, előfordulhat, hogy a tesztelő adathalmazban olyan esetek vannak, amilyenekkel a modell a tanuló adathalmazban még nem találkozott. Hogy ezt a problémát elkerüljük, ún. rétegelt tördelést (rétegelt mintavétel) alkalmazunk (*stratified splitting, stratified sampling*, Kuhn és Johnson, 2013, Hastie és mtsai, 2013, Morais és mtsai, 2019, Boehmke és Greenwell, 2020). Ekkor a válaszváltozó minden szintjén külön-külön vesszük a fenti felosztási szabály valamelyikét, így biztosítva, hogy a tanuló adathalmazba megfelelő számú mintaelem essék a minden csoportból.

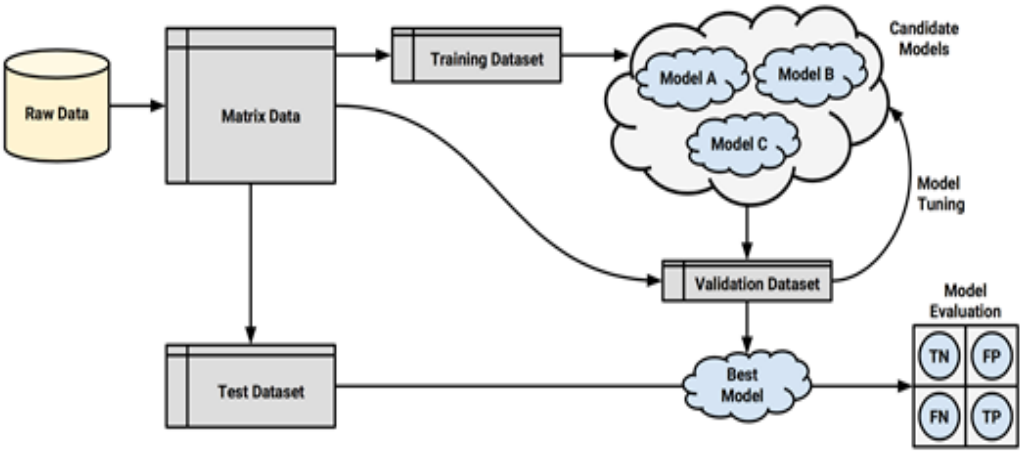
Ahhoz, hogy valóban hatásos modellt fejlesszünk a teszthalmazon való értékelés alapján, még egy validációs lépést iktathatunk be. Ha ugyanis a legjobb modellt nagyszámú ismételt tesztelés alapján választjuk ki, mégpedig attól függően, hogy melyik tanuló-tesztelő pároson értük el a legjobban teljesítő modellt, akkor a modell teljesítménye nem lesz torzításmentes.

Ezért a tanuló adathalmazt is két részre osztjuk, beiktatva egy ún. belső validálási adathalmazt. A belső validálási adathalmaz a modell fejlesztése során használt tesztelő adathalmaz, a tanuló adathalmaz része. Miután elkészült a legjobban teljesítő modellünk, azt teszteljük a tesztelő adathalmazon, mégpedig egyszer, a végső lépésben. A képzés, a tesztelés és a validálás közötti felosztás lehet például Tanuló: 60% + Validáló: 20% + Tesztelő: 20% (8. Ábra, Lantz, 2015).

a



b



8. Ábra Tanuló és tesztelő adathalmazok validációs adathalmazzal (b) vagy anélkül (a) (Forrás: Lantz, 2015)

4.5 A legfontosabb gépi tanulási módszerek

A gépi tanulási algoritmusok kutatásainkban alkalmazott legfontosabbjait, azok típusai, illetve feladatai szerint a 4. Táblázatban foglaltam össze. Itt csak azokat említem meg, melyeket kutatásom során használtam, illetve jelen munkámban bemutatok az 5. fejezetben.

4. Táblázat A legfontosabb gépi tanulási módszerek

	Feladatok				
Típus	Osztályozás	Becslés	Csoportosítás	Többcélú	Adatredukció
Felügyelt	K-legközelebbi szomszéd (K-NN) Naiv Bayes (NB) Klasszifikációs döntési fák (DT) Logisztikus regresszió (LogR) Diszkriminancia-elemzés (lineáris, nemlineáris, DA, LDA, QDA)	Regresszió (lineáris, nemlineáris, többszörös, (LR, NLR, MLR) Regressziós döntési fák (RT)		Támogatóvektor-gépek (SVM) Parciális legkisebb négyzetek módszere (PLSR)	
Felügyelés nélküli			Hierarchikus és k-közép klaszterezés (HC, KMC)		Főkomponens-elemzés (PCA)
Félig felügyelt				Főkomponens-regresszió (PCR)	
Metatanulás				Zsákolás és gyorsítás (BB) Neurális hálózatok (ANN) Véletlen erdő (RF)	

4.6 A modellek jóságának mérőszámai

A modellek értékeléséhez különféle mérőszámokat vezetünk be (*performance measurements*).

Becslési (regressziós) feladat esetén a modell becslésének jóságát az átlagos eltérés-négyzetek négyzetgyöke (*Root Mean Square Error*, RMSE), vagy a magyarázott varianciahányad (R^2), vagy az átlagos abszolút hiba (*Mean Absolute Error*, MAE) alapján állapítjuk meg (Rao, 1948, Powers, 2011, Kuhn és Johnson, 2013, Ghatak, 2017, Boehmke és Greenwell, 2020, Tharwat, 2021).

A klasszifikációs modellek jóságát a legegyszerűbben a helyes, illetve helytelen pozitív, illetve negatív hozzárendelés fogalmainak segítségével mérhetjük (5. Táblázat, James és mtsai, 2013).

5. Táblázat Kereszttáblázat (*confusion matrix*) a klasszifikáló modellek jóságának vizsgálatához szükséges definíciókhoz

		a modell válasza negatív	pozitív	
a valóság	negatív	helyes negatív (TN)	hibás pozitív (FP)	a specificitást ebből a sorból számoljuk
	pozitív	hibás negatív (FN)	helyes pozitív (TP)	az érzékenységet ebből a sorból számoljuk
		a negatív predikciós értéket ebből az oszlopból számoljuk	a precizitást ebből az oszlopból számoljuk	

A fenti alapfogalmak segítségével képzett, a modellek jóságának mérőszámait a 6. Táblázatban foglaljuk össze.

6. Táblázat A modellek jóságának mérőszámai

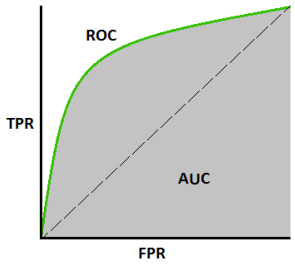
Elnevezés	Definíció	Angol elnevezés
Specificitás	$Sp = \frac{TN}{FP + TN}$	Specificity or True Negative Rate (1- Type I error)
Szenzitivitás	$Se = \frac{TP}{TP + FN}$	Sensitivity or Recall or True Positive Rate (power, 1-Type II error)
Pozitív predikciós érték (precizitás)	$PPV = \frac{TP}{TP + FP}$	Precision or Positive Predicted Value (1-false discovery rate)
Negatív predikciós érték	$NPV = \frac{TN}{TN + FN}$	Negative Prediction Rate
Prevalencia	$Prev = \frac{TP + FN}{N}$	Prevalence
Detekciós szint	$DR = \frac{TP}{N}$	Detection Rate
Detekciós prevalencia	$DPrev = \frac{TP + FP}{N}$	Detection Prevalence
Pontosság	$A = \frac{TP + TN}{N}$	Accuracy
Kiegyensúlyozott pontosság	$BA = \frac{Se + Sp}{2}$	Balanced Accuracy
F1-szkór	$F1 = \frac{2PPV * TPR}{PPV + TPR}$	F1 score
Cohen-féle kappa	$\frac{P(MM) - P(\mathrm{VM})}{1 - P(VM)}$	Cohen's Kappa

TP = *True Positives* (helyes pozitív felismerés), TN = *True Negatives* (helyes negatív felismerés), FP = *False Positives* (téves pozitív felismerés), and FN = *False Negatives* (téves negatív felismerés); $P(MM =$*P*(megfigyelt megfelelés) a modellbecslés megfelelési rátája, $P(VM) =$ *P*(várt megfelelés) pedig az a megfelelési arány, amely az osztályozás véletlen választásakor várható.

A *Cohen-féle kappa* statisztika (Cohen, 1960, Landis és Koch, 1977) a pontosságot úgy méri, hogy figyelembe veszi annak a valószínűségét, hogy a helyes döntés véletlenszerűen született. Ennek akkor van nagy jelentősége, ha például osztályozás esetén az egyik csoportunk elemszáma igen magas a többihez képest, így már akkor is jó eredményt érnénk el, ha minden elemet ebbe a csoportba sorolnánk, ami persze nem célunk. A *P*(várt megfelelés) a csoportok elemszámától függ. Az 1-et nem meghaladó értékeket felvehető Cohen-féle kappára ún. ökölszabályt alkalmazunk: 0,4 felett már közepes, 0,6 felett már jó, 0,8 felett kiválónak minősül (Cohen, 1960). Kiegyensúlyozatlan mintából

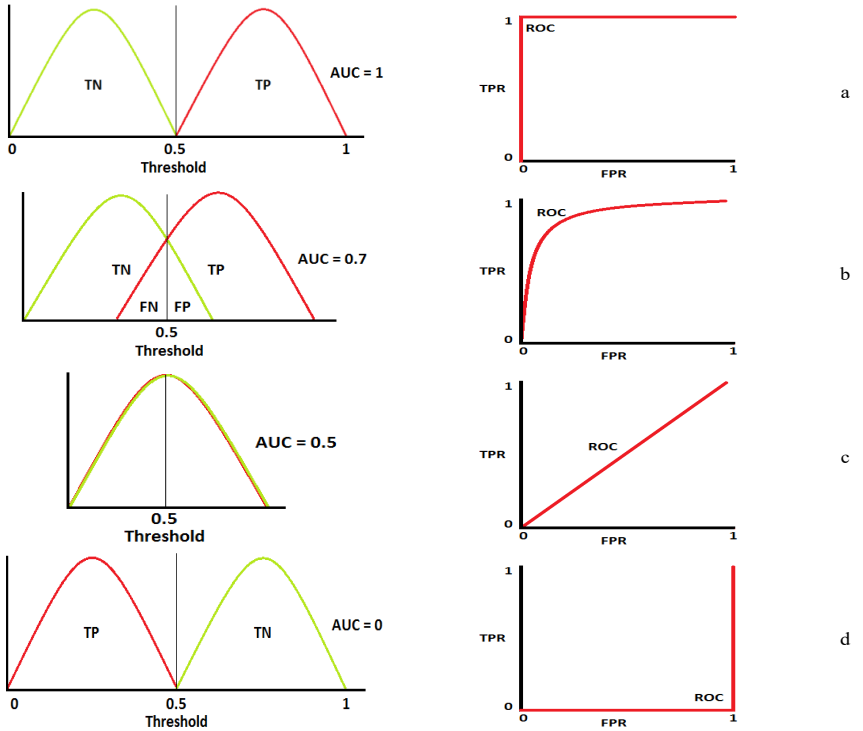
származó adatok esetén az F1 szkór jobb becslést ad a modell jóságára (Ghatak, 2017), mint a pontosság (*accuracy*).

Ha a helyes pozitív arányt (érzékenységet) ($0 \leq TPR \leq 1$) ábrázoljuk a hibás pozitív arány ($0 \leq FPR \leq 1$, 1-specificitás) függvényében, akkor a modellek egy másik fontos értékelését, az ún. *ROC-görbét* (*receiver operating characteristic curve,* Ghatak, 2017, Boehmke és Greenwell, 2020, James és mtsai, 2013) kapjuk (9. Ábra). A ROC-görbe kiegyensúlyozatlan mintavételből származó adathalmaz esetén különösen informatív adaléka a modell értékelésének.



9. Ábra ROC-görbe (FPR: hibás pozitív arány; TPR: helyes pozitív arány (Forrás: Sarang Narkhede, url)

Az ún. AUC (AUC: *area under the ROC curve*) a ROC-görbe és az *X* tengely által bezárt terület nagysága, azaz a ROC görbe 0-tól 1-ig vett határozott integrálja (9. Ábra, Fawcett, 2006). Minél magasabb az AUC értéke, annál jobb a modell elválasztó ereje (ökölszabály: 0,7 felett közepes, 0,8 felett jó, 0,9 felett kiváló, Hajian-Tilaki, 2013, Hosmer és mtsai, 2013, Sing és mtsai, 2005, 10. Ábra).



10. Ábra A modell elválasztási képessége és a ROC-görbe kapcsolata néhány példán keresztül: a. AUC=1, tökéletes elválasztás; b. AUC=0,7, a modell nem hibátlan, de megfelelő; c. AUC=0,5, a modell semmitmondó; d. AUC=0, a modell teljesen hibás (minden pozitívat negatívnak, minden negatívat pozitívnak ismer fel) (Forrás: Sarang Narkhede, url)

R környezetben a modellek értékelését a ‘modelr’ csomag ‘rsquare()’, ‘rmse()’, illetve ‘mae()’ függvényeivel (Wickham, 2023), vagy a ‘caret’ csomag ‘postResample()’, ‘MAE()’, vagy ‘RMSE()’ fügvényeivel kaptam meg (Kuhn, 2008, 2020). Az információs kritériumok értékeit a ‘modelr’ csomag ‘AIC()’ és ‘BIC()’ függvényeivel számítottam (Wickham, 2023). A ‘broom’ csomag ‘glance()’ függvényei ezek mindegyikét kiadja (Robinson és mtsai, 2023). ROC görbét és AUC értéket állít elő a ‘pROC’ csomag ‘roc()’ függvénye is (Robin és mtsai, 2011).

4.7 A hagyományos statisztikai módszerek és a gépi tanulás kapcsolata

A gépi tanulás feladatait, módszereit, az eredmények jóságának vizsgálatát tekintve sokszor hagyományos statisztikai módszerekre ismerünk rá. Felmerül tehát a kérdés, hogy pontosan mi is a különbség a hagyományos statisztikai módszerek alkalmazása és a gépi tanulás módszereinek alkalmazása között.

A gépi tanulás módszerei a hagyományos statisztikai módszerekből indulnak ki, azok módosításával, továbbfejlesztésével, kiegészítésével építkeznek. Akár hagyományos statisztikai, akár gépi tanulásos módszerrel közelítünk meg egy problémát, a közös, hogy a rendelkezésre álló adatokat megfelelő feldolgozást követően használva modelleket építünk, illetve meglévő modelleket fejlesztünk, a modellparamétereket becsüljük valamilyen módon, az eredményeket validáljuk, különböző modelleket, paraméterbecsléseket hasonlítunk össze, hogy kiválasszuk a kitűzött céljainknak legjobban megfelelőket, valamint értelmezzük, interpretáljuk a kapott eredményeket, hogy választ adhassunk a munka kezdetén feltett tudományos kérdésre.

A hagyományos statisztikai módszerek és a gépi tanulás módszerei között az alapvető különbségek az alkalmazási területekben és a modell alapjául szolgáló adathalmaz típusában, kiterjedésében, illetve minőségében rejlenek. A legfontosabb különbségek a hagyományos és a gépi tanulási módszerek között, hogy legtöbb esetben (de nem kizárólagosan) a gépi tanulási módszerek mintaelemszám-igénye már a tanuló és tesztelő adatokra való bontásból adódóan is jóval meghaladja a hagyományos statisztikai módszerekét, illetve az, hogy a gépi tanulási módszerek nagy adathalmazok birtokában jellemzően sokkal bonyolultabb összefüggéseket is képesek feltárni, mint a hagyományos statisztikai módszerek. Természetesen számos olyan gépi tanulási módszer van, amelynek nincsen megfelelője a hagyományos statisztikai módszerek között (zsákolás, gyorsítás, neurális hálózatok, véletlen erdő), néhány hagyományos statisztikai módszert sem sorolhatunk a gépi tanulás módszerei közé.

A távérzékeléssel, a mérőműszerek hatalmas fejlődésével, illetve az adatátviteli-adattárolási módszerek gazdagodásával az agrárkutatásokban egyrészt óriási adatmennyiségek halmozódtak és halmozódnak fel, másrészt bizonyos területeken az adatok mennyisége továbbra is szűkös a hozzáférés időigényessége, illetve költségessége miatt. (Termés továbbra is többnyire egyszer van egy évben…) Az agrárkutatásokban megjelenő adatstruktúrák típusainak heterogenitása ezzel a fejlődéssel igen megnőtt.

A gépi tanulás egyfelől hatalmas előretörésnek tekinthető a hagyományos statisztikai módszerekhez képest. Másfelől az alkalmazhatósági területük különbözősége miatt a gépi tanulás mellett a hagyományos statisztikai módszerek használata továbbra is szükséges és kívánatos. Hangsúlyosan érvényes ez az agrárkutatásokbeli alkalmazásukhoz, ahol a kettő megközelítés szorosan megfér egymás mellett, kiegészítik, gazdagítják egymást, bővítve ezzel a megoldási eszköztárat a nagyon eltérő adatstruktúrák kezelésére.

5 GÉPI TANULÁSI MÓDSZEREK

Az alábbiakban a gépi tanulási algoritmusok kutatásainkban használt legfontosabbjait (4. Táblázat) mutatom be. Hangsúlyozni szeretném, hogy a módszerek közül többeket alkalmazhatjuk hagyományos statisztikai módszerként is, és azokban az esetekben, amikor ezt az adataink minősége, illetve mennyisége úgy kívánta, mi is így alkalmaztuk.

5.1 Felügyelt (*supervised*) gépi tanulási módszerek

5.1.1 A K-legközelebbi szomszéd módszer

A *K-legközelebbi szomszéd módszer* (*K-Nearest Neighbour*, KNN) alapötlete az, hogy egy ismeretlen (hiányzó) cellaértéket az ehhez a cellához legközelebb lévő k db cella értékéből becsüljük, feltételezve, hogy a hiányzó cellaérték a legközelebbi szomszédok cellaértékeihez hasonló a leginkább (Hastie és mtsai, 2013, Kuhn és Johnson, 2013, Lantz, 2015, Tang és Ishwaran, 2017, Boehmke és Greenwell, 2020). Például képfelismerés esetén egy pixel hiányzó színét a körülötte lévő pixelek színeiből következteti ki.

A k érték választásának a felül-, illetve alulillesztettség-problémája szempontjából van jelentősége. Nagy k érték választásakor a zaj hatása (a szórás nagysága) csökkenthető, ám nő annak a valószínűsége, hogy kis, ámde fontos mintázatokat nem veszünk észre. A k-t gyakran a $k = \sqrt{N_T}$ közelében választjuk, ahol N_T a tanuló adathalmaz elemszáma. További, az alkalmazásokban gyakran előforduló megoldás, hogy nagy k-t választva a cellaértékeket súlyozzuk: a közelebbiek nagy, a távolabbiak kisebb súlyt kapnak (Boehmke és Greenwell, 2020). A k legközelebbi szomszédot folytonos adattípusú ismérvek esetén legtöbbször Euklideszi-távolsággal határozhatjuk meg. (Természetesen speciális feladatok esetében más távolságdefiníciót is használhatunk.) Amennyiben a kimenet diszkrét (kategória-) változó, a módszer a k legközelebbi szomszéd szavazatainak értékelése alapján konszenzussal dönt, szavazatazonosság esetén véletlen módon.

A módszer előnye, hogy nem követel meg az adatok eloszlására semmilyen feltételt, hátránya, hogy az ismérvek szerepe, fontossága rejtve marad. Használata adat-előkészítéskor (*preprocessing*) elterjedt (Cunningham és Delany, 2007, Kuhn és Johnson, 2013, Lantz, 2015). Mi is használtuk ilyen értelemben, a hiányzó értékek pótlására (6.2. fejezet).

R programozási környezetben adat-előkészítéshez a `'dplyr'` (Wickham és mtsai, 2023), a `'tidyverse'` (Wickham és mtsai, 2019), a `'mice'` (van Buuren és Groothuis-Oudshoorn, 2011), a `'gdata'` (Warnes és mtsai, 2023), valamint a `'MASS'` (Venables és Ripley, 2002), vizualizációhoz a `'ggplot2'` (Wickham, 2016), és a `'vcd'` (Meyer és mtsai, 2023b), tördeléshez az `'rsample'` (Frick és mtsai, 2024), illetve a `'caTools'` (Tuszynski, 2021), csomag függvényeit, KNN modellekhez a `'caret'` (Kuhn, 2008, 2020) csomag `'train()'` függvényét (method = `'knn'`), vagy a `'class'` (Venables és Ripley, 2002) csomagból a `'knn()'` függvényt használtam. Legutóbb az R 4.4.1 (2024-06-14) -es 'Race for Your Life' verzióját használtam. Korábban az ezt megelőző verziókkal dolgoztam a 2.4.0-tól kezdődően (2006-10-03) (R Core Team, 2023, CRAN, 2021).

5.1.2 A naiv Bayes-módszer

A *naiv Bayes-módszer* (*Naive Bayes Method*, NBM) olyan felügyelt módszer, amely a Bayes-tételen alapul azzal a feltétellel, hogy az ismérvváltozók $(X_1, X_2, \dots X_n)$ egymástól páronként függetlenek (Lantz, 2015):

$$P(Y|X_1, X_2, \dots X_n) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X_1, X_2, \dots X_n)}.$$

Mivel itt a nevező konstans, ezért a klasszifikációs szabály Y becslésére az alábbi lesz:

$$\hat{Y} = arg\max_{Y} P(Y)\prod_{i=1}^{n} P(X_i|Y).$$

A naiv Bayes-módszer speciális esete, amikor $P(X_i|Y)$-ról feltesszük, hogy normális eloszlású Y-tól függő várható érték és szórás paraméterekkel. További speciális esetek, amikor a $P(X_i|Y)$-t multinomiális, Bernoulli, vagy egyéb más eloszlásúnak feltételezzük (Hastie és mtsai, 2013, Kuhn és Johnson, 2013).

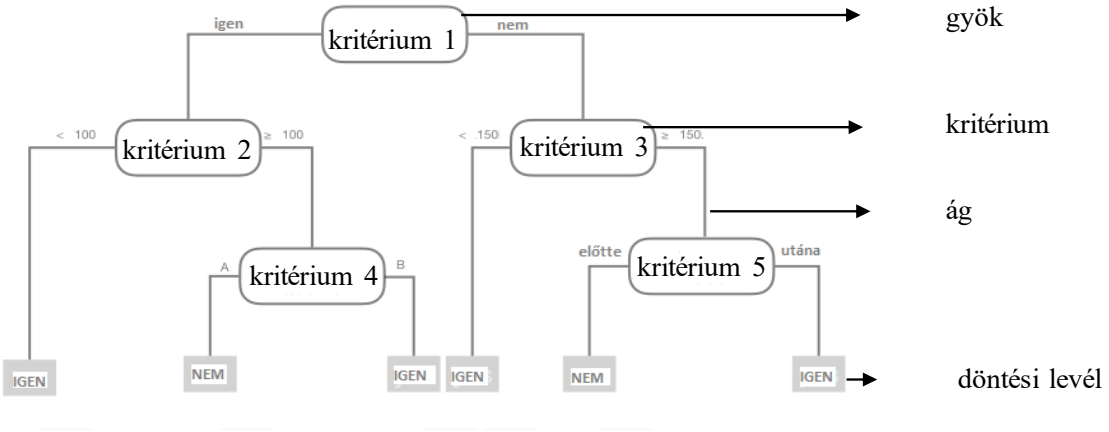
Megjegyezzük, hogy bár a módszer függetlenségi feltétele nagyon erős, ezért becslésre kevésbé, ám osztályozásra igen jól használható. Ennek az az egyszerű oka, hogy nem baj, ha a valószínűség becslése nem pontos, amennyiben az osztályozás sikeres. (Mindegy, hogy valamit azért sorol egy osztályba, mert annak valószínűsége 98%-os vagy azért, mert 77%-os.) A módszer előnye, hogy jól működik szennyezett, illetve hiányzó adatok esetén, továbbá hátránya, hogy feltételezi, hogy az ismérvek egyformán fontosak (Lantz, 2015).

A naiv Bayes-módszert R környezetben a `'caret'` (Kuhn, 2008) csomag `'train()'` függvényével (method=`'cv'`), illetve a `'e1071'` (Meyer és mtsai, 2023a) csomag `'naiveBayes()'` függvényével használtam.

5.1.3 A döntési és regressziós fák

A *döntési fákat* (*Decision Trees*, DT, klasszifikációs és regressziós fák, *Classification and Decision Trees*, CART) osztályozási és becslési feladatokra is használhatjuk attól függően, hogy a címkék, azaz a válaszváltozó típusa diszkrét vagy folytonos (Breiman és mtsai, 1984, Hastie és mtsai, 2013, Lantz, 2015, Ghatak, 2017, Boehmke és Greenwell, 2020, Kuhn és Johnson, 2013). A döntési fák gyökérből, a fák levelein lévő kritériumokból, a fák ágain a kritériumokra adott válaszokból, illetve a végső levelekből áll, melyeken a modell eredménye található. A 11. Ábra egy döntési fára mutat példát. A kritériumok az ismérveket bontják fel eldöntendő kérdések formájában.

A döntési fákat mélyítve egy ismérvváltozót akár több részre is felbonthatunk. A módszer jól alkalmazható akár nagyszámú ismérvváltozó esetén is, de előfordulhat, hogy azokból végül csak néhány, vagy akár csak egyetlen dominálja a döntést. Ez elvezet ahhoz a kérdéshez, hogy milyen mély fát építsünk, azaz meddig növeljük a fa komplexitását. A komplexitás növelésével ugyanis a felülillesztettség problémája merülhet fel, amit például a modellfejlesztés idejében való leállításával (a modellparaméterek megfelelő beállításával) kerülhetünk el.



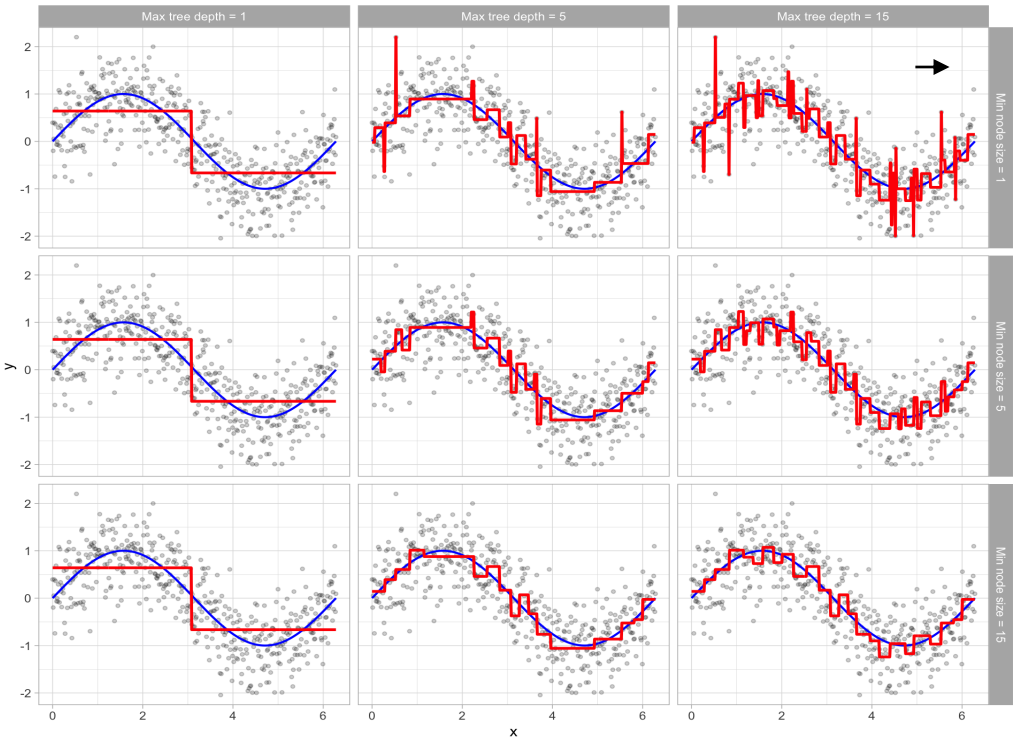
11. Ábra Példa egy döntési fára diszkrét és folytonos ismérvváltozókkal (saját szerkesztés)

A modell megfelelő paramétereinek meghatározásával a fa komplexitását a következőképpen állíthatjuk be:

- meghatározzuk a fa maximális mélységét (*max tree depth*)
- meghatározzuk a minimális ágméretet, azaz az egyes ágakra jutó minimális mintaelemszámot (*min node size*).

Bár a fenti két paraméter nem független egymástól, ezeket külön-külön is megadhatjuk.

A 12. Ábra egy regressziós modell 3-3 paraméterbeállításának eredményét mutatja (maximális mélység = 1, 5, illetve 15 és minimális ágméret = 1, 5, illetve 15.

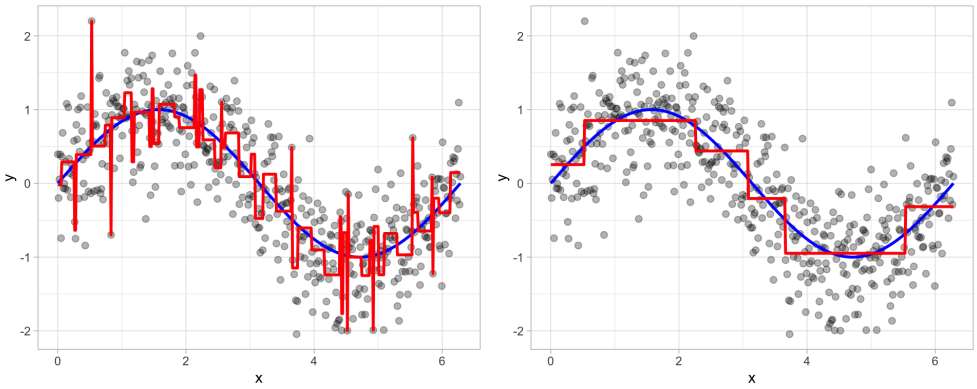


12. Ábra Regressziós döntési fa eredménye három maximális mélység (oszlopok: 1, 5, illetve 15) és három minimális ágméret (sorok: 1, 5, illetve 15) beállítással (forrás: Boehmke és Greenwell, 2020)

A fa metszésével (*pruning*) csökkenthető a fa komplexitása (Lantz, 2015, Boehmke és Greenwell, 2020). Először is építünk egy túlságosan komplex fát, ezután egy megfelelő α komplexitási paraméterrel büntetjük a komplexitást, azaz nem pusztán az SSE hibatagot minimalizáljuk, hanem a

$$SSE + \alpha|T| = \sum_{i=1}^{n}(X_i - \bar{X})^2 + \alpha|T|$$

összeget, ahol $|T|$ a fa ágainak számát jelöli (13. Ábra).



13. Ábra Túlillesztett regressziós döntési fa modellel (bal) és annak metszett módosításával (jobb) kapott illesztés (forrás: Boehmke és Greenwell, 2020)

Felmerül a kérdés, hogyan lehet kiválasztani, hogy mely ismérvváltozó kerüljön egy csomópontba. Ebben az ún. vágási mérőszámok, az entrópia és a Gini-index segítenek:

Az *entrópia* (Hastie és mtsai, 2013, Lantz, 2015, Boehmke és Greenwell, 2020) a bizonytalanság mérőszáma, amelynek segítségével az ismérvváltozókból való információnyerés mértékét határozhatjuk meg:

$$E = -\sum_{i=1}^{c} p_i log_2(p_i),$$

ahol p_i az i-edik csoportba esés valószínűsége, c a csoportok száma. A cél az entrópia csökkentése, miközben a fát fejlesztjük.

A *Gini-index* a fa tisztaságának mérőszáma (Gini, 1912, Breiman, és mtsai, 1984):

$$G = 1 - \sum_{i=1}^{c} p_i^2,$$

ahol p_i az i-edik csoportba esés valószínűsége (Hastie és mtsai, 2013, Lantz, 2015, Boehmke és Greenwell, 2020). Értéke 0, ha csupán egyetlen csoport van, vagy ha minden mintaelemet egyetlen csoportba sorolunk, ekkor a fa „tiszta". Értéke 1, ha a mintaelemeket véletlen módon soroljuk csoportokba, és 0,5-et vesz fel, ha az egyes csoportokba egyenletesen soroljuk be a mintaelemeket. A Gini-indexet kiszámoljuk minden ismérvváltozó minden értékére és ismérvváltozónként súlyozott összeget számolunk egy adott csomópontra. A cél az egyes csomópontokban azt az ismérvváltozót használni, amelynek a Gini-indexe a legkisebb.

A döntési fák előnye, hogy folytonos és diszkrét, illetve vegyes típusú adatokra is alkalmazható, nem érzékeny a hiányzó adatokra és viszonylag kis adatsorra is működik. Ez utóbbi előnye az agrártudomány azon területein, ahol az adat-előállítás különösen időigényes- és/vagy költséges (termés többnyire csak évente egyszer van...), igen fontos szempont. Hátránya a módszernek, hogy ahogy egyre komplexebb a fa, az interpretáció is nehézkessé válik, valamint az, hogy könnyen alul-, illetve felülilleszthető (Lantz, 2015).

Továbbfejlesztett döntési fákkal még foglalkozunk a 5.3.2. fejezetben.

Döntési fákat R környezetben a `'caret'` (Kuhn, 2008, 2020) és `'rpart'` (Therneau és Atkinson, 2023) csomagok alkalmazásával az `'rpart()'` függvény segítségével állítottam elő.

5.1.4 Diszkriminanciaelemzés

A lineáris diszkriminanciaelemzés (LDA), a döntési fákhoz hasonlóan klasszifikációs módszer. Mint lineáris módszer, sok közös vonást mutat a széles körben használt MANOVA-val, bár céljaik (irányuk) és értelmezésük eltérő. A MANOVA azt vizsgálja, hogy a modelljében magyarázó változóként szereplő faktor(ok) szintjei között a különbségek szignifikánsak-e az eredeti n tulajdonság (ismérv) optimalizált lineáris kombinációján (DV), míg az LDA a magyarázó változóként szereplő tulajdonságok (ismérvek) olyan térbeli vetítéssel képzett hipersíkjait (DV_i) keresi, amelyek a válaszváltozóként szereplő faktor szintjeit (a csoportokat) a lehető legjobban szét tudják választani. Az előzetes csoportosítási tényező (faktor) a MANOVA esetében tehát magyarázó (független), míg az LDA esetében válasz- (függő) változó, míg az ismérvek a MANOVA esetében a függő, az LDA esetében a magyarázó (független) változók (Rao, 1948, Hastie és mtsai, 2013, Nisbet és mtsai, 2018).

A módszer nemcsak az előzetes besorolásnak az ismérvváltozókkal való magyarázatára alkalmazható, hanem a kész modell használatával új objektumok besorolására is.

A módszer megköveteli az ismérvváltozók normalitását és a homoszkedaszticitást, bár e két feltétel esetleges sérülésére nem túlságosan érzékeny, amennyiben a többi feltétel nem sérül. Érzékeny azonban a kollinearitásra erősen redundáns információt hordozó ismérvek esetén, a kiugró értékekre, illetve a csoportok nagyon különböző mintaelemszámára (Rao, 1948, Nisbet és mtsai, 2018).

A klasszifikáció eredményét a 4.6. fejezetben ismertetett mutatók szerint értékelhetjük. Az ún. Wilk-féle Λ megmutatja, hogy a csoportosító változó varianciájának hány százalékát nem magyarázzák a független változók, azaz $1-\Lambda$ a magyarázott varianciahányad. Erre egy Khi-négyzet-tesztet végzünk, melynek nullhipotézise, hogy a csoportokra vonatkozó diszkriminanciafüggvény értékei megegyeznek. Szignifikáns eredmény mutat sikeres elkülönítést. A kanonikus korreláció η a Wilk-féle Λ-ból is előállítható, de a négyzete a csoporton belüli (SS_b) és teljes eltérésnégyzetek (SS_{Total}) hányadosaként is:

$$\eta = \sqrt{\frac{SS_b}{SS_{Total}}} = \sqrt{1-\Lambda}.$$

A kanonikus korreláció a magyarázó változók és a diszkriminanciafüggvény közötti korrelációt fejezi ki. A korreláció erőssége a magyarázó változóknak, illetve azok lineáris kombinációjának a diszkriminatív erejét mutatják.

A nemlineáris (pl. kvadratikus) diszkriminanciaelemzés (QDA) hasonló az LDA-hoz, de nem feltételezi, hogy a csoportok kovarianciamátrixa megegyezik. Ebben az esetben a csoportok közötti határvonal egy hipersík helyett egy nemlineáris (pl. kvadratikus) felület.

Diszkriminanciaelemezésre mutatunk példát a 6.4., valamint a 6.6. fejezetben.

R programozási környezetben adat-előkészítéshez a `'caTools'` (Tuszynski, 2021) csomag függvényeit, diszkriminanciaelemzéshez a `'MASS'` (Venables és Ripley, 2002) csomag `'lda()'` és `'qda()'` függvényeit, a `'DFA.CANCOR'` (O'Connor, 2023) csomag `'DFA()'` függvényét, valamint a `'klaR'` (Wihs és mtsai, 2005) csomag `'greedy.wilks()'` függvényét használtam.

5.1.5 Lineáris és nemlineáris, valamint többszörös regresszió

A regresszióanalízis egy felügyelt becslő módszer, melynek modellje:

$$Y = f(X_1, X_2, \dots X_n) + \varepsilon,$$

ahol Y a folytonos válaszváltozó (amit becsülni szeretnénk), $X_1, X_2, \dots X_n$ folytonos ismérvek (magyarázó változók), ε pedig normális eloszlású, 0 várható értékű hibatag, mely a magyarázó változóktól független és azonos szórású (homoszkedaszticitási feltétel). Az f függvény lehet lineáris vagy nemlineáris (*Nonlinear Regression*, NLR). Speciálisan tehát a lineáris regressziós modell (*linear regression*, LR):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon,$$

alakú, ahol β_0 a modell konstans tagja, β_i $(i = 1,2, \dots, n)$ pedig az egyes változók együtthatói.

A módszer a β_i $(i = 0,1,2, \dots, n)$ paramétereket becsüli a rendelkezésre álló adathalmazból. A becslés hibáját a következőképpen definiálhatjuk:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y - f(X_1, X_2, \dots X_n)\right)^2}$$

A modellparaméterek becslése az $RMSE$ (*root mean square error*) minimalizálásával történhet a legkisebb négyzetek elve alapján (*ordinary least squares*, OLS, Lantz, 2015, Ghatak, 2017, Boehmke és Greenwell, 2020).

Egyszerű lineáris regresszióról beszélünk (*simple linear regression*, SLR), ha $n = 1$. Ha pedig $n > 1$, akkor többszörös (lineáris vagy nemlineáris) regresszióról (*multiple linear regression*, MLR). A többszörös lineáris regresszió feltétele, hogy a magyarázó változók egymástól páronként függetlenek legyenek. E feltétel sérülését multikollinearitásnak nevezzük. További feltétel, hogy a magyarázó változók térben és időben is függetlenek legyenek egymástól, azaz, hogy ne legyen autokorreláció (Kutner és mtsai, 2005, Kuhn és Johnson, 2013, Hastie és mtsai, 2013, Faraway, 2015).

Multikollinearitás esetén a többszörös lineáris regresszió eredménye többféle torzulást is mutathat, mint például:

- nagyon nagy standard hiba az együtthatók becslésében
- a modell szignifikáns, de egyik együttható sem az
- egy új magyarázó változó hozzáadása a régiek együtthatóit nagyban módosítja
- az együtthatók előjele nem igazolja a szakmai elvárásokat
- ugyanarra a problémára vonatkozó, de más-más adatsorra nagyon különböző együtthatók jönnek ki.

A multikollinearitás tesztelését a VIF-mutató (*Variance Inflation Factor*; $1 \leq \mathrm{VIF} < +\infty$) értelmezésével végezhetjük (Ghatak, 2017). A VIF érték egy adott változóra megmutatja, mennyivel nagyobb a hiba varianciája a magyarázó változó egy egységnyi értékére ahhoz képest, hogy ha nincs multikollinearitás. Ha ez a változó független a

többitől, akkor VIF $= 1$. Másfelől a VIF azt mutatja, hogy hányszoros mintaelemszámra lenne szükség olyan hatásos modellhez, amilyen multikollinearitás nélkül lenne ekkora mintával. Ökölszabályként mondhatjuk, hogy $VIF < 5$ esetén multikollinearitás fennállhat, de az nem súlyos, ám $VIF > 5$ esetén már az együtthatók becslése nem megbízható (O'Brian, 2007, James és mtsai, 2013). Megjegyezzük, hogy a multikollinearitás teszteléséhez gyakran az ún. tolerancia-mutatót használjuk, mely a VIF reciproka.)

Multikollinearitás esetén megfontolhatjuk egyes nagyon erősen korreláló változók valamelyikének elhagyását, hiszen ezek a többihez képest kevés új információt hordoznak. Alkalmazhatunk ún. lépésenkénti (*stepwise*) regressziót is (James és mtsai, 2013), amely a változókat fontossági sorrendben egymás után vonja be a modellbe (vagy ellenkező irányban rakja ki a modellből), bár erős multikollinearitás esetén ez a módszer gyakran semmitmondó modellhez vezet. Összefüggő két változó helyett a két változó standardizált értékeinek átlagából képzett változó használata is megoldás lehet (James és mtsai, 2013). Multikollinearitás esetén igen hatásos az ún. főkomponens-regresszió (*Principal Component Regression,* PCR, Hastie és mtsai, 2013, Boehmke és Greenwell, 2020, James és mtsai, 2013, ld. 5.2.2. és a 5.2.3. fejezet), illetve a parciális legkisebb négyzetek regresszió (*Partial Least Squares Regression,* PLSR, Kuhn és Johnson, 2013, Hastie és mtsai, 2013, James és mtsai, 2013, ld. 5.2.3. fejezet).

A modell jóságát az $RMSE$ hibán kívül az átlagos abszolút hibával (MAE), illetve a determinációs együtthatóval is (R^2) is kifejezhetjük (Boehmke és Greenwell, 2020):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y - f(X_1, X_2, \dots X_n)|$$

$$R^2 = \frac{\sum(f(X_1, X_2, \dots X_n) - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

A regresszióanalízis fontos lépése az ún. regressziós diagnosztika (Ghatak, 2017), melynek keretében

- Student-féle t-próbával teszteljük a determinációs együttható (R^2) szignifikanciáját,
- összehasonlítjuk a kapott modellt az ún. nullmodellel (*base model*), azaz az Y válaszváltozó átlagával, az $\bar{Y}$ (átlag) konstans modellel F-próbával (ANOVA), valamint
- teszteljük az összes együttható becslését szintén Student-féle t-próbával.

Érdemes meggondolni azoknak a változóknak az elhagyását, amelyek együtthatói nem szignifikánsak. (Lényegében ezen az alapon működik az előbb említett stepwise-módszer is.)

Megjegyezzük, hogy még ha a regressziós diagnosztika minden lépésében szignifikáns eredményt kapunk, akkor is lehet még szisztematikus hiba a regressziós modellünkben, melyet a hibatagokra vonatkozó feltételek (normalitás, a magyarázó változóktól való függetlenség, homoszkedaszticitás, zérus várható érték) ellenőrzésével fedhetünk fel.

Többszörös regresszióra (lépésenkénti módszerrel) példát láthatunk a 6.3. fejezetben, lineáris és nemlineáris regresszióra a 6.4. és a 6.5. fejezetben.

Az azonos adathalmazra illesztett regressziós modellek összehasonlítása történhet a magyarázott varianciahányad (R^2), a modellek nullmodellel való összehasonlításakor kapott F értékei, a paraméterek szignifikanciái, az átlagos hibanégyzetek gyökei ($RMSE$),

a (módosított) Akaike, illetve Bayesian információs kritériumok (AIC, AIC_c, BIC) alapján, illetve a modellek hibatagjaihoz tartozó eltérésnégyzetösszegek hányadosa, azaz

$$F = \frac{SS_{ResidModel1}}{SS_{ResidModel2}} > 1$$

alapján.

Egymásba ágyazott modellek, azaz amikor a B modell paraméterhalmaza az A modell paraméterhalmazának részhalmaza, az összehasonlítást végezhetjük a

$$\frac{SS_{ResidModel\_A} - SS_{ResidModel\_B}}{df_{ResidModel\_A} - df_{ResidModel\_B}} \quad \text{és a} \quad \frac{SS_{ResidModel\_B}}{df_{ResidModel\_B}}$$

kifejezések F eloszlást követő hányadosával is, ahol df a szabadsági fokot jelöli (Motulsky és Christopoulos, 2003).

A regressziós modellek paramétereit (speciálisan például lineáris modellek konstans, illetve meredekségi paramétereit) is összevethetjük egyrészt egy hipotetikus számmal a Student-féle t-eloszlást követő

$$t = \frac{becsült\ paraméter - hipotetikus\ paraméter}{SE},$$

másrészt egymással, hasonlóan a

$$t = \frac{paraméter\_1 - paraméter\_2}{\sqrt{SE_1^2 + SE_2^2}}$$

statisztikával, ahol SE, SE_1 és SE_2 a modellek standard hibái. (Megjegyezzük, hogy az egyszerű lineáris modellek meredekségeinek fenti összevetése megegyezik a magyarázó változót kovariáns, a modellek azonosítóját kétszintű faktorként magába foglaló ANCOVA modell interakciós együtthatójának szignifikanciavizsgálatával (Motulsky és Christopoulos, 2003).

Speciálisan érdekes lehet, hogy egy faktor különböző szintjeire épített regressziós modell helyettesíthető-e egy globális (az aggregált adatokra illesztett) modellel. Ez esetben a

$$\frac{SS_{AggregModel} - SS_{SepModels}}{df_{AggregModel} - df_{SepModels}} \text{ és a}$$

$$\frac{SS_{\mathrm{SepModels}}}{df_{\mathrm{SepModels}}}$$

hányadosát képezzük, ahol $SS_{\mathrm{SepModels}}$ a k darab modell eltérés-négyzetösszegének, $df_{\mathrm{SepModels}}$ pedig a szabadsági fokainak összege (Motulsky és Christopoulos, 2003).

A regresszióanalízis előnye az egyszerűség, a könnyű interpretálhatóság, hátránya az erős feltételein kívül, hogy nagyon érzékeny a kiugró értékekre (Lantz, 2015).

19. Példa modellek összehasonlítására

A galakto-oligoszacharidok (GOS) a tejiparban használt olyan prebiotikumok, amelyeket általában alacsony enzimfelhasználhatósággal és ezért magas működési költségekkel tudnak előállítani. Pázmándi és munkatársai (2017) az enzimek többlépcsős ultraszűréssel történő visszanyerésének és az ismételt reakciólépésekben történő újbóli felhasználásának lehetőségét vizsgálták.

Különböző koncentrációjú laktózból álló savóból származó szubsztrátokon a reakció során a laktózból glükóz, galaktóz és GOS, valamint egyéb szénhidrátok

keletkeznek, melyek Y tömegszázalékának változását az első 24 órában telítődési modellel írtuk le.

$$Y = p_1 * \left(1 - exp(-p_2 * t)\right) + \varepsilon,$$

ahol t a reakcióidő, $p_1 \cdot p_2$ pedig a $t = 0$ időpontban a görbe meredeksége ($p_2 > 0$), ε normális eloszlású, zérus várható értékű hibatag.

A membránkezelés hatására a többlépcsős szűrés későbbi szakaszaiban a pH-érték emelkedését figyelték meg, ami eltér az alkalmazott kezelés optimális értékétől, ezért ekkor egy pH-korrekciót is alkalmaztak kontroll (korrekció nélküli kezelés) mellett. A kérdés az volt, hogy a pH növekedése befolyásolta-e a katalitikus teljesítményt. A modellt a pH-korrekcióval és az anélkül kapott megfigyelésekre, valamint a két adathalmazt együttesen tartalmazó bővített adathalmazra is illesztettük (globális modell). Ezután azt teszteltük, hogy a két modell (pH-korrekcióval és anélkül) helyettesíthető-e egyetlen globális modellel (Motulsky és Christopoulos, 2003), ami azt indikálná, hogy a pH megváltozásának nem volt szignifikáns hatása a produkcióra. Szignifikáns esetben a két különböző pH érték esetén illesztett modellek becsült p_1 és p_2 paramétereit is összehasonlítottuk fent leírt speciális Student-féle t próbával.

Az enzimek visszanyerése természetesen nem jelenti azt, hogy nincs veszteség. Ezért azok újabb felhasználásához az enzimek pótlására van szükség. Cao és munkatársai (2017) a projekt folytatásaként a különböző adagokban végzett enzimpótlás esetén vizsgálták a keletkezett szénhidrátok mennyiségét. A folyamatokat (kis módosítással) ismét telítődési modellel írtuk le:

$$F(t) = C_0 + p_1 * \left(1 - exp(-p_2 * t)\right) + \varepsilon,$$

ahol $F(t)$ a szénhidrátok koncentrációja, C_0 pedig a $t = 0$ időpontban való érték. A különböző szénhidrátok kezdeti $p_1 \cdot p_2$ becsült reakciósebessége nagyon jól illeszkedett egy, az origón átmenő lineáris függvényre, melynek változója az enzimaktivitás volt. Ez azt jelentette, hogy a modellek becsült kezdeti sebességével jól reprezentálható az enzimaktivitás.

20. Példa modellparaméterek összehasonlítására

Hajnal és munkatársai (2013) három év alapján kilenc, valamint Szalay és munkatársai (2019) 24 év alapján három kajszibarackfajta (*Prunus armeniaca*) virágrügy-differenciálódását vizsgálták laboratóriumi körülmények között. Évente nyolc alkalommal gyűjtöttek vesszőket, és a virágok a fejlődésük nyolc lépésben leírt stádiumait elért százalékos arányát jegyezték le. A fejlődés folyamatára szigmoid görbét illesztettünk:

$$Y(t) = 100/\left(1 - exp(-s * (t - m))\right) + \varepsilon,$$

ahol t a Juliánusz-napokban kifejezett idő, m az inflexiós pont, $s > 0$ meredekségi tényező (a görbe meredeksége az inflexiós pontban $s/4$), ε zérus várható értékű hibatag.

A görbék becsült inflexiós pontjainak összehasonlításával a fajták fejlődésének időbeli eltolódására, a sebességi tényezők összehasonlításával az egyes fejlődési fázisok eltérő ütemeire következtethettünk.

A projekt folytatásaként Szalay és munkatársai (2016) tizenegy év alapján három, valamint Bakos és munkatársai (2024) három év alapján tizenhat kajszibarackfajta (*Prunus armeniaca*) virágrügyeinek fagytűrését vizsgálták a nyugalmi időszakban mesterséges fagyasztási kísérletben. Októbertől februárig hat időpontban szedtek hajtásokat a fákról, melyeket minden egyes mintavételi időpontban négy vagy öt különböző hőmérsékleten

kezeltek egy napon keresztül, ami után meghatározták a rügy károsodásának mértékét. LT50-nek nevezik azt a hőmérsékletet, amelynél a rügyek 50%-a fagykárosodást szenved egy adott időpontban és a fajta egy adott fenológiai stádiumában. Az LT50 érték meghatározása egy adott faj esetén döntő jelentőségű a fajra jellemző fagytűrés jellemzésére.

A beállított hőkezelések eredményeképpen olyan fagykárosodási értékeket kapunk, amelyek egy szigmoid görbe mentén helyezkednek el. Az LT50 értékek, azaz a görbék inflexiós pontjai egyrészt becsülhetők lineáris regressziós modell segítségével azt feltételezve, hogy a fagyállósági szigmoid görbe 20% és 80% közötti szakasza lineárisnak tekinthető (Gu, 1999, Szalay és munkatársai, 2016). Egy másik lehetőségkét az LT50 értékek becslésére simított (spline) regressziót alkalmaztunk ún. GAM-módszerrel (*General Additive Model*, Hastie és mtsai, 2013).

A spline-illesztés lényege, hogy szakaszonként illesztünk (elkerülendő a túlillesztést, nem magas fokú) polinomiális görbét egy adathalmazra, mégpedig úgy, hogy meghatározott csomópontokban (*knots*) a polinomiális görbék simán (folytonosan deriválhatóan) kapcsolódnak össze (James és mtsai, 2013, Boehmke és Greenwell, 2020).

A becsült LT50 értékek becslése lehetőséget nyújtott az egyes fajták fagyérzékenységi profiljának jellemzésére.

Regressziós modellek összehasonlítására további példa található a 6.5. fejezetben.

Lineáris és nemlineáris regressziós modelleket R környezetben legegyszerűbben a `'caret'` (Kuhn, 2008, 2020) csomagbeli `'lm()'` függvénnyel, illetve a `'stats'` (R Core Team, 2023) csomagbeli `'nlm()'` függvénnyel fejlesztettem.

5.1.6 Logisztikus regresszió

A lineáris és nemlineáris regresszióanalízis kimenete minden esetben folytonos változó. A logisztikus regressziót (*Logistic Regression*, LogR) bináris vagy multinomiális válaszváltozó (címke) esetén, azaz osztályozási feladatokra alkalmazzuk (Hosmer és mtsai, 2013, Hastie és mtsai, 2013, Ghatak, 2017, Boehmke és Greenwell, 2020). Abból indulunk ki, hogy egy bináris változó pozitív kimenetelének valószínűségét becsüljük az alábbi alkalmas formában:

$$P(X) = \frac{Exp(\beta_0 + \beta_1 X)}{1 + Exp(\beta_0 + \beta_1 X)},$$

ugyanis $\lim_{a \to -\infty} \frac{Exp(a)}{1+Exp(a)} = 0$ és $\lim_{a \to +\infty} \frac{Exp(a)}{1+Exp(a)} = 1$ alapján a fenti kifejezés 0 és 1 közötti értéket vehet fel. A fenti alakot logit formába átrendezve kapjuk, hogy

$$ln\left(\frac{P(X)}{1 - P(X)}\right) = \beta_0 + \beta_1 X.$$

Ezzel visszavezettük a feladatot egy egyszerű lineáris regresszióra, melynek paraméterei β_0 és β_1, melyeket a maximum likelihood (ML) módszerrel becsüljük (ld. 2.2. fejezet).

Ez azt eredményezi, hogy a becsült együtthatók segítségével az 1 értékű címkéknek a valószínűsége 1-hez lesz közel, a 0 értékűeknek 0-hoz.

Multinomiális esetben a $P(X)$ valószínűséget

$$P(X) = \frac{Exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + Exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

alakban írjuk fel, és ezután szintén a ML-módszerrel kalibráljuk az együtthatókat (Hastie és mtsai, 2013, Boehmke és Greenwell, 2020).

Logisztikus regressziós modellt R környezetben bináris output esetén a `'stats'` (R Core Team, 2023) csomagbeli `'glm()'` függvény segítségével (family=`'binomial'`), multinomiális esetben pedig a `'nnet'` (Venables és Ripley, 2002) csomagbeli `'multinom()'` függvénnyel fejlesztettem.

5.1.7 Támogatóvektor-gépek

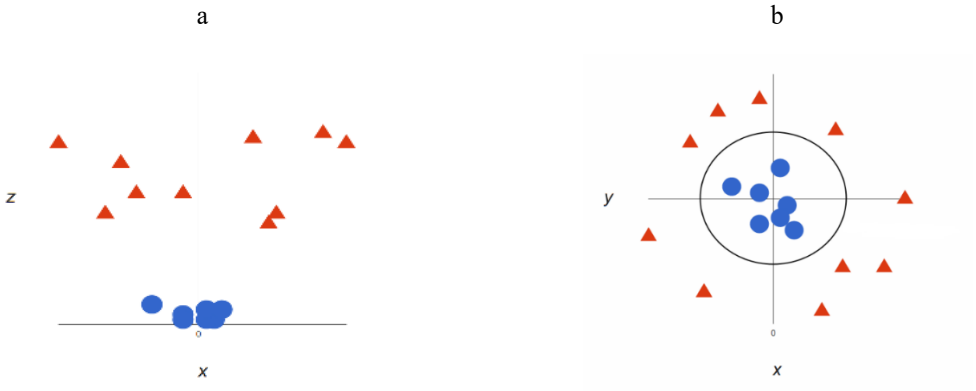
Bár a támogatóvektor-gépek (*Support Vector Machines*, SVM) osztályozási és regressziós feladatot is képesek ellátni, azokat a gyakorlatban legtöbbször osztályozásra használják. Az SVM az n dimenziós térben elhelyezkedő pontokhoz egy olyan hipersíkot keres, amely jól elválasztja az előzetes címkézés szerinti csoportokat (Hastie és mtsai, 2013, Kuhn és Johnson, 2013, Lantz, 2015, Nisbet és mtsai, 2018, Boehmke és Greenwell, 2020, James és mtsai, 2013, Cristianini és Shawe-Taylor, 2000).

Lássunk tehát példákat olyan osztályozási feladatokra, melyek megoldására az SVM-módszer kiválóan alkalmas. Láthatjuk, hogy a 14. Ábra bal oldala két olyan hipersíkot is tartalmaz, mely a piros és kék csoportokat hibátlanul elválasztja. AZ SVM ezek közül azt választja, amely a legtávolabb áll a legközelebb lévő pontoktól, azaz az ábrabeli példa esetében a fekete hipersíkot.

14. Ábra (a) A lineáris SVM modell megoldása a fekete hipersík, mely a piros és kék csoportokat hibátlanul elválasztja, és az ilyen tulajdonságú hipersíkok között a legközelebbi pontoktól való távolsága ennek a legnagyobb. (b) Nemlineáris elválasztási feladat. (Forrás: Stecanella, 2021)

Az SVM-módszer azonban nem csupán lineáris elválasztásra képes. Vegyük például a 15. Ábra jobb oldalán látható nemlineáris elválasztási feladatot. Ezt az elválasztást speciális módon tudjuk megoldani. Az SVM alapötlete, hogy bevezetünk egy harmadik (z) tengelyt, amelyen a $z = f(x, y) = x^2 + y^2$ értékeket ábrázoljuk (15. Ábra, a). Látható, hogy ekkor a $z = 1$ egyenes sikeresen elválasztja a két csoportot. Az eredeti ábrán tehát $x^2 + y^2$=1 kört rajzolva az elválasztás megtörtént (15. Ábra, b). Vegyük észre, hogy az eredetileg nemlineáris feladatot ezzel az ötlettel lineáris elválasztási feladatra vezettük vissza, azaz ezzel a trükkel a feladat komplexitását csökkentettük. Az f függvényt

(másodfokú) bázisfüggvénynek (*kernel function*) nevezzük, az új dimenzió bevezetésének ötletét bázistrükknek (*kernel trick*, James és mtsai, 2013).



15. Ábra Nemlineáris elválasztási feladat. Új tengely bevezetésével (a) az SVM nemlineáris hipersíkot állít elő (b). (Forrás: Stecanella, 2021)

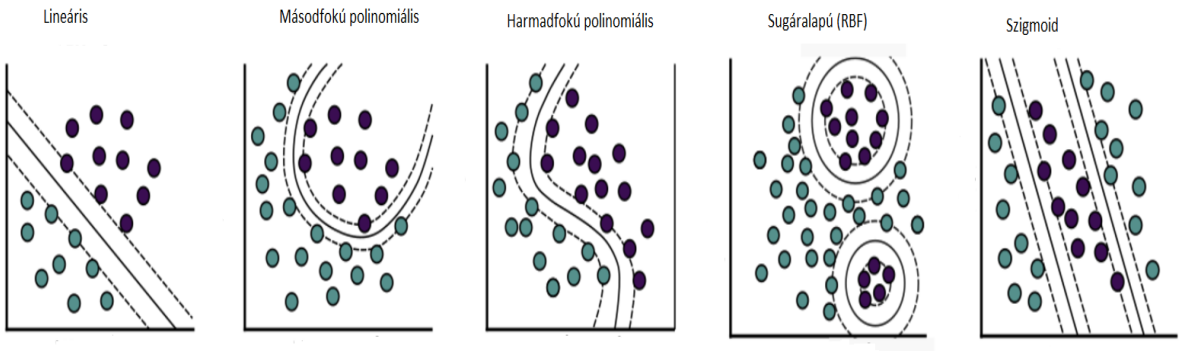
A leggyakrabban használt bázisfüggvények:

- lineáris $K(x_i, x_j) = x_i \cdot x_j$
- d-paraméterű (d-fokú) polinomiális: $K(x_i, x_j) = (x_i \cdot x_j + 1)^d$
- γ-paraméterű sugáralapú (*radial basis function,* RBF):

$$K(x_i, x_j) = \exp\left(-\gamma \cdot \|(x_i - x_j\|^2\right)$$

- b, c-paraméterű szigmoid $K(x_i, x_j) = \tanh(b(x_i, x_j) + c)$

A 16. Ábra a különböző bázisfüggvények alkalmazására mutat be néhány példát.



16. Ábra Példák bázisfüggvény alkalmazásokra (lineáris, másod-, illetve harmadfokú polinomiális, sugáralapú (RBF), szigmoid) (Forrás: Anshul Trivedi, url)

Az SVM-módszernek van még egy fontos paramétere, az ún. költségérték (*cost value*, C). A C érték azt határozza meg, hogy a modellben mennyire büntetjük az elválasztó sávon belüli vagy az elválasztó hipersík rossz oldalán lévő pontokat. Ha a C értéket nagynak választjuk, akkor az SVM megpróbál olyan hipersíkot és elválasztó sávot találni, hogy nagyon kevés pont legyen a sávon belül, ami túlságosan összetett modellt jelenthet keskeny sávszélességgel, különösen akkor, ha a pontok eleve nem könnyen szeparálhatóak.

Egy alacsonyabb C nagyobb hibát ad a tanuló halmazon, de nagyobb sávszélességet ad eredményül, ami robusztusabb megoldás lehet.

Az SVM-módszer előnye, hogy szennyezett adatokra is jól működök, és nem hajlamos túlillesztettségre. Hátránya, hogy a bázisfüggvényt és paramétereit meg kell adni, és erre a módszer nagyon érzékeny, így esetenként rengeteg modellt le kell gyártani és összehasonlítani, valamint, hogy az osztályozásra nem ad ki valószínűségeket, illetve, ha nagyon komplex, akkor már nagyon nehéz interpretálni (*black box*, Lantz, 2015).

Az SVM alkalmazására a 6.6. fejezetben mutatunk példát.

SVM klasszifikációt R környezetben az `'e1071'` (Meyer és mtsai, 2023a) csomagbeli `'svm()'` függvény, vagy a `'caret'` (Kuhn, 2008, 2020) csomag `'train()'` függvényével (method='svmRadial'), vagy a `'kernlab'` (Karatzoglou és mtsai, 20023, 2004) csomag `'svmlinear()'` függvényével végeztem.

5.2 Felügyelés nélküli (*unsupervised*) gépi tanulási módszerek

5.2.1 Klaszterelemzés

A klaszterelemzés célja előre nem címkézett adatok hasonlóságon alapuló csoportosítása. A módszer az egyes mintaelemek közötti távolságmetrikán alapul, így erősen függ az előzetes távolságdefiníciótól (Hastie és mtsai, 2013, Boehmke és Greenwell, 2020, James és mtsai, 2013). Az ismérvváltozók lehetnek diszkrét (bináris, ordinális, nominális), illetve folytonos típusúak is. Az adatok típusától, illetve a feladat céljától függően számos olyan távolságfüggvényt definiálhatunk, melyek megfelelnek a távolság kritériumának (nemnegatív, monoton, eltolás-invariáns, valamint teljesíti a háromszög-egyenlőtlenséget), ám az adataink hasonlóságát, illetve különbözőségét más-más módon ragadják meg. Megkülönböztetünk hasonlóságot, illetve különbözőséget kifejező távolságdefiníciókat is. A hasonlóságra normális eloszlású változók esetén a leggyakrabban a Pearson-féle korrelációt alkalmazzák:

$$d(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}.$$

Rangszámokon alapuló asszociációs mérőszámokat (Spearman, 1904, Kendall, 1938, Somers, 1962 stb.) is használhatunk (ld. 2. Példa).

A klaszterelemzésben gyakrabban alkalmazzák a különbözőségen alapuló távolságdefinícókat (Podani, 1997). Az intervallumskálás változókra a p paraméterű Minkowski-távolság:

$$d(x,y) = \sqrt[p]{\sum_{i=1}^{N} |x_i - y_i|^p}$$

speciális esete $p = 2$ esetén az ún. Euklideszi, $p = 1$ esetén a Manhattan-távolság. A Csebisev-távolságot a $p \to \infty$ esettel a

$$d(x,y) = \lim_{p\to\infty} \sqrt[p]{\sum_{i=1}^{N} |x_i - y_i|^p} = \max_i |x_i - y_i|$$

képlettel definiáljuk. Amennyiben bináris adataink vannak, akkor a távolságdefiníciót ezek előfordulásai alapján határozzuk meg, azaz gyakorisági adatok segítségével. Fejezzük ki két mintaelemben, x-ben és y-ban a 0 és 1 értékek helyspecifikus előfordulását az a, b, c, d gyakoriságokkal az alább módon:

		előfordulás x-ben	
		1	0
előfordulás y-ban	1	a	b
	0	c	d

Ekkor az ún. Hamming-távolság a $d(x,y) = \sqrt{b+c}$ formulával definiálható, amely csupán a nem egyezéseket veszi figyelembe. Ám attól függően, hogy a 0, illetve 1 helyspecifikus értékek, azok együttes, vagy csak az egyik elemben való előfordulása, illetve mindkét elemben való hiánya milyen fontos a feladat célját tekintve, számos más távolságdefiníció alkalmazható.

Ha ugyanis a különbözőségeknek nagyobb hangsúlyt kívánunk adni, mint az egyezőségeknek, akkor használhatjuk a Rogers-Tanimoto-távolságot:

$$d(x,y) = \frac{a+d}{a+2b+2c+d}.$$

Ha az 1,1 helyspecifikus megjelenés (mindkét mintaelemről állítható egy meghatározott tulajdonság) fontosabb, ám a 0,0 (egyik elemről sem állítható egy meghatározott tulajdonság) kevésbé érdekes számunkra, alkalmazható például a Lance-Williams-távolság:

$$d(x,y) = \frac{b+c}{2a+b+c}.$$

Nagyon gyakran használják még a Jaccard-féle hasonlósági mérőszámot is, mely a Lance-Williamshoz hasonlóan a d értéket nem is veszi figyelembe:

$$d(x,y) = \frac{a}{a+b+c}.$$

Ha a változóink kettőnél több értéket vehetnek fel, akkor a 2*2-es táblázatnál nagyobb gyakoriságtáblázatot állítunk elő. Tegyük fel, hogy az x mintaelem k, az y pedig l különböző értéket vehet fel. Ekkor egy $k * l$ –es ún. kontingenciatáblázatot készítünk, melynek celláiba az f_{ij} gyakoriságokat, a sorok, illetve oszlopok végére azok összegeit, $f_{i.}$-t, illetve $f_{.j}$-t írjuk. Ezek segítségével képezhetjük a Khi-négyzet-távolságot:

$$d(x,y) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{\left(f_{ij} - \frac{f_{i.} * f_{.j}}{n}\right)^2}{\frac{f_{i.} * f_{.j}}{n}},$$

ahol $n = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij}$.

A Gower-távolság (Gower, 1971, Podani, 1999) kevert típusú változókra alkalmas. A különböző típusú IV_k, $k = 1,2, \dots, s$ változókra eltérő definíciót adunk meg:

$$d_{IV_k}(x_i, y_i) = \begin{cases} \text{ha IV}_\text{k} \text{ folytonos változó, akkor } \frac{|x_i - y_i|}{max(IV_k) - min(IV_k)} \\ \text{ha IV}_\text{k} \text{ diszkrét változó, akkor } \begin{cases} 0, \text{ha } x_i = y_i \\ 1, \text{ha } x_i \neq y_i. \end{cases} \end{cases},$$

$$d(x, y) = \frac{\sum d_{IV_k}(x_i, y_i)}{s}.$$

A hierarchikus klaszterelemzés (*hierarchical clustering*, HC) első lépésében n különböző objektumunk van, melyek mind egy-egy csoportot (klasztert) képviselnek. Minden egyes lépésben kiszámoljuk az általunk választott távolságdefinícióval képzett összes páronkénti távolságot az objektumok között, és azok közül a legkisebbet választva a két objektumot (csoportot) összevonjuk, új objektum (csoport) jön létre. Az objektumok csoportok száma minden lépésben eggyel csökken. Ha két legkisebb távolság van, akkor ezek egyikét véletlen módon választhatjuk. A folyamat addig folytatható, amíg már egyetlen csoportunk marad, ennek minden mintaelem eleme. Természetesen sem az n számú, sem az egyetlen csoport képzése nem volt célunk, így a kapott mintázat alapján a folyamat végén kiválaszthatjuk a legmegfelelőbb, azaz a leginkább interpretálható csoportszámot.

Amint az az algoritmusból kiderült, egy objektum lehet egyetlen mintaelem, illetve az összevonások eredményeképpen egy csoport is. Azaz a csoportok közötti távolságfogalomra (*linkage*) is szükség van az algoritmus alkalmazásához. Két csoport távolságát kifejezhetjük a két csoport közepeinek távolságával (*centroid linkage*), illetve a két csoport elemeinek átlagos páronkénti távolságával (*average linkage,* UPGMA). Másfelől két csoport összevonása történhet a két legközelebbi elem távolsága (*single linkage*), a két legtávolabbi elem távolsága (*complete linkage*), vagy ezek mediánja alapján is, mindig ezek közül a legkisebbet választva. Az egyik leghatásosabb összevonási módszer az ún. Ward-módszer (Ward, 1963), amely minimalizálja a csoportokon belüli eltérés-négyzeteket. Ezt úgy éri el, hogy minden lépésben minden lehetséges összevonásra képezi az i-edik csoportbeli j-edik elem k-adik változóbeli, x_{ijk} -vel jelölt értékek segítségével az SS_{Teljes} teljes, valamint az SS_{hiba} eltérés-négyzetösszegeket

$$SS_{hiba} = \sum_{i=1}^{N_G} \sum_{j} \sum_{k} \left| x_{ijk} - \frac{1}{N_G} \sum_{j=1}^{N_G} x_{ijk} \right|^2$$

$$SS_{Teljes} = \sum_{i=1}^{N_G} \sum_{j} \sum_{k} \left| x_{ijk} - \frac{1}{N_G} \sum_{i} \sum_{j=1}^{N_G} x_j \right|^2,$$

valamint ezekből az

$$R^2 = \frac{SS_{Teljes} - SS_{hiba}}{SS_{Teljes}}$$

értékeket, és ezek közül a legkisebbhez tartozó összevonást végzi el abban a lépésben. (Itt G az összevonás utáni csoportot, N_G ennek elemszámát jelöli.) Ha több legkisebb R^2 van, akkor azok közül véletlenül választ.

A hierarchikus klaszterelemzés eredményének vizualizálása dendrogram segítségével történik. A dendrogram faszerkezetű, a levelei az egyes mintaelemek, és az összevonások minden lépése leolvasható róla. Ezen kívül az ágak hosszúsága az egyes lépésekben az összevonásra kerülő csoportok távolságát is reprezentálják.

A k-közép klaszterelemzés (*K-means clustering*, KMC) esetében a hierarchikus módszerrel ellentétben már első lépésében meg kell adnunk a kívánt csoportszámot, k-t (Hastie és mtsai, 2013, Lantz, 2015, Boehmke és Greenwell, 2020). Az ismérvváltozók lehetnek diszkrét (bináris, ordinális, nominális), illetve folytonos típusúak. A módszer először is választ k darab induló mintaelemet (ezeket az angol szakirodalom *centroid*-nak, vagy *seed*-nek nevezi). Ezeket a felhasználó is megadhatja, de ennek hiányában véletlenszerűen lesznek kijelölve. A módszer lényege, hogy minden egyes lépésben minden egyes mintaelemet ahhoz a centroidhoz rendeli, amelyhez a legközelebb áll. Így már az első lépésben éppen k csoport keletkezik. Ezután kiszámolja a keletkezett csoportok csoportközéppontjait, és azok lesznek az új centroidok. Minden pont ezután újra besorolásra kerül a legközelebbi centroidhoz, és így tovább, egészen addig, amíg a csoportba sorolás már nem, vagy előre meghatározott módon „alig" változik tovább. A módszer befejezéséhez választhatjuk a csoporton belüli eltérésnégyzeteket is: a lépések során ezt minimalizáljuk, és leállunk, ha ez adott nagyságrendnél jobban nem csökken (azaz a csoportosítás már csak „alig" változik).

A klaszterezés eredményeinek validálását külső és belső (*internal és external measurements*) mérőszámokkal végezhetjük (Halkiditis és mtsai, 2001). A klaszterelemzés ugyan felügyelés nélküli módszer, de előzetes csoportbesorolás ismeretében is végezhetjük (felügyelt formában). A külső validálás során az előzetes csoportbesorolásnak való megfelelést mérjük például ANOVA, Rand-index (Rand, 1971), Pearson-féle gamma, khi-négyzet-próba, Meila-féle variációs információ (Meila, 2007) segítségével. A belső validálást az elválasztás (*separation*), illetve a kompaktság (*compactness*) mérésével végezhetjük. Az előbbi azt jelenti, hogy a különböző csoportok távolsága a lehető legnagyobb legyen, az utóbbi, hogy a csoportelemek minél közelebb legyenek egymáshoz. A kettő aránya a maximalizálandó ún. Dunn-index (Dunn, 1974).

A klaszterezés minőségének jellemzésére használt másik jól interpretálható mutató a Silhouette-mérték ($SM(x)$, Rousseeuw, 1987), melyet minden egyes x mintaelemre kiszámolhatunk, értéke -1 és +1 között van:

$$SM(x) = \frac{b - a}{max\ (a, b)},$$

ahol a az x csoportjában az x-től vett páronkénti távolságok átlaga, b pedig x-nek, a hozzá legközelebbi csoport elemeivel való páronkénti távolságának átlaga. $SM(x)$ értéke 1 közelében azt jelenti, hogy az x nagyon közel van a csoportjabeli pontokhoz, és a legközelebbi másik csoporttól jól elválik, azaz besorolása nagyon jó. $SM(x)$ értéke -1 közelében azt jelenti, hogy az x inkább egy másik csoportba tartozik, azaz a besorolása gyenge. $SM(x)$ értéke 0 közelében arra mutat rá, hogy egymást átfedő csoportjaink vannak. Amennyiben a modellünket szeretnénk értékelni, képezhetjük a Silhouette-mértékek minden elemre vett átlagát. Ezzel a mérőszámmal különböző klaszterezési eredményeket hasonlíthatunk össze.

A hierarchikus és k-közép klaszterelemzés előnyeit és hátrányait a 7. Táblázatban foglaljuk össze (Lantz, 2015).

7. Táblázat A hierarchikus és k-közép klaszterelemzés előnyei és hátrányai

Módszer	Előnyök	Hátrányok
Hierarchikus	nem kell előre megadni a csoportszámot	nem rugalmas: ha egy elemet besorolt, az már nem változik
	folytonos és gyakorisági adatokra is alkalmas	függhet az adatok sorrendjétől
	a kiugró adatokra kevésbé érzékeny	hiányzó adatokra nagyon érzékeny
k-közép	nagyon egyszerű és gyors	előre meg kell adni a csoportszámot
	nagy adathalmazra is jó	a kiugró adatokra érzékeny
	rugalmas, minden lépésben új döntést hoz	függ a kezdőpontoktól
	folytonos és gyakorisági adatokra is alkalmas	átfedő csoportok esetén nem mindig konverges

Hierarchikus klaszterelemzést R környezetben az `'cluster'` (Maechler és mtsai, 2023) csomagbeli `'hclust()'` függvény segítségével végeztem. A k-közép-módszerhez a `'kmeans()'` függvényt használtam a `'stats'` (R Core Team, 2023) csomagból.

21. Példa klaszterelemzésre: a távolságmátrixot az UPOV-előírás szerint állítottuk elő

Király és munkatársai (2015) a Kárpát-medencében található 60 különböző régi almafajta, illetve genotípus morfológiai és biológiai jellemzését végezék el az UPOV (*Union Internationale pour la Protection des Obtentions Végétales*; UPOV, 2005) leírásai alapján. A nyugalmi és vegetációs időszak folyamán összesen három alkalommal való felvételezés alapján 53 morfológiai és 3 fenológiai jellemzőt jegyeztek fel 2007 és 2011 között. A fajták végleges számkód szerinti besorolását több év megfigyelései és mérési adatai alapján határozták meg.

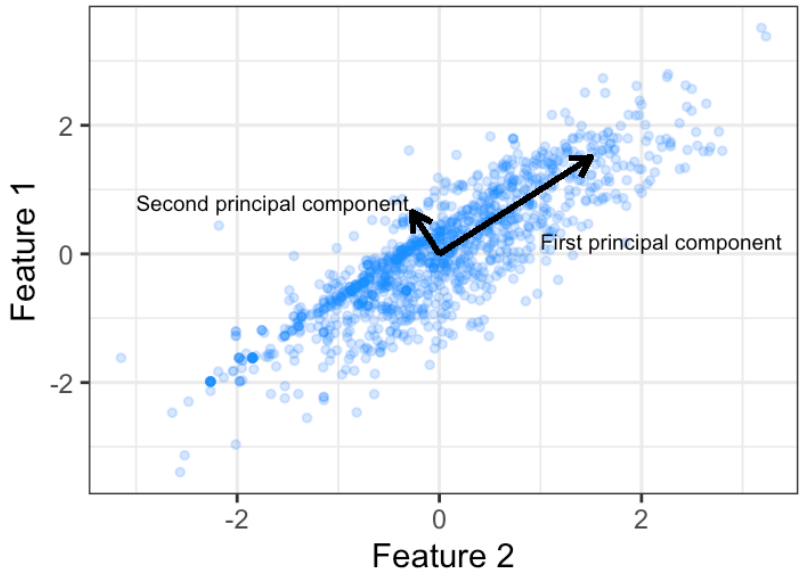
Az adatok klaszterelemzéshez való előkészítésekor figyelembe kellett venni, hogy az UPOV két fajtát akkor tekint megkülönböztethetőnek, ha legalább egy tulajdonság tekintetében egyértelmű különbség van, amelyet minőségi (QL) és pszeudo-kvalitatív (PQ) tulajdonságok esetében egy, mennyiségi (QN) tulajdonságok esetében pedig két kategóriányi különbség jelez. A távolságmátrixot tehát úgy állítottuk elő, hogy ha egy fajtapár egy QL-, vagy PQ-típusú UPOV kódjában nem volt különbség, vagy QN-típusú kódjában nem volt egy kategóriánál nagyobb különbség, akkor abban a változóban a fajtapár különbsége zérus volt. Ha különbségük egy változóban nem zérus volt, akkor 1. Egy fajtapár távolsága az egyes kódok közti különbségek összege volt.

A hierarchikus klaszterelemzést és a Ward-módszerrel végeztük el az így elkészített távolságmátrix segítségével.

5.2.2 Főkomponens-elemzés

A főkomponens-elemzés (*Principal Component Analysis,* PCA) ún. dimenziócsökkentő eljárás, melynek célja a változók minél kisebb információveszteséggel való tömörítése (Hastie és mtsai, 2013, Ghatak, 2017, Boehmke és Greenwell, 2020, James és mtsai, 2013). Kiindulunk s számú, egymással korreláló ismérvváltozóból, jelöljük ezeket most $X_1, X_2, \ldots, X_s$ –sel, és képezzük ezek lineáris kombinációját: $F_1 = \sum_{i=1}^{s} a_i X_i$, melyet első főkomponensnek nevezünk. Az $a_1, a_2, \ldots, a_s$ együtthatókat úgy határozzuk meg, hogy azok négyzetösszege 1 legyen, és az F_1 az X_i változók varianciájának a lehető legnagyobb hányadát hordozza. Minden további lépésben ehhez hasonlóan az X_i változók

lineáris kombinációjával egy további főkomponenst állítunk elő, mégpedig oly módon, hogy az együtthatók négyzetösszege 1 legyen, az F_m az X_i változóknak az $F_1, F_2, \ldots, F_{m-1}$ főkomponensek által nem magyarázott varianciahányadának a lehető legnagyobb részét hordozzák, és a főkomponensek páronként korrelálatlanok, azaz egymásra merőlegesek legyenek. A főkomponens-elemzés geometriailag térbeli forgatásnak felel meg, ahol az első főkomponens az adatok legnagyobb szórásának megfelelő hipersík, a második a következő legnagyobb szóráshoz tartozó, az előzőre merőleges hipersík, és így tovább. A 17. Ábra egy két ismérvből (Feature 1 és Feature 2) álló adathalmazon végzett főkomponens-elemzés eredményének sematikus ábrázolása. Az első főkomponens (PC1) az adatpontok legnagyobb szórásának megfelelő irányban fekszik, a második főkomponens (PC2) erre merőleges.



17. Ábra Egy kétváltozós adathalmazon végzett főkomponens-elemzés eredményének sematikus ábrája (Forrás: Boehmke és Greenwell, 2020)

Ha most a főkomponensek segítségével az eredeti változóinkat szeretnénk kifejezni ($X_j = \sum_{i=1}^{s} L_i F_i$), akkor az így kapott $L_1, L_2, \ldots, L_s$ együtthatókat loadingoknak nevezzük. Az egyes X_j változók loadingjainak négyzetösszege az X_j változó kommunalitása.

A célunk most már az, hogy az így kapott s számú főkomponensből csak azt a $t < s$ számút tartsuk meg, amelyek együtt az X_j változók varianciájának elegendően nagy hányadát hordozzák. Ezáltal az s dimenziós problémát úgy redukáljuk t dimenzióssá, hogy közben az egymással korreláló X_j változóink helyett az egymástól független főkomponenseket tartjuk meg.

A főkomponens-elemzésbe olyan változókat érdemes bevonni, amelyek kommunalitása nem túl kicsi. Ökölszabályként a 0,25-nél alacsonyabb kommunalitású változókat, ha más súlyos érv nem mond ennek ellen, akkor kihagyhatjuk az elemzésből (Tabachnick és Fidell, 2019, Field, 2012). Főszabályként annyi főkomponenst választunk, amennyihez tartozó sajátérték 1-nél nagyobb. Az 1-nél kisebb sajátértékű főkomponensek már kevesebb varianciát hordoznak, mint akár egyetlen eredeti változó, ezért ezeket többnyire elhagyjuk. A megfelelő számú főkomponens megválasztásában a sajátértékeknek a főkomponensek számától függő pontdiagramja is segíthet (*scree plot*), ezen ugyanis egy könyökpontot találva az x tengelyről leolvasható a megtartandó főkomponensek ajánlott száma.

Hogy az adataink mennyire alkalmasak dimenziócsökkentésre, annak eldöntésében a Bartlett-próba, illetve a Kaiser-Meyer-Olkin (KMO, Kaiser, 1970) mutató segít. A Bartlett-féle szfericitáspróba a változók korrelációs mátrixát ($R = (r_{ij})$) az identitásmátrixszal hasonlítja össze, és azt méri, hogy elegendően összefüggőek-e a változóink. A próba számított értéke:

$$\chi^2 = \left(n - 1 - \frac{2s + 5}{6}\right) ln(detR),$$

ahol n a mintaelemek száma, $ln(detR)$ pedig a korrelációs mártix determinánsának természetes alapú logaritmusa. Akkor lehetünk elégedettek, ha szignifikáns eredményt kapunk.

A KMO-mutató kiszámításához először képezzük a parciális korrelációs együtthatókat:

$$a_{ij} = -\frac{r_{ij}}{\sqrt{r_{ii} r_{jj}}}.$$

Ennek segítségével a KMO-mutató az alábbi módon definiálható:

$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}.$$

Minél közelebb van a KMO-mutató az 1-hez, annál kisebbek a parciális korrelációk, vagyis annál kevesebb az egyetlen változó által hordozott varianciahányad, tehát a változók annál jobban tömöríthetők. Ökölszabályként elmondhatjuk, hogy ha a KMO-mutató 0,5 alatt van, akkor az adathalmazunk nem alkalmas dimenziócsökkentésre, de 0,6 alatt se várhatunk túl jó eredményt. 0,6 felett elfogadható, 0,7 felett megfelelő, 0,8 felett jó, 0,9 felett kiváló eredményre számíthatunk (Kaiser, 1974, Hutcheson és Sofroniou, 1999, Field és mtsai, 2012).

Az eredmények értelmezéséhez igen előnyös, ha az eredményként kapott, a (sztenderdizált) változókhoz tartozó loadingokat egy megfelelő térbeli forgatással szemléletesebbé tesszük. Többféle forgatási módszert alkalmazhatunk. Ha ortogonális módszert választunk, akkor a főkomponensek merőlegesek, azaz egymástól páronként függetlenek maradnak. Ilyen például az ún. *varimax* forgatás, amely egyetlen főkomponensen belül maximalizálja a loadingok varianciáját, ezáltal egy főkomponensen belül többnyire nagy és kis abszolútértékű loadingokat kapunk. Ez megkönnyíti az értelmezést, hiszen a loadingok a főkomponensnek az egyes változókkal vett Pearson-féle korrelációi (R). Így tehát a magas abszolútértékű loadingok mutatják, lényegében mely változók által hordozott variancia került a főkomponensbe. Azok a változók, amelyek loadingja az első főkomponensben alacsony abszolútértékű, egy későbbi főkomponensben (jellemzően, de nem minden esetben) magas abszolútértékű loadingot, azaz egy másik főkomponenssel való magas korrelációt mutatnak. Így a főkomponensekről az elforgatott loadingjai mutatják meg, lényegében mely összefüggő változókat tömörítették, és melyeket hagytak egy másik főkomponensre. Ökölszabályként elmondhatjuk, hogy 0,3 alatti loadingokat ($R^2 < 0{,}09 \approx 0{,}1$) nem szükséges figyelembe venni, elég az ennél magasabbakat értelmezni (Field, és mtsai, 2012).

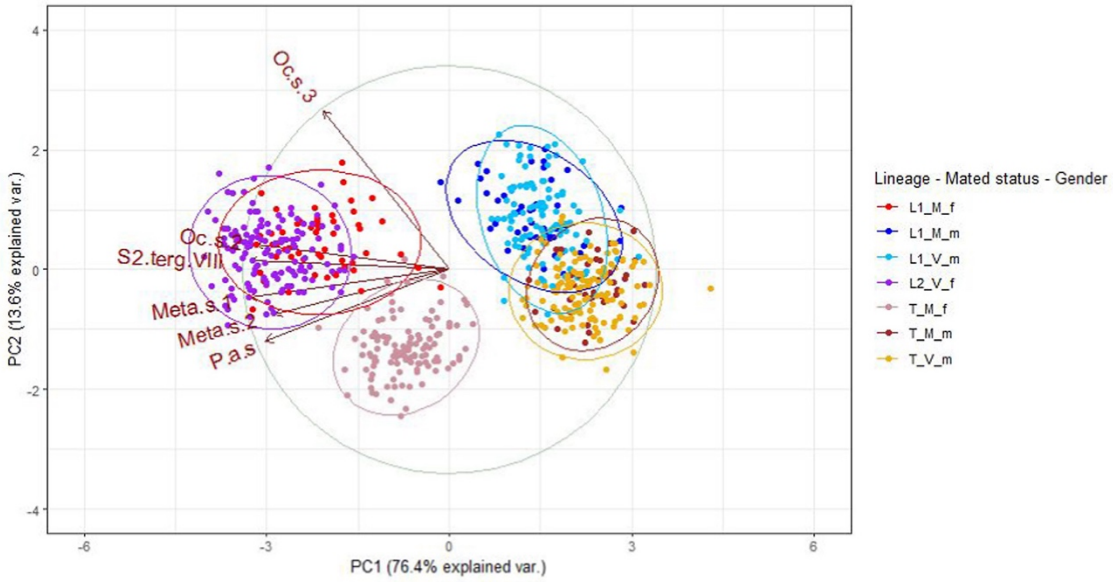
Ha végül előállítjuk az eredményül kapott a_i együtthatókkal a sztenderdizált változók lineáris kombinációt ($\sum_{i=1}^{s} a_i X_i$), akkor az ún. főkomponens-szkórokat kapjuk. Ha e szkórokat az első néhány főkomponens által alkotott koordinátarendszerben ábrázoljuk, akkor az egyes mintaelemek hasonlósága (közelsége) és különbözősége (távol

állása) grafikusan is szemléltethető. Ha még ebbe az ábrába az eredeti változók által meghatározott irányvektorokat is ábrázoljuk (*biplot*), akkor a hasonlóságokat és különbözőségeket a változók aspektusából is magyarázni tudjuk.

22. Példák biplotra

Az első példát a 10. Példa kapcsán már megismert publikációból mutatjuk be (Musa és mtsai, 2023). Abban a példában az L1, L2 és T dohánytripsz fajtakomplex változatok (*Thrips tabaci* Lindeman; Thysanoptera: Thripidae) nőstényeinek petéit vizsgálták azok morfometriai tulajdonságai szerint. Ebben a példában 808 kifejlett egyed (253 L1, 150 L2 és 405 T) nyolc vizsgált morfometriai jellemzőjéről lesz szó. Ezekkel mint magyarázó változókkal a változatokkal mint előzetes besorolással főkomponens-elemzést végeztünk és az eredményeket biploton ábrázoltuk (18. Ábra). A kapott két főkomponens, PC1 és PC2, a teljes variancia 90,0%-át magyarázta.

A kifejlett dohánytripszek morfológiai adataira kapott két főkomponens, PC1 és PC2, a teljes variancia (76,4 + 13,6 =) 90,0%-át magyarázta. Egyrészt jól kivehető az első főkomponens mentén a nőstények (piros, lila, rózsaszín) morfológiai elkülönülése a hímektől Másrészt a nőstények közül is kiválik a T változat a második főkomponens mentén. Ugyanígy a hímek között is elkülönülnek az L1 változatok (világoskék és sötétkék) a T változatoktól (sárga és barna) a második főkomponens mentén. Az egyes főkomponensek loadingjait elemezve, valamint a biplot segítségével a változatok morfológiai elkülönülése részleteiben is magyarázható.

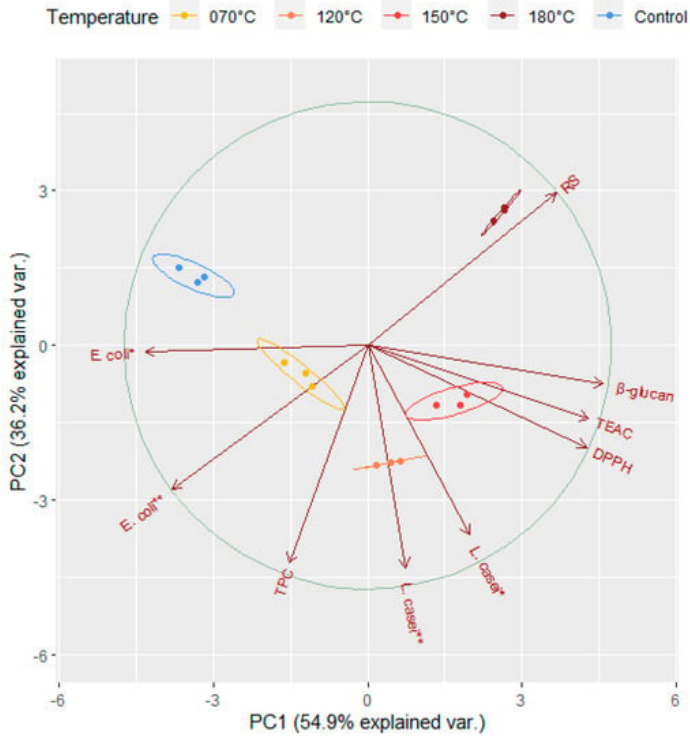


18. Ábra A különböző dohánytripsz (*T. tabaci*) -változatok kifejlett egyedeire vonatkozó morfometriai változók főkomponens-elemzésének eredményeképpen kapott biplotja. A barna nyilak a mért változókat jelölik, a tengelyekre vetített hosszuk a megfelelő főkomponensek loadingjainak felelnek meg. A különböző színű ellipszisek pedig a hím (m) és nőstény (f), szűz (V) és párosodott (M), L1, L2 és T változatokat ábrázoló pontok 95%-os konfidencia-intervallumai.

Második példánkban Kiss és munkatársai (2015) kutatását idézzük, akik négy különböző hőmérsékleten (plusz kontroll) való kezelések hatását vizsgálták Reishi gyógygomba (*Ganoderma lingzhi*) szárított termőtestére α-glükán, β-glükán összes glükán; *E. coli*, *L. casei* (kezelés kezdetén és végén), antioxidáns-kapacitás (DPPH, TEAC), összes fenolok (TPC) és a redukáló cukrok mért értékei alapján.

A különböző hőmérsékleti kezeléseket tekintettük előzetes csoportbesorolásnak, és a fenti változókkal főkomponens-elemzést végeztem, majd biplotot készítettem (19. Ábra). Két főkomponens a teljes variancia 91,1%-át magyarázta (54,9% és 36,2%). A diagramon a nyilak a mért változókat, a különböző színű pontok pedig a különböző hőkezelést kapott mintaelemeket jelölik. A mintavételi pontokat és egy adott változó feléjük mutató nyilát együtt úgy kell értelmezni, hogy az a változó nagymértékben jellemzi azokat a mintaelemeket. A nyilaknak a tengelyre vetített nagyságai az egyes PC-khez való hozzájárulásuk erősségét jellemzik (azok a megfelelő PC-k loadingjai). A hasonló vagy ellentétes irányba mutató nyilak pozitív vagy negatív korrelációjú változókat jeleznek, míg a közel egymásra merőlegesek alacsony vagy semmilyen korrelációra utalnak. A színes ellipszisek 99%-os konfidenciaszintre vonatkoznak a mintavételi pontok körül.

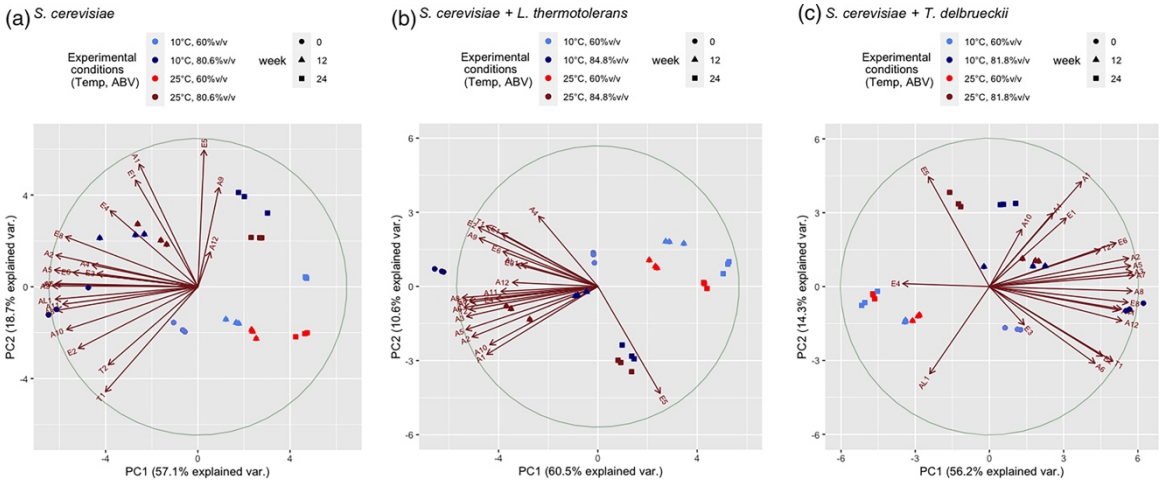
Az ábra az első főkomponens mentén egyértelműen mutatja a hőkezelés hőmérsékletének emelésének hatását, míg a kontroll jól elkülönül a második főkomponens mentén. Az antioxidáns-kapacitás és a β-glükán 150 °C-on, a redukáló cukrok 180 °C-on voltak a legmagasabbak, míg az *L. casei* és a fenolok 120 °C-on vagy e hőmérséklet közelében voltak a legmagasabbak. Az *E. coli* 70 °C-on volt a legmagasabb. A tanulmányban kapott eredmények a rövid távú hőkezelés jelentős hatását mutatták ki a Reishi gomba szárított termőtestére azzal, hogy fokozta az antioxidáns kapacitást, a β-glükán oldhatóságot és a prebiotikus tulajdonságot.



19. Ábra A Reishi gombák szárított termőtestén mért változók főkomponens-elemzésének biplotja. A barna nyilak a mért változókat jelölik, a különböző színű pontok a különböző hőfokon végzett kezeléshez tartozó mintaelemeket jelölik. A színes ellipszisek a mintavételi pontok körül 99%-os konfidenciaszintre vonatkoznak. *: kezelés kezdetén **: kezelés végén

Harmadik biplot példánk Fejzullahu és munkatársai (2024) kutatásáról szól, amelyben alkoholos almapárlatok 27 összetevő alapján leírt aromaprofilját vizsgálták *Lachancea thermotolerans* (Lt) és *Torulaspora delbrueckii* (Td) élesztőknek az *S.*

cerevisiai (Sc) mellett az erjesztésbe való bevonásával. Főkomponens-elemzést végeztem az erjesztés idejének, az alkalmazott hőmérsékletnek és az alkoholtartalomnak mint csoportosító változóknak összevont szintjeivel a három vizsgált esetre: erjesztés Sc Lt-vel, Sc Td-vel és Sc-vel önmagában, és az eredményeket biploton szemléltettem (20. Ábra).



20. Ábra A főkomponens-elemzés biplotjai a) *S. cerevisiae*, b) *S. cerevisiae* + *L. thermotolerans* és c) *S. cerevisiae* + *T. delbrueckii* által nyert almapárlatok megfigyelt változóira. A barna nyilak a mért változókat (illékony vegyületek) jelölik, tengelyekre vetített hosszuk a megfelelő főkomponens loadingjainak felelnek meg. A különböző alakú és színű pontok pedig az érlelési időhöz (0, 12, 24 hét), a hőmérséklethez (10 ºC és 25 ºC) és az alkoholtartalomhoz (60% és >80%) tartozó almapárlat-mintáknak felelnek meg.

Azok a mintavételi pontok, amelyek felé a nyilak mutatnak, a nyílnak megfelelő összetevőből nagy mennyiségben tartalmaznak, míg azokra a mintaelemekre, amelyek ellenkező irányban helyezkednek el, ezek az összetevők nem jellemzőek. A közel azonos irányba mutató nyilak pozitív korrelációban, az ellentétes irányba mutatóak negatív korrelációban álló összetevőkre vonatkoznak, a merőlegesek egymással nem korreláló összetevőket jelenítenek meg.

A főkomponens-elemzés és a biplot segítségével a háromféle módon erjesztett párlatok aromaprofiljai összehasonlíthatóvá váltak.

Főkomponens-elemzés alkalmazására 6.3. és a 6.4. fejezetben mutatunk példát.

Főkomponens-elemzést R környezetben több csomag, illetve függvény is kínál, ilyenek a `'stats'` csomag `'prcomp()'` vagy `'princomp()'` függvényei, a `'FactoMineR'` (Sebastien és mtsai, 2008) csomag `'PCA()'` függvénye, az `'ade4'` csomag (Dray és Dufour, 2007, Bougeard és Dray, 2018, Chessel és mtsai, 2004, Dray és mtsai, 2007, Thioulouse és mtsai, 2018) `'dudi.pca()'` függvénye, a `'psych'` csomag `'principal()'` függvénye, valamint az `'ExPosition'` (Beaton és mtsai, 2014) csomag `'epPCA()'` függvénye. Vizualizációhoz a `'factoextra'` (Kassambra és Mundt, 2020) csomag számos kiváló lehetőséget nyújt.

5.2.3 A főkomponens-regresszió és a parciális legkisebb négyzetek módszere

A főkomponens-elemzés kiválóan alkalmazható olyan módszerek előkészítésére, amelyek megkövetelik a változók függetlenségét. Amint láttuk, ilyen például a többszörös lineáris regresszió (ld. 5.1.5. fejezet). Amennyiben a magyarázó változók között korreláció áll fenn, a problémát multikollinearitásnak nevezzük.

Multikollinearitás esetén igen jól alkalmazható a többszörös regresszióanalízis előtt elvégzett dimenziócsökkentő főkomponens-elemzés (PCA, *preprocessing,* ld. 5.2.2. fejezet), hiszen a módszer eredményéül kapott főkomponensek az eredeti változók varianciájának nagy hányadát megtartják úgy, hogy azok már egymással korrelálatlanok (Boehmke és Greenwell, 2020). A főkomponens-szkórokat mint magyarázó változókat használhatjuk ezután a többszörös lineáris regresszióhoz. Ezt a módszert főkomponens-regressziónak (*Principal Component Regression*, PCR) nevezzük. Ennek tehát első lépése felügyelés nélküli, második lépése felügyelt módszer. Előfordulhat azonban, hogy a PCA-val előállított főkomponensek nem jól magyarázzák a regresszióanalízis válaszváltozóját. További hátránya e módszernek, hogy esetenként nehéz az interpretációja, valamint, hogy nagyon érzékeny a zavarásra (*confounding*).

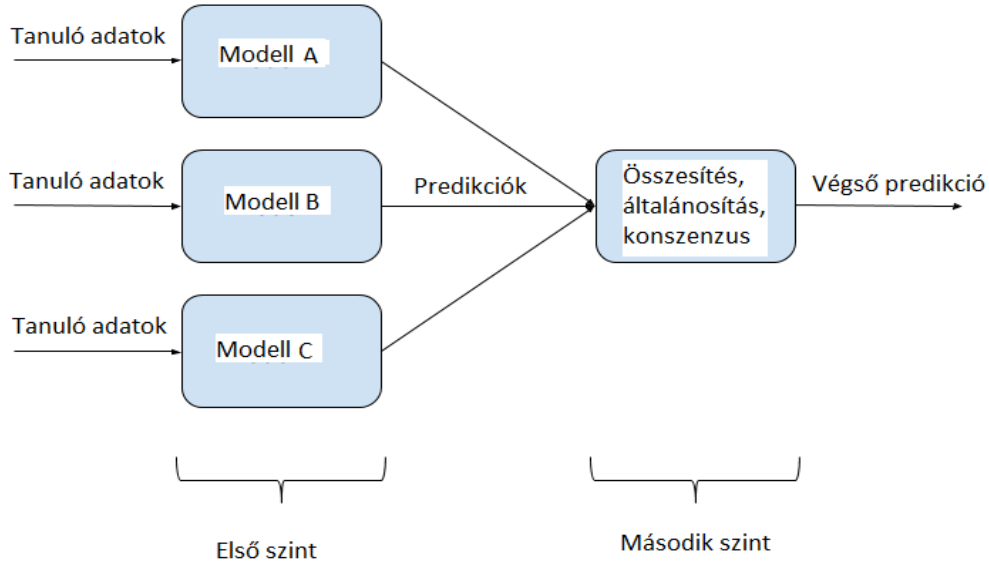
A parciális legkisebb négyzetek módszere (*Partial Least Squares Regression*, PLSR) a főkomponens-regressziónak a felügyelt változata, mégis e fejezetben tárgyaljuk a módszerek közti összefüggések könnyebb áttekinthetősége céljából. E módszer a PCR-nél hatásosabb lehet a többszörös regresszió alkalmazásánál, hiszen már a dimenziócsökkentés során használhatjuk a regresszióanalízis felügyelt tulajdonságát, azaz a válaszváltozó által hordozott információt (Tobias, 1996, Hastie és mtsai, 2013, Boehmke és Greenwell, 2020, James és mtsai, 2013). A módszer ugyanis a PCR-rel ellentétben a főkomponenseket nem a változók varianciáinak maximalizálásával, hanem a válaszváltozóval való korreláció maximalizálásával hozza létre. Így, ha ezután az így kapott főkomponensekkel regresszióanalízist végzünk, azok esetenként (de nem minden esetben) sokkal jobban tudják magyarázni a válaszváltozót, mint ha a főkomponenseket a válaszváltozók figyelembevétele nélkül állítottuk volna elő (PCR). Ez nagy előnye a PLSR-nek a PCR-rel szemben. A PLSR segítségével a legfontosabb ismérvváltozók is kiválaszthatóak (*Variable Importance,* Boehmke és Greenwell, 2020). Van azonban a PLSR-módszernek a PCR-rel szemben hátránya is, mégpedig az, hogy túlillesztésre hajlamos, míg ez a PCR-ről nem mondható el.

Főkomponens-regresszió alkalmazására a 6.3. fejezetben láthatunk egy példát.

Főkomponens-regressziót R környezetben a `'pls'` (Liland és mtsai, 2023) csomag `'pcr()'` függvényével végeztem, míg PLSR modellt ugyanebben a csomagban a `'plsr()'` függvénnyel építettem.

5.3 Metatanuló módszerek (e*nsemble methods, meta-learning methods*)

A metatanuló módszerek ugyanazon az adathalmazon futtatott számos modell együttes kiértékeléséből egy konszenzusmegoldást dolgoznak ki azzal a céllal, hogy a modell pontosságát növeljék, illetve az előrejelzés hibáit csökkentsék (21. Ábra, Hastie és mtsai, 2013).

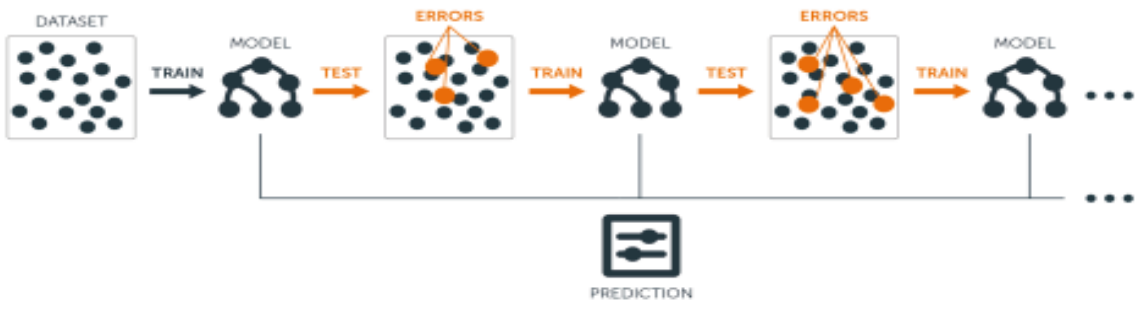


21. Ábra A metatanuló módszerek sematikus ábrája (Forrás: saját szerkesztés)

A legegyszerűbb metatanuló módszer az ún. *modellátlagolás*, melynek során az ugyanazon az adathalmazon futtatott számos modell eredményeinek egyszerű átlagolásával készítünk előrejelzést. Így például osztályozási feladat esetében a modellek szavazatait összesítjük, regressziós (becslő) feladatok esetében átlagot számolunk.

A *zsákolás* (*Bootstrap-Aggregating; Bagging*) egy egyszerű 'minta-a-mintából'-módszer (Hastie és mtsai, 2013). Lényege, hogy a meglévő adathalmazból véletlenszerűen választva nagyszámú új adathalmazt képezünk (*bootstrapping*, ld. 6.5. fejezet), és ezekre mint tanuló adathalmazokra külön-külön futtatjuk a modellünket. Végül a több véletlen adatrészhalmazon tanult modelleredményeket összesítjük (*aggregating*) valamely irányelv szerint (Breiman, 1996a Kuhn és Johnson, 2013, Ghatak, 2017, Boehmke és Greenwell, 2020, James és mtsai, 2013).

A lépésenként javítva tanuló (*Gradient Boosting Method,* GBM) módszer lényege, hogy ún. gyengén tanuló modellekből ún. erősen tanuló modelleket fejlesszen. A módszer a tanuló adathalmazt módosítja lépésről lépésre az előző lépés kimenetének függvényében (22. Ábra). A hibásan besorolt, illetve pontatlanul becsült mintaelemek nagyobb súlyt kapnak az új tanuló adathalmazban, mivel ezek hívják fel a figyelmet a modell pontatlanságára (Kuhn és Johnson, 2013, Hastie és mtsai, 2013, Géron, 2017, Ghatak, 2017, Boehmke és Greenwell, 2020, James és mtsai, 2013).



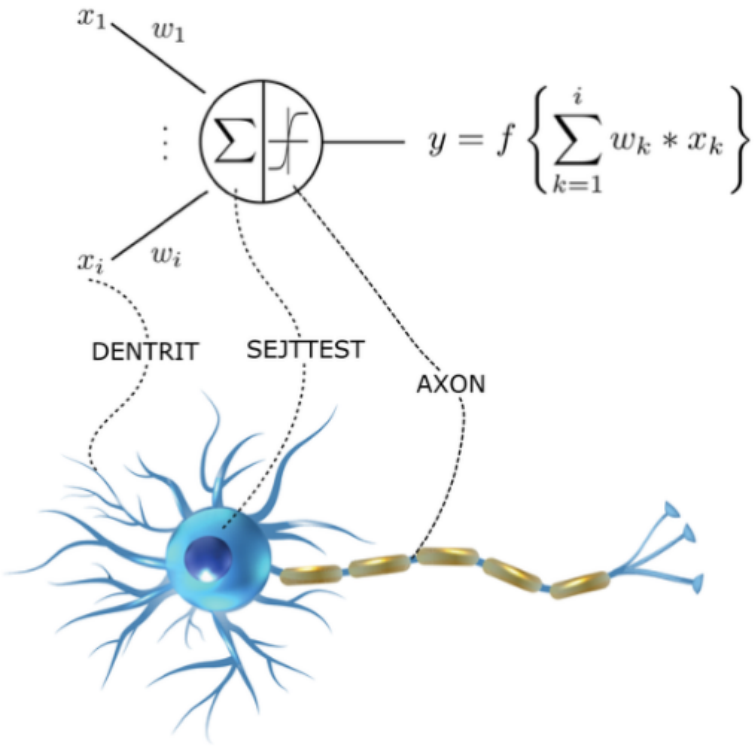
22. Ábra A lépésenként javítva tanuló (*Gradient Boosting Method,* GBM) módszer sematikus ábrája (Forrás: Boehmke és Greenwell, 2020)

A zsákolást és a lépésenkénti javítást gyakran együtt használják (*Bagging and Boosting*, BB). A zsákolás a felülillesztési problémára, a lépésenként javítva tanulás a torzításra (*biasing*) lehet megoldás (Lantz, 2015, Hastie és mtsai, 2013).

A zsákolást és a lépésenként javítva tanuló módszerhez R környezetben a `'gbm'` (Ridgeway és Developers, 2024), a `'h2o'` (Fryda és mtsai, 2024), valamint az `'xgboost'` (Ghatak, 2017, Chen és mtsai, 2024) csomagokat használtam.

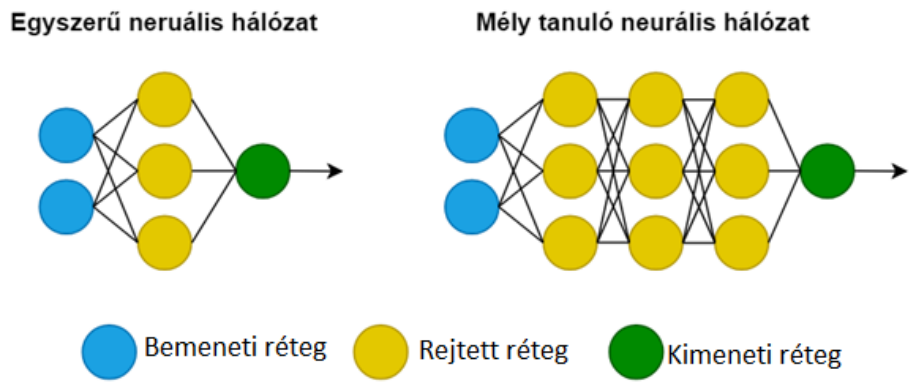
5.3.1 Neurális hálózatok

A neurális hálózatok (*Artificial Neural Networks,* ANN) az idegsejtek működésének modelljén alapuló eljárás. Az idegsejt a dendriteken keresztül fogadja az ingereket (üzeneteket, információt), ezeket súlyozva összegzi a tartalmukat, s ennek megfelelően alakítja ki az ingerekre megfelelő válaszát (23. Ábra). A neurális hálózatok e biológiai modell szerint fogadják a változók értékei formájában az információt, ezeket a változókat valamely súlyozással összegzik, előállítanak egy aktiváló függvényt, és végezetül ennek eredményét adják ki eredményül (Fausett, 1994, Bishop, 1995., Kuhn és Johnson, 2013, Lantz, 2015).



23. Ábra Az idegsejtek működése és ennek modelljére épülő neurális hálózati struktúra. (Forrás: Tóth Bálint Pál, Élet és Tudomány, 2016/09/15)

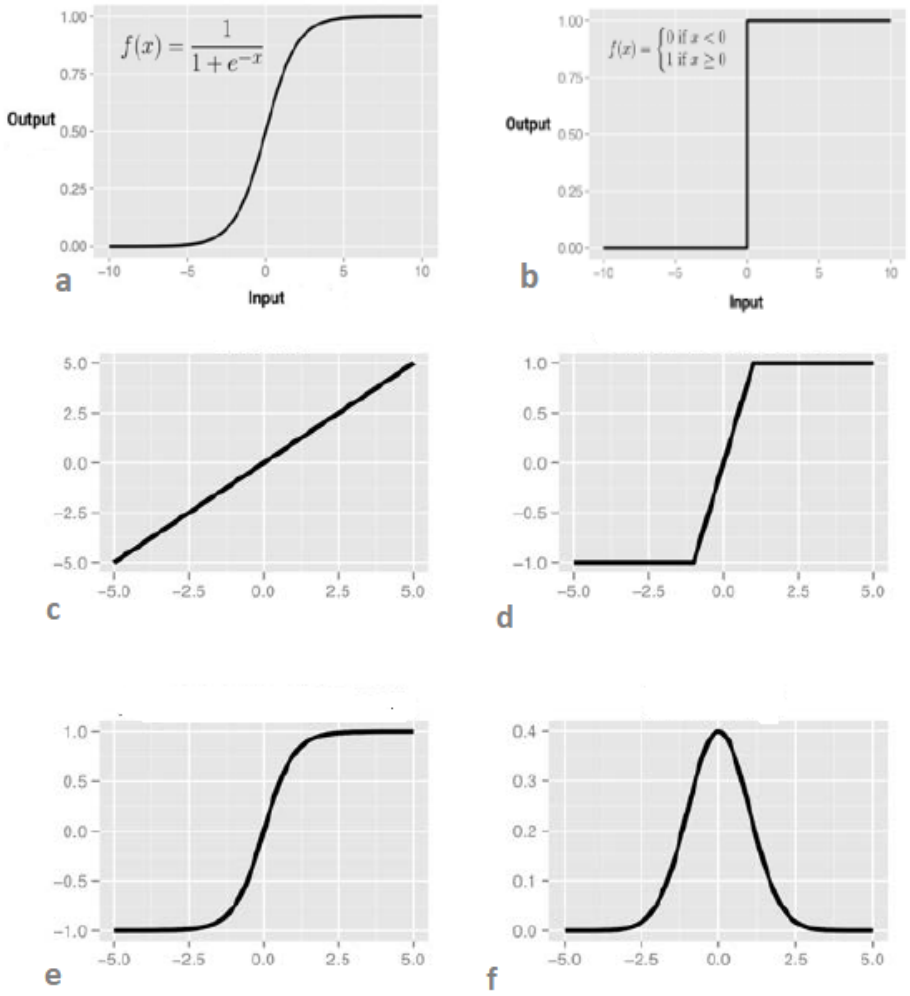
A modell bemeneti rétegből (*input*), rejtett rétegből (*hidden layer(s)*) és kimeneti rétegből (*output*) áll (24. Ábra). A rétegeket összekötő szakaszokon különféle súlyok szerepelnek. Amennyiben több rejtett rétegből áll a hálózat, mélytanuló hálózatról beszélünk (*deep learning, deep neural networks,* DNN, Hastie és mtsai, 2013, Chollet és Allaire, 2018, Nisbet és mtsai, 2018, Boehmke és Greenwell, 2020).



24. Ábra A neurális hálózatok sematikus szerkezete (forrás: Kovács Róbert, mesterin.hu 2020. 09. 22.)

A neurális hálózatot az aktiváló függvény, a hálózati topológia, valamint a tanuló adathalmazt létrehozó algoritmusa definiálja (Boehmke és Greenwell, 2020).

Az *aktiváló függvény* leggyakoribb fajtája az ún. szigmoid (25. Ábra a.), de ezen kívül több más típusú függvényt is alkalmazhatunk, (25. Ábra b, c, d, e, f).



25. Ábra A neurális hálózatok leggyakrabban használt aktiváló függvényei. a. szigmoid; b. egylépcsős; c. lineáris; d. monoton szakaszfüggvény; e. tangens hiperbolikus; f. Gauss (forrás: Lantz, 2015)

A neurális hálózatok az aktiváló függvény választásától függően nagyon különbözőek lehetnek: lineáris esetben a modell nagyon hasonlít egy egyszerű lineáris regresszióra, a Gauss-függvényt választva az ún. sugárirányú függvénymodellt kapjuk (*radial basis function*, RBF). Amint látjuk, ezeknek a függvényeknek a lényeges része az értelmezési tartomány [-5; +5] részhalmazán belül van, ezért az adatokat előkészítésül erre a tartományra vetítjük normálással vagy standardizálással.

A *hálózati topológiát* meghatározó legfontosabb három tényező: a rétegek száma, az, hogy engedjük-e, hogy az információ visszafelé is áramolhasson, illetve a csomópontok száma egy-egy rétegben (Lantz, 2015).

Az egyszerű (egyrétegű) neurális hálózatok egyszerű (lényegében lineáris) kapcsolatokat tudnak modellezni. Ennél összetettebb, illetve nemlineáris összefüggések modellezéséhez több rétegre van szükség. A rétegek számának növelésével azonban az interpretálás is nehezebbé válik, ezen kívül a műveletigény is nagyon megnőhet, ahogy azt például mágneses magrezonancia (NMR) módszerek alkalmazásánál több helyen is problémaként említik (Amargianitaki és Spyros, 2017, Masetti és mtsai, 2021, Kalogiouri és Samanido, 2021, ld. 6.6. fejezet).

Attól függően, hogy az információ kizárólag előrefelé (az inputtól az output irányába), vagy visszafelé is áramolhat, megkülönböztetünk előremutató (*feedforward*) vagy visszacsatoló (*feedback*) neurális hálózatot (Boehmke és Greenwell, 2020). Annak ellenére, hogy az előbbi egy erős megszorítás, ezek a modellek igen jól működnek. A visszacsatoló modellek igen komplex kapcsolatokat képesek modellezni, és ezért késleltetett válaszreakciókat is tudnak kezelni, műveletigényük viszont sokkal magasabb.

Természetesen a csomópontok száma a bemeneti és kimeneti rétegekben az ismérvváltozók számától, illetve a lehetséges címkék számától függ. A rejtett rétegekben azonban a csomópontok számát a modellezőnek kell megadnia. Sajnos megbízható, mindig alkalmazható ökölszabály erre nem létezik, mivel az optimális szám többek között függ a bemeneti réteg csomópontjainak számától, a tanuló adathalmaz méretétől, az adatok szennyezettségétől, valamint a tanulási feladattól. A csomópontok számát növelve a túlillesztettség, illetve a túl nagy műveletigény problémája léphet fel.

A neurális hálózatok elméletében óriási áttörést jelentett Rumelhart és munkatársainak (1986) a tanuló adathalmaz létrehozásának algoritmusáról szóló munkája, az ún. hiba-visszaterjesztés módszere (*backpropagation method,* Lantz, 2015). Ezzel a módszerrel a gép tanulását oly módon sikerült felgyorsítani, hogy az lehetővé tette a mélytanulást (*deep learning*), azaz a többrétegű neurális hálózatok építését, amit addig a nagy műveletigénye miatt a gyakorlatban nemigen lehetett alkalmazni. A módszer iteratív: a nulladik lépésben a bemenő változók súlyozása véletlen. Ezután minden egyes lépés két részből áll: az elsőben az adott súlyokkal történő összegzést követően az aktiváló függvény egy kimeneti eredményt bocsát ki. A második lépésben ezt a kimeneti eredményt a címkeváltozó értékeivel összehasonlítva a bemenő adatok súlyait a hiba csökkentése érdekében módosítjuk. Ezt addig folytatjuk, amíg a folyamatot egy, a hiba nagyságrendjére vonatkozó kritérium alapján leállítjuk.

Kérdés, hogy mi alapján lehet a súlyokat a hiba csökkentése érdekében módosítani. Erre szolgál az ún. konjugált gradiens módszer (*gradient descent method,* Ghatak, 2017, Boehmke és Greenwell, 2020), melynek lényege, hogy az aktiváló függvény deriválásával meghatározzuk a legnagyobb meredekség irányát és nagyságát, tehát azt a súlyvektort, amellyel a hiba legjobban csökkenthető.

A neurális hálózatok előnyösen alkalmazhatóak nagy bonyolultságú feladatokra (ilyenek a nagy nyelvi modellek, *Large Language Models,* LLM), hiszen rendkívül komplex összefüggéseket is jól tudnak kezelni viszonylag kevés feltétellel (Lantz, 2015).

Ám igen nagy műveletigényűek, és bár nagyon nagy pontosságra képesek, az eredmények interpretálása nehézkes (fekete doboz típusúak (*black box*), azaz jól működnek, de nem tudjuk, valójában miért).

Neurális hálózatok alkalmazására a 6.6. fejezetben mutatunk példát.

Neurális hálózatokat R környezetben a `'neuralnet'` (Fritsch és mtsai, 2019) csomag alkalmazásával építettem a `'neuralnet()'` függvény segítségével, valamint a `'keras'` (Allaire and Chollet, 2019) csomagbeli `'keras_model_sequential()'` függvénnyel.

5.3.2 Véletlen erdő

A véletlenerdő-módszer (*Random Forest*, RF) döntési fák sokaságára épül (Breiman, 2001, Liaw és Wiener, 2002, Lantz, 2015, Nisbet és mtsai, 2018, Boehmke és Greenwell, 2020, Kuhn és Johnson, 2013, James és mtsai, 2013). Az eredeti adathalmazból véletlen mintavétellel a mintaelemek halmazából részhalmazokat képezünk (*bootstrapping,* ld. 5.3. fejezet). A modellezés folyamán azonban nemcsak a mintaelemek halmazából, hanem az ismérvváltozók halmazából is véletlenszerűen választunk részhalmazokat (*splitting*). Az így kiválasztott adathalmazra döntési fa modellt építünk. A véletlen erdő fontos paramétere az a szám, hogy a k darab ismérvváltozóból hányat választunk be egy fa építésére egy csomópontban (*mtry*). Leggyakrabban ezt hozzávetőlegesen $k/3$-nak, vagy $\sqrt{k}$-nak választjuk. Egy másik fontos paraméter, hogy hány fát szeretnénk építeni (*ntree*). Természetesen, mivel a módszer egyenként döntési vagy regressziós fákat épít (ld. 5.1.3. fejezet), így örökli az ott ismertetett paramétereket is: egy-egy fa maximális mélységét (*tree depth*), illetve egy-egy fa minimális ágméretét (*nodesize*), mely kettő nyilván szorosan összefügg egymással (Ghatak, 2017).

Az elkészített nagyszámú fából konszenzus fa készül, mégpedig osztályozás esetén szavazással, regresszió esetén átlagolással. Ezen kívül más konszenzuskritériumok is megfogalmazhatóak, például kiegyensúlyozatlan elemszámú csoportok esetén figyelembe vehetjük a csoportba esés valószínűségét is.

A véletlen erdő tehát több egymással összefüggő paraméterrel működik, melyek értékeire a módszer érzékeny. Ezért ajánlatos a modellt hangolni (*tuning,* Kuhn és Johnson, 2013, Boehmke és Greenwell, 2020). Ez a gyakorlatban azt jelenti, hogy az egyes paraméterek értékeire egy-egy vektort definiálunk, és az ún. rácskeresés (*grid search,* ld. 6.6. fejezet, illetve 17. Példa) segítségével a paraméterkombinációk rácspontjaira elvégezzük a modellfejlesztést, és végül az eredmény alapján kiválasztjuk az optimális paraméterkombinációt.

Az eredmény értékelését a véletlen erdőre vonatkozó mutatók segítségével végezhetjük.

A választott halmazon kívüli hiba (*Out-Of-Bag Error*, OOB hiba, ld. 2.2. fejezet, Breiman, 1996b) a modell hibájának mérésére szolgál: azoknak a mintaelemeknek a modellbecslését hasonlítja össze a válaszváltozóik értékével, melyek nem voltak benne a kiválasztott adathalmazban (Breiman, 1996b).

Az OOB hiba segítségével képezhető a változók fontossági mérőszáma (*Mean Decrease*, MDA, Breiman, 2001), mégpedig úgy, hogy egy változóra kiszámoljuk, mennyivel nő az előrejelzés hibája, ha az OOB adatok minden változóját változatlanul hagyjuk, de a kérdéses változó értékeit véletlen permutáljuk. Ezt minden egyes fa építésénél kiszámoljuk minden egyes változóra (ld. 6.2. fejezet).

A távolságmérték (*Proximity Measure*, Breiman, 2001) két mintaelem távolságát mutatja meg egy arány segítségével: a két elem az épített fák hányadában esett arra a döntési levélre. Azt sejtjük, hogy hasonló elemekre ez a szám magas lesz (közel 1-hez).

Véletlen erdő alkalmazására 6.2., valamint a 6.6. fejezetben mutatok példát.

A véletlenerdő-módszer előnye, hogy osztályozásra és regresszióra is alkalmas, szennyezett és hiányos adathalmazokra is jól működik, nem érzékeny a kiugró adatokra, és segítségével a legfontosabb ismérvváltozók is kiválaszthatóak (*Variable Importance,* Kuhn és Johnson, 2013, Boehmke és Greenwell, 2020), de hátránya, hogy komplexitása miatt az interpretációja nehézkes (*black box*), valamint nagy figyelmet és sok időt igényel a modell hangolása (Lantz, 2015).

A véletlenerdő-módszert R környezetben a `'randomForest'` (Liaw és Wiener, 2002, Breiman és mtsai, 2018) csomag alkalmazásával használtam a `'randomForest()'` függvény, valamint a `'ranger'` (Wright és Ziegler, 2017) csomag `'ranger()'` valamint a `'h2o'` csomag (Fryda és mtsai, 2024) `'h2o.randomForest()'` függvényének segítségével.

6 EREDMÉNYEK

Ebben a fejezetben nyolc publikációból származó elemzést mutatok be. Különböző szakterületek szakértő kollégáival való közös publikációkról van szó, melyekből az elemzési munkát emelem ki, a társszerző kollégák szakterületéhez kapcsolódó módszereket és megállapításokat a háttérben hagyva. Az eredményeket, illetve azok egy részét olyan formában közlöm, hogy azokból mi az, ami a felhasznált elemző-modellező módszer – egyes esetekben a szokásostól eltérő módon való – használatának köszönhetően e munka szerzőjének hozzájárulásával kerülhetett napvilágra. Ezen elemzések leírásánál tehát az adatok feldolgozására, illetve az elemzés módszertanára helyezem a hangsúlyt, és nem minden, a publikációban közölt módszerre, illetve eredményre térek ki. A kíváncsi olvasó a hiányzó részleteket a publikációk felkeresésével érheti el.

6.1 Az aszály hatása a talajban élő mikroízeltlábúakra

E fejezet az alábbi publikáción alapszik:

Flórián N., **Ladányi M.**, Ittzés A., Kröel-Dulay G., Ónodi G., Mucsi M., Szili-Kovács T., Gergócs V., Dányi L., Dombos, M. 2019. 'Effects of single and repeated drought on soil microarthropods in a semi-arid ecosystem depend more on timing and duration than drought severity'. PLoS ONE 14(7): e0219975. https://doi.org/10.1371/journal.pone.0219975 **D1**

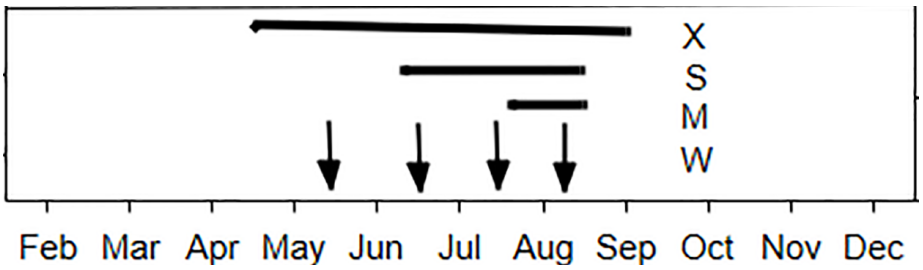
6.1.1 Motiváció és célkitűzés

A talajnedvesség az egyik legfontosabb tényező, amely befolyásolja a talaj élővilágát. A száraz és félszáraz ökoszisztémákban a talaj mezofaunája alkalmazkodott az átmeneti aszályos eseményekhez, de eddig csak korlátozottan ismertük a jövőbeli éghajlatváltozási forgatókönyvek szerint előre jelzett különböző nagyságú és gyakoriságú aszályok hatásait. Tanulmányunkban arra kerestük a választ, hogy a rugófarkúak és az atkák hogyan reagálnak a különböző nagyságú, és ismétlődő szimulált aszályos eseményekre egy magyar félszáraz homoki sztyeppén végzett terepkísérletben.

6.1.2 A kísérlet beállítása

A talajban élő ízeltlábúak aktivitásában bekövetkező változásokat talajcsapdázással követték nyomon két éven keresztül egy magasság, kitettség és a domináns növényfajok tekintetében térben változékony homokos talajon. Az első évben (2014) extrém aszályos előkezelést (teljes letakarás), az ezt követő évben pedig kevésbé pusztító kezeléseket (súlyos aszály, mérsékelt aszály, vízpótlás) alkalmaztak. Ezeken a helyszíneken hat, belsőleg homogénnek tekinthető blokkon egyenként nyolc darab parcella volt. Az első évben (2014) a blokkok felében szélsőséges aszályos előkezelést alkalmaztak (két szint: szélsőséges aszály (X) és kontroll (C)). Az ezt követő évben (2015) az előzetesen X- és C-kezelt parcellákon a talajba jutó csapadékmennyiség négy szintjét szimulálták: súlyos aszály (S), mérsékelt aszály (M), kontroll (C, környezeti csapadék) és víz hozzáadása (W) a 26. ábra szerint. Ily módon a két tényezőt (azaz a szélsőséges szárazságot és az ezt követő különböző csapadékeloszlást) egy teljes faktoriális tervben kombinálták, így nyolc kezeléskombinációt (CC, CS, CM, CW, XC, XS, XM, XW) kaptak hat ismétléssel, összesen 48 vizsgálati parcellán. 2015-ben a csapdákat áprilistól novemberig minden hónapban kiürítették a kezeléseknek megfelelő tevékenységek ütemezését követve, ezért a

talajban élő mikroízeltlábúak ún. *aktivitássűrűségi értékei* (AD, azaz az egy csapdában egy csapdázási időszakban talált egyedek száma) az egyhónapos gyűjtésekre vonatkoztak.



26. Ábra A kezelések időzítése és időtartama. A vízszintes sávok jelzik a különböző aszályos kezelések (a parcella letakarásának) időszakát: az 1. kezelés esetében: X: szélsőséges aszály 2014-ben; a 2. kezelés esetében: S: súlyos aszály, M: mérsékelt aszály 2015-ben, függőleges nyilak jelzik a vízpótlási eseményeket.

6.1.3 A kutatás hipotézise

A kutatás hipotézise az volt, hogy a különböző szintű aszályt szimuláló kezelések (extrém (5 hónap), súlyos (2 hónap), mérsékelt (1 hónap)) negatívan befolyásolják a talajban élő mikroízeltlábúak aktivitássűrűségét (AD), míg az extra csapadékkezelésre a talajfauna pozitív reakcióját feltételezték.

6.1.4 Módszer

Az egyes parcellák adatai igen nagy heterogenitást mutattak még az igen gondos adattisztítást követően is. Ennek oka részben abban áll, hogy az ugróvillás fajok az év során változó egyedsűrűséggel jelennek meg, másrészt abban, hogy egyes mezofauna fajok a talajban jellemzően csoportosan jelennek meg. Csupán e két ok is elegendő ahhoz, hogy megértsük, hogy az AD adataink a kezelésektől függetlenül is nagyságrendben térhettek el egymástól, ezért ezek óvatlan összehasonlításával igen torzító eredményeket kaphattunk volna.

E probléma megoldására az elemzéshez kétféle megközelítést használtunk: (1) normálás (relatív aktivitássűrűség, RAD) és (2) ordinális skálázás (aktivitássűrűség-különbség ADD) alkalmazása.

- A relatív aktivitássűrűség elemzése

A különböző nagyságrendű 2015-ös adatokat a relatív aktivitássűrűség (RAD) képzésével tettük összehasonlíthatóvá. Egy adott hónapra és egy adott parcellára vonatkozó (a logaritmusfüggvénnyel transzformált) aktivitássűrűséget elosztottuk az adott kezelésnek az év során észlelt összes (transzformált) aktivitássűrűségével, minden egyes mezofaunális csoportra külön-külön:

$$RAD_{tmb} = \frac{\ln\left(AD_{tmb} + 1\right)}{\sum_{m=4}^{11} \overline{\ln\left(AD_{tm} + 1\right)}}$$

ahol a t a kezelési tényezőkombinációkat (CC, XC, CW, XW, CM, XM, CS, XS), m a hónapokban kifejezett időt ($m = 4{,}5{,}\ldots{,}11$), b a blokkokat jelöli ($b = 1{,}2{,}\ldots{,}6$), $\overline{\ln\left(AD_{tm} + 1\right)}$ az m hónapban a t kezeléshez tartozó összes b blokkra számított egy egységnyivel eltolt és logaritmus-transzformált aktivitássűrűség átlaga.

A 2014-ben kapott adatokat ln(x+1) transzformációval korrigáltuk, ezután a szélsőséges szárazsággal kezelt (X) és a kontroll (C) kezelések hatásának összehasonlítását Student-féle *t*-próbával végeztük minden egyes mezofaunális csoportra külön-külön a Bonferroni-féle elsőfajú hiba korrekciójával.

A relatív aktivitássűrűség összehasonlítására a 2015-ös adatok esetében MANOVA modellt (ld. 3.1. fejezet) használtunk dinamikusan változó faktorokkal a kezelések kezdetének figyelembevételével minden egyes mezofaunális csoportra külön-külön. A 2014-es adatok esetében, mint láttuk, egyetlen faktorunk volt (F1), két szinttel: kezelt (X) és kontroll (C). A 2015-ös áprilisi és májusi adatokat még ugyanezzel a faktorral elemeztük. A júniusi adatok esetében már két faktorunk volt, a második faktor (F2) szintén kétszintű volt, kontroll és öntözött (W). Júliusban belépett a második faktor esetében a súlyos aszálykezelés szintje (S), augusztustól pedig az enyhe aszálykezelés szintje (M).

Szignifikáns MANOVA eredmény esetén az F1 és F2 hatásainak meghatározására egytényezős ANOVA-t alkalmaztunk (ld. 3.1. fejezet) az elsőfajú hiba Bonferroni-féle korrekciójával minden hónapra vonatkozóan. A hibatagok normalitását a ferdeség és a csúcsosság, míg a varianciák homogenitását a maximális és minimális varianciák aránya alapján ellenőriztük.

- Az aktivitássűrűség-különbség (ADD) elemzése

A talaj mezofaunájához tartozó fajok, feltéve, hogy jelen vannak, általában képesek rövid időn belül rendkívül nagy populációméretet létrehozni. A két érték közötti releváns különbség irányának és nagyságának kifejezésére először az aktivitássűrűségi adatokat a 8. Táblázatban bemutatott aktivitássűrűségi kategóriákat meghatározó ordinális skálára alakítottuk át.

8. Táblázat Az aktivitássűrűség (AD) kategóriái a felszínen élő ugróvillás és atkafajokra vonatkozóan.

A felszíni ugróvillás és atka fajok aktivitássűrűsége	kategória	értelmezés
0–10	alacsony	A faj véletlenszerű megjelenést mutat.
11–100	közepes	A faj jelen van, de az aktivitássűrűsége nem magas.
>100	magas	A faj nagy egyedszámmal van jelen.

Az ordinális adatokból határoztuk meg a különbségek irányát és nagyságát kifejező faktort, az ún. aktivitássűrűség-különbségeket (ADD). E faktor értékeit a CW, CM, CS (2014-ben kontroll, 2015-ben kezelt) és CC (mindkét évben kontroll), valamint XW, XM, XS (mindkét évben kezelt) és XC (2014-ben kezelt, 2015-ben kontroll) faktorkombinációk aktivitássűrűség-értékeinek összehasonlításai alapján határoztuk meg a 9. Táblázat szerint. A táblázatban szereplő kritériumok a talajban élő ízeltlábúak szakértőinek és a statisztikai elemzők közös munkája. Az egyes értékek a következő releváns, illetve nem releváns aktivitássűrűség-különbségeket fejezik ki:

- A 8. Táblázat szerinti alacsony, közepes, illetve magas kategórián *belüli* aktivitássűrűség- különbségek **nem tekinthetők relevánsnak**, *kivéve*, ha a nagyobb aktivitássűrűség-érték legalább kétszerese a kisebbnek.
- A szomszédos kategóriák közötti aktivitássűrűség-értékek különbsége már **relevánsnak tekinthető**, *kivéve*, ha azok különbsége nem éri el a 10 (alacsony-közepes), illetve 50 (közepes-magas) értéket. Ezzel kizárjuk, hogy pl. a 10 (alacsony) és 15 (közepes), aktivitássűrűség-értékek közti (csupán 5) különbséget relevánsnak tekintsük.

- Az alacsony és magas kategóriába tartozó aktivitássűrűség-értékek különbsége **releváns**.

9. Táblázat Az aktivitássűrűség-különbségek irányát és nagyságrendjét kifejező aktivitássűrűség-különbség (ADD) kiszámításának módja a 2015. évi kezelési és kontrolltényező-pár értékek összehasonlítása alapján. A különbségek kritériumait a nyilak felett mutatjuk be. A különbség nagysága: „nincs releváns különbség” (0); “egy nagyságrendű különbség” (±1), “két nagyságrendű különbség” (±2). Az x a két összehasonlított aktivitássűrűség-érték közül az alacsonyabb értéket jelenti.

A különbség rendje	Az aktivitássűrűség kategóriája		
	alacsony	közepes	magas
Nem releváns különbség	0 ↔ abszolút különbség<10 ↔ 0	0 ↔ abszolút különbség<50 ↔ 0	0
Egy nagyságrendű (releváns) különbség	abszolút különbség>10 ±1	abszolút különbség>x*2 ±1	abszolút különbség>50 ±1
Két nagyságrendű (releváns) különbség		±2	

A 2015-ös egyes kezelések esetében a kezelések kezdete előtt és után (azaz az első vízpótlás (W) előtt és után, illetve a mérsékelt (M) és a súlyos (S) aszály kezdete előtt és után) képeztük azoknak az eseteknek a súlyozott arányát, amikor a különböző mezofaunális csoportok aktivitássűrűségkülönbség-értékei relevánsan alacsonyabbak voltak a kezelt parcellákon (a kontrollhoz képest). Ezeket az arányokat Z-tesztekkel és Fischer-féle egzakt tesztekkel hasonlítottuk össze, hogy kiderüljön, hogy a kezelés hatására releváns mértékben különbözött-e az aktivitássűrűség.

6.1.5 Eredmények

A szélsőséges aszályos kezelés évében (2014) a kezelt parcellákon az összes ugróvillások aktivitássűrűsége alacsonyabb volt, mint a kontroll parcellákon. A 2014-ben fogott, a talaj felszínén élő ugróvillás populációk összlétszáma 48%-ra csökkent, és a talajban élő ugróvillások aktivitássűrűsége is 90%-kal csökkent. A magas varianciák miatt szignifikáns csökkenés azonban csak a növényzetben élő ugróvillás populációk fogásában volt kimutatható, amelyek esetében a csökkenés 94%-os volt (Student-féle t-próba, $p < 0{,}001$). Ezzel szemben az összes atkacsoport (*Oribatida, Mesostigmata, Prostigmata, Astigmata*) aktivitássűrűsége szignifikánsan, 49%-kal ($p < 0{,}05$), 67%-kal, 70%-kal és 97%-kal (mindháromra $p < 0{,}001$) nőtt a szélsőséges szárazság hatására (Student-féle t-próba).

A szélsőséges aszályos kezeléssel ellentétben a 2015-ben alkalmazott kevésbé intenzív aszályos kezelések (súlyos vagy enyhe) nem mutattak ilyen egyértelmű csökkenést. A relatív aktivitássűrűségben (AD) nem tudtunk szignifikáns változást

kimutatni a különböző aszálykezelések között a kontrollkezeléshez képest egyik mikroízeltlábú esetében sem.

Második megközelítésünkben megvizsgáltuk a kezelések és a kontrollok közötti ordinálisan skálázott különbségeket (ADD), amelyeket a kezelés végrehajtása előtti és utáni hónapokban kaptunk.

A súlyos szárazság, a mérsékelt szárazság és a víz hozzáadása esetén sem találtunk szignifikáns különbséget a kontrollhoz képest az aktivitássűrűség-különbségek (ADD) tekintetében a mikroízeltlábúakra vonatkozóan, kivéve a talaj felszínén élő ugróvillások esetében. A 10. táblázatban azoknak az eseteknek az arányát mutatjuk be a talaj felszínén élő ugróvillásokra vonatkozóan, amikor az aktivitássűrűség-értékek relevánsan alacsonyabbak voltak a kontrollhoz képest, azaz amikor az aktivitássűrűségkülönbség-értékek negatívak voltak. A súlyos aszály negatív hatása a kontroll évet követően szignifikánsan nagyobb arányban jelentkezett, ami azt jelzi, hogy a populációknak nehézséget okozott az ilyen mértékű abiotikus stressz. A 2015-ben súlyos és korábban extrém aszállyal kezelt parcellákban (XS-XC) szintén gyakrabban fordultak elő negatív hatások, bár ezek a hatások statisztikailag nem voltak szignifikánsak, azaz az előző évben is súlyos aszálynak kitett parcellákon a következő évi súlyos aszály már nem tudott szignifikáns különbséget okozni.

10. Táblázat A 9. Táblázat alapján a 2015-ös adatokból előállított aktivitássűrűség-különbség (ADD) értékeinek súlyozott arányára vonatkozó összehasonlítások az egyes kezelések előtt és után a talaj felszínén élő ugróvillás fajok esetében a Z-próba értékeivel és szignifikanciáival, valamint a Fischer-féle egzakt teszt eredményével. A százalékos adatok az ordinális skálázásból származnak (8. Táblázat). Kezelési kódok: az első karakter a 2014. évi előkezelést jelzi (kontroll: C, extrém szárazságkezelés: X), a második karakter pedig a 2015-ös, egymást követő kezelést (kontroll: C, vízpótlás: W, mérsékelt szárazság: M és súlyos szárazság: S).

Az összehasonlított kezeléspárok	Azoknak az eseteknek az aránya, amikor az aktivitássűrűség-értékek relevánsan alacsonyabbak voltak a kontrollhoz képest		Egyoldali Z próbastatisztika abszolútértéke	Fischer-féle egzakt teszt *p* értéke
	Kezelés előtt	Kezelés után		
CM-CC	11% (18)	30% (10)	1,25 ns	0,32
XM-XC	6% (16)	14% (7)	0,68 ns	0,51
CS-CC	0% (16)	62% (13)	3,69 ***	<0,001
XS-XC	21% (14)	43% (14)	1,21 ns	0,42
CW-CC	42% (12)	21% (14)	1,11 ns	0,40
XW-XC	18% (11)	33% (15)	0,86 ns	0,66

***: $p < 0{,}001$; ns: nem szignifikáns

6.1.6 Következtetés

A talaj mezofauna csoportjai a közép-magyarországi száraz homoki sztyeppék változékony környezetében alkalmazkodtak az extrém körülményekhez. Bár a szélsőséges események megváltoztathatják aktivitássűrűségüket, úgy tűnik, rövid időn belül képesek megbirkózni a változásokkal. Az extrém szárazság nagyobb hatással volt a talaj mezofauna aktivitássűrűségére, mint a súlyos vagy mérsékelt szárazság. A víz hozzáadása és az aszályos kezelések esetében a kezelés időtartama és időzítése fontosabbnak tűnt a talaj mezofaunája szempontjából, mint annak súlyossága (azaz a talajnedvesség csökkenésének mértéke).

Arra következtettünk, hogy a száraz homoktalaj mezofaunája képes túlélni szélsőségesen száraz körülmények között, populációik gyorsan helyreállnak, de a nagyon hosszú aszályos időszakokkal nem biztos, hogy képesek megbirkózni.

6.2 A kímélő talajművelésnek a lefolyásra és a talajveszteségre gyakorolt hatása

E fejezet az alábbi publikáción alapszik:

Madarász B., Jakab G., Szalai Z., Juhos K., Kotroczó Zs., Tóth A., **Ladányi M.** 2021 Long-term effects of conservation tillage on soil erosion in Central Europe: A random forest-based approach, Soil and Tillage Research, 209(104959), ISSN 0167-1987, https://doi.org/10.1016/j.still.2021.104959. **D1**

6.2.1 Motiváció és célkitűzés

Magyarországon a termőföldek több mint egyharmada erodálódik az intenzív mezőgazdaság (PT) okozta talajdegradáció következtében. A kímélő talajművelés (CT) költséghatékony és víztakarékos gazdálkodási rendszer lehet a megoldás e problémára, mivel a talajerózió mérséklése mellett javítja az élelmezésbiztonságot és csökkenti az erőforrások degradációját még olyan időjárási anomáliák esetén is, mint az aszály és az intenzív esőzések. Azonban a kímélő talajművelésnek a lefolyásra (RO) és a talajveszteségre (SL) gyakorolt átfogó hatásai máig nem kielégítően ismertek. A kutatás célja a kímélő talajművelésnek a talaj a lefolyásra (RO) és a talajveszteségre (SL) gyakorolt hosszú távú hatásának feltárása volt.

6.2.2 A kísérlet beállítása

A kísérlet 2003-ban indult a SOWAP (*Soil and Surface Water Protection Using Conservation Tillage in Northern and Central Europe*) projekt részeként a szántásos (PT) és a talajkímélő (CT) szántóföldi művelés összehasonlító tanulmányozására. A PT és a CT eróziós kísérletéhez 2 × 2 azonos méretű, 24 × 50 m-es lejtős parcellát alakítottak ki. A parcellapárok között az egyetlen különbség a talajművelés típusa volt, ám ezek minden más szempontból (vetésforgó, növénytípus, műtrágyázás, gyomirtás és növényvédelem) ugyanabban a kezelésben részesültek. Az elemzésbe 16 éves (2004–2019) megfigyelésből származó adatokat vontunk be, ezalatt a vetésforgóban kukoricát (nyolcszor), őszi búzát (háromszor), repcét (kétszer), napraforgót (kétszer) és tavaszi árpát (egyszer) termesztettek.

A lefolyás mérésére egy speciális, kétcsatornás gyűjtőrendszert fejlesztettek ki, amely lehetővé tette mind a gyakori, alacsony intenzitású események, mind a ritka (<1% valószínűségű), nagy intenzitású csapadékesemények lefolyásának gyűjtését.

Egy lefolyásesemény olyan csapadékesemény, amely lefolyáshoz vezet, és legalább hat órás különbség van az előző és az azt követő esemény között. Minden lefolyásesemény után a lefolyt tömeg rövid ideig tartó ülepítése után megmérték a folyékony rész térfogatát, amiből meghatározhatóvá vált a talajveszteség (SL) és a lefolyás (RO). A növénytakaró mértékét havonta, vizuális felméréssel, 1 m^2-en határozták meg egy 1,0 × 1,0 m-es keret segítségével, öt ismétléssel. A meteorológiai adatokat a kísérleti helyszínen lévő automata meteorológiai állomás gyűjtötte ötpercenként. A kísérlet tizenhat éve alatt a meteorológiai állomást elektromos és felszerelési hibák, állatkárok és villámcsapás sújtották. Ennek következtében a csapadékadatok mintegy 20 %-a sajnálatos módon hiányzott.

6.2.3 Módszer

Az általános és általánosított lineáris modellek az összefüggések komplexitása és a kapcsolatok nemlineáris tulajdonságai miatt nem alkalmasak a talajerózió előrejelzésére, ezért egy robusztusabb, nemparametrikus módszert választottunk. A tizenhat éves nagyparcellás szántóműveléses kísérlet adatait felhasználva a teljes adathalmazra és a

kukoricakultúra részadathalmazra vonatkozóan a lefolyást és a talajveszteséget négy különálló véletlen erdő (RF) klasszifikációs modellel becsültük.

A tizenhat évet felölelő adatsorunk 560 feljegyzést tartalmazott 140 lefolyás eseményről és a kapcsolódó csapadék-, lefolyás- (RO) és talajveszteségadatokról (SL). Négy faktor- (F) és három folytonos (C) típusú magyarázó változót választottunk ki: (F1) kultúrnövény (6 szinttel: a legtöbb a kukorica volt 244 rekorddal, búza, napraforgó, repce, tavaszi árpa, parlag), (F2) talajművelés (2 szinttel: PT és CT), (F3) növénytakaró (5 szinttel: 1-től 5-ig terjedő skála 0 és 100% között 20%-os lépésekkel), (F4) az esemény hónapja (12 szinttel januártól decemberig), valamint (C1) a csapadék mennyisége (mm), (C2) időtartama (h) és (C3) intenzitása (30 perces maximális intenzitás, I_{30}, mm/h). A lefolyás (RO) és a talajveszteség (SL) függő változóként szolgált. Az adathalmazunk, bár hosszú távú volt, szűkössége miatt regressziós becslésre nem volt alkalmas, ezért a lefolyás (RO) és a talajveszteség (SL) változókat négy-négy kategóriára osztottuk az eloszlásuk és a szakemberek által megítélt súlyosságuk figyelembevételével. Az RO és SL kategóriákat, valamint a hiányzó és a teljes csapadékadatok számát a 11. Táblázat tartalmazza. A hiányzó értékeket a *k*-legközelebbi szomszéd módszer segítségével imputáltam (Tang és Ishwaran, 2017, ld. 5.1.1. fejezet).

11. Táblázat A lefolyás (RO) és a talajveszteség (SL) négy kategóriája; az RO és SL adatok összes kultúrára (Összes), illetve a kukoricakultúrára (Összes kukorica) vonatkozó száma; a csapadékváltozókból (mennyiség, időtartam, intenzitás) hiányzó értékek száma az összes kultúra, illetve a kukoricakultúra adataiból (Hiányzó és Hiányzó kukorica)

	Kategóriák	Kockázati csoportok	Összes	Hiányzó	Összes kukorica	Hiányzó kukorica
Elfolyás (RO, mm)	$RO = 0$	1 (alacsony)	245	55	69	3
	$0 < RO < 0{,}3$	2 (közepes)	203	59	95	21
	$0{,}3 < RO < 1{,}2$	3 (magas)	52	10	34	4
	$1{,}2 \leq RO$	4 (nagyon magas)	60	8	46	4
		összesen	560	132	244	32
Talajveszteség (SL, t/ha)	$SL = 0$	1 (alacsony)	181	39	45	2
	$0 < SL < 0{,}03$	2 (közepes)	193	59	92	17
	$0{,}03 < SL < 0{,}15$	3 (magas)	97	23	49	8
	$0{,}15 \leq SL$	4 (nagyon magas)	89	11	58	5
		összesen	560	132	244	32

Az adathalmazt véletlenszerűen, rögzített arányban (75% és 25%) tanuló és tesztelő részhalmazokra osztottam (*stratified sampling*, ld. 4.4. fejezet). Az RF-modell osztályozási teljesítményének mérésére a tesztadathalmazra a 6. Táblázatban (ld. 4.6. fejezet) ismertetett mérőszámokat használtam. Az *mtry* és *ntree* hiperparamétereket (ld. 5.3.2. fejezet) az OOB hiba (ld. 2.2. fejezet) és a négyzetes hibaátlag négyzetgyökének (RMSE, ld. 4.6. fejezet) minimalizálásával optimalizáltam, figyelembe véve az egyéb osztályozási teljesítmény-mértékeket is (pontosság, F1-érték, AUC, ld. 4.6. fejezet). Az *ntree* esetében az optimális értékek a különböző RF-modellekre 400, 500 vagy 550, az *mtry* esetében 4 vagy 5 volt.
Az egyes lépésekben a magyarázó változók pusztán a Gini-index (ld. 5.1.3. fejezet) átlagos csökkenésének elvén alapuló választása azonban torzított lehet, ha a magyarázó változókban a hiányzó adatok különböző mértékűek, illetve, ha a faktortípusú magyarázó változók szintjeinek száma eltér, mely kritériumok esetünkben fennálltak (Breiman, 1984, Strobl és mtsai, 2005). Ezért a megbízhatóság és torzításmentesség biztosítása céljából a fák tördeléséhez felosztási szabályként (*splitting rule*) a pontosság (*accuracy*) átlagos

csökkentésének és a Gini-index átlagos csökkenésének elvét egyszerre figyelembe vevő módosított (kombinatorikus) kiválasztási szabályt alkalmaztam (Strobl és mtsai, 2006).

A változók fontosságát a teljes átlagos pontosságcsökkenés (*mean decrease accuracy*, MDA, ld. 5.3.2. fejezet) alapján a modell egészére és az RO, illetve SL kategóriákra külön-külön is kiszámoltam a következőképpen (Strobl és mtsai, 2008; Louppe és mtsai, 2013; Gregoruttiet és mtsai, 2015): egy adott i fa esetében először is az OOB mintakészletből kiszámoltam a pontosságot (*accuracy*, $Ac_{i,\,\text{original}}$). Ezután egy adott j magyarázó változó értékeit véletlenszerűen permutáltam, míg az összes többi változó értékét változatlanul hagytam, és a permutált adatokkal kapott modell pontosságát is kiszámítottam ($Ac_{i,j,\,\text{permutált}}$). Végül vettem e két pontossági érték különbségét, és ezeknek a különbségeknek az összes fán (N_{trees}) mért átlagát képeztem:

$$MDA_j = \frac{1}{N_{trees}} \sum_{i=1}^{N_{trees}} \left(Ac_{i,\,original} - Ac_{i,j,\,permutált} \right).$$

Így kaptam egy változófontossági mérőszámot a j változóra, az ún. teljes átlagos pontosságcsökkenés értékét (Breiman, 2001). Az MDA-t a kimeneti kategóriák szerinti bontásban külön-külön is előállítottam.

A megfigyelt és az előre jelzett kategóriák közötti egyezés mértékét a Cohen-féle Kappa (Cohen, 1960, ld. 4.6. fejezet) segítségével fejeztem ki. Mint láttuk, az RO és SL kategóriáinak 1-től 4-ig terjedő értékei ordinális skálát képeznek az alacsonytól a közepes, magas és nagyon magas kockázatig terjedően. Így a téves besorolást nem lehet egyformán kezelni, ha az előrejelzés alacsony vagy magas kockázatot becsült, amikor az egy nagyon magas kockázatú eseménynél tévedett (azaz nem tekinthető ugyanakkorának a becslés hibája, ha egy vagy két kategóriányit téved). Ezért inkább a súlyozott Cohen-féle Kappát (Cohen, 1968) használtam, amely növekvő súlyokat rendel a növekvő téves besorolási hibákhoz (Fleiss és Cohen, 1973; Fleiss és mtsai, 2003).

6.2.4 Eredmények

Bár nem tudjuk felsorolni, pláne nem mérni az erózió mértékét befolyásoló összes tényezőt, feltételeztük, hogy a talajművelés, a növényfaj, a hónap, a növénytakaró, valamint a csapadék mennyisége és intenzitása alkalmas az eróziós kockázat bizonyos pontosságú előrejelzésére. A RO és SL adatok diszkretizálásával a véletlen erdő (RF) klasszifikációs módszer képes volt az eróziós kockázat előrejelzésére a rendelkezésre álló adatokból. A teljes adathalmazra vonatkozóan az RO és SL esetében az RMSE értékek 0,76, illetve 0,79 voltak. A tesztelő és a tanuló adathalmazra számított pontosságok aránya mind az RO, mind az SL esetében 0,99 volt, ami megerősítette, hogy a túlillesztés nem torzította az eredményeinket (11. Táblázat).

A tesztelő adathalmazra alkalmazott RF modell pontossága 0,71 és 0,76 volt az RO és SL esetében. Ezek az értékek magasnak tekinthetők, figyelembe véve az RO és SL változókra ható tényezők összetettségét. A kategóriák érzékenysége, pontossága, kiegyensúlyozott pontossága, AUC-értékei és F1-értékei alapján az RF-modell a legjobban az 1. és 2. kategóriákat jósolta meg, vagyis amikor nulla vagy kismértékű elfolyás, vagy talajveszteség történt. E csoportok érzékenységi értékei meghaladták a 80%-ot az RO és a 72%-ot az SL esetében, de minden kategóriában a pontosság 0,7, a kiegyensúlyozott pontosság 0,8, az AUC-érték 0,85, az F1-érték 0,74 felett volt. A szélsőséges talajveszteség előrejelzése volt azonban a legfontosabb, mivel a teljes SL 56%-át (PT) és 78%-át (CT) a hét legnagyobb intenzitású csapadékesemény okozta. E negyedik, legsúlyosabb kategóriának az érzékenysége az SL esetében 0,73 volt, míg az RO esetében csak 0,5. A

pontosság, a kiegyensúlyozott pontosság és az AUC meghaladta a 0,72, 0,73, illetve 0,92 értéket mind az RO, mind az SL esetében. Az F1-érték ebben a kategóriában csupán 0,59 volt az RO esetében, az SL esetében azonban elérte a 0,76-ot. Ezért arra a következtetésre jutottam, hogy a legmagasabb kockázatú RO kisebb sikerrel jelezhető előre, mint a legnagyobb kockázatú SL. A mérsékelt RO és az SL (3. kategória) kevésbé megbízhatóan becsülhető az RF segítségével, az érzékenység 0,50 és 0,08 volt az RO és az SL esetében.

12. Táblázat A lefolyás (RO, mm, balra) és talajveszteség (SL, t/ha, jobbra) függő változókra és az összes kultúrnövényre alkalmazott véletlen erdő (RF) modell osztályozási minőségi mutatói és a különböző változók relatív fontossága TP: helyes pozitív; TN: helyes negatív; FP: hamis pozitív; FN: hamis negatív döntés; N = a megfigyelések teljes száma; ROC: receiver operating characteristic curve; AUC: Area Under the ROC curve; művelés (szántás vagy kímélő talajművelés); kultúrnövény (búza, napraforgó, repce, tavaszi árpa, parlag), növénytakaró (1-től 5-ig terjedő skála 0 és 100% között 20%-os lépésekkel), az esemény hónapja (januártól decemberig).

Összes kultúrnövény	Elfolyás (RO)				Talajveszteség (SL)			
	Pontosság (Accuracy)	Súlyozott Kappa	OOB hiba	Pontosság arány (Accuracy Ratio) Teszt/Tanuló	Pontosság (Accuracy)	Súlyozott Kappa	OOB hiba	Pontosság arány (Accuracy Ratio) Teszt/Tanuló
	0,71	0,69***	48,46%	0,99	0,76	0,75***	38,81%	0,99
	Csoportonkénti statisztikák							
Kategóriák	1	2	3	4	1	2	3	4
RO[mm], SL [t/ha]	RO=0	0<RO<=0,3	0,3<RO<1,2	1,2<=RO	SL=0	0<SL<=0,03	0,03<SL<0,15	0,15<=SL
OOB hiba	0,32	0,39	0,75	0,75	0,18	0,38	1,00	0,71
Szenzitivitás ($TP_i/(TP_i+FN_i)$)	0,80	0,81	0,50	0,50	0,90	0,78	0,08	0,73
Specificitás ($TN_i/(TN_i+FP_i)$)	0,88	0,81	0,92	0,97	0,82	0,82	1,00	0,98
Pozitív predikciós érték (precizitás)	0,77	0,70	0,57	0,73	0,80	0,71	1,00	0,79
Negatív Pozitív predikciós érték $NPV_i=(TN_i/(TN_i+FN_i))$	0,90	0,89	0,90	0,91	0,92	0,87	0,91	0,97
Prevalencia ($(TP_i+FN_i)/N$)	0,32	0,35	0,17	0,16	0,44	0,36	0,09	0,11
Detekciós szint (TP_i/N)	0,26	0,28	0,09	0,08	0,39	0,29	0,01	0,08
Detekciós Prevalencia ($(TP_i+FP_i)/N$)	0,34	0,40	0,15	0,11	0,49	0,40	0,01	0,10
Kiegyensúlyozott pontosság ($(Szenzitivitás_i+Specificitás_i)/2$)	0,84	0,81	0,71	0,73	0,86	0,80	0,54	0,85
AUC	0,86	0,85	0,80	0,93	0,93	0,88	0,71	0,92
F1 érték $(2\frac{PPV*Szenzitivitás}{PPV+Szenzitivitás})$	0,78	0,75	0,53	0,59	0,85	0,75	0,14	0,76
	A változók fontossága (*MDA*)							
Kategóriák	1	2	3	4	1	2	3	4
Művelés	**44,75**	**24,44**	**20,49**	**19,03**	**35,41**	**18,3**	**7,21**	8,54
Kultúrnövény	**21,75**	**16,78**	**7,22**	**15,84**	**14,65**	9,53	**5,17**	**17,42**
Növénytakaró	0,99	8,55	6,29	12,54	8,08	5,68	0,00†	10,65
Hónap	10,76	11,28	5,1	6,79	11,23	9,89	0,86	**12,18**
Csapadék (mm)	5,45	7,46	0,74	3,16	11,00	7,33	1,55	2,16
Csapadékhullás időtartama (óra)	**11,84**	7,27	2,01	6,53	13,37	**19,31**	3,31	8,33
Csapadékintenzitás	8,47	**15,88**	**14,09**	**13,47**	**24,41**	**25,3**	**5,93**	**16,44**

*** szignifikáns $p < 0{,}001$; a három legmagasabb fontosságú változók kategóriánként félkövérrel jelölve
† Egy alacsony prediktív erővel rendelkező változó pontossága a permutáció során véletlenszerűen kis mértékben növekedhet, ami így negatív fontossági pontszámot eredményezhet; ezeket nullának tekintettük.

A tévesen besorolt RO és SL események 74%-a, illetve 78%-a került a szomszédos csoportokba, többnyire alacsonyabb kategóriába (RO: 66%; SL: 73%), igen magas súlyozott Cohen-féle Kappa-értékekkel (RO: 0,69; SL: 0,75, $p < 0{,}001$). Ez arra utalt, hogy (érthetően) nem sikerült robusztus határokat találni a 2. és 3. kategória között, és ezért a modell alulbecsülte ezeket a változók nagyfokú változékonyságából adódóan.

Az átlagos pontosságcsökkenés (MDA) azt mutatja, hogy mennyivel csökken a pontosság, ha egy változó prediktív erejét értékei véletlen permutálásával elrontjuk. A csoportváltozók fontossági értékei azt mutatták, hogy a talajművelés és a növényfajoknak a talajveszteségre (RO) gyakorolt hatása volt a legfontosabb. Ezzel szemben az elfolyás (SL) meghatározásában az egyik legfontosabb változó a csapadék intenzitása volt, amelyet az első három csoportban a talajművelés követett. A 10%-os lejtésnél a csapadék mennyisége viszonylag kevésbé volt fontos a talajveszteség kockázatának szempontjából. A talajművelés fontossága a nagyobb talajveszteségű eseményeknél csökkent.

A növénytakaró változó fontossága átlagosan 6,6 volt, ami alacsonyabb volt a vártnál. Az RF azonban nem ad információt arról, hogy egy változó főhatásként vagy egy kölcsönhatás részeként fontos-e. Ennek megfelelően a növénytakaró alacsony értékeit a növénykultúrával és a hónappal együtt kell értékelni (fontossági átlagaik rendre 13,6, illetve 8,5), mivel ezek együtt jelentős tényezőt alkotnak.

A modellezést elvégeztem csak a kukoricára vonatkozó részadathalmazra is, amely a vetésforgóban a leggyakoribb volt. Bár így kevesebb adattal dolgoztunk (11. Táblázat), az RF-modell megbízhatósága nem csökkent számottevően (13. Táblázat). A kukorica adathalmazra korlátozott és a tesztadathalmazra alkalmazott RF-modell általános pontossága a RO és SL esetében 0,73, illetve 0,71 maradt. Az érzékenység, a pontosság, a kiegyensúlyozott pontosság, az AUC-értékek és a kategóriák F1-értékei alapján azonban az RF modell a negyedik (legsúlyosabb) kategóriát becsülte a legsikeresebben, amelyet az első (legkevésbé súlyos) kategória követett. A legmagasabb és legalacsonyabb kockázatú csoportok érzékenységi értékei meghaladták a 64%-ot az RO és a 82%-ot az SL esetében, és mindegyik csoportban meghaladték a 0,7 pontosságot, a 0,8 kiegyensúlyozott pontosságot, a 0,9 AUC és a 0,75 F1-értéket.

A legrosszabbul teljesítő kategória ismét a harmadik volt. Azonban, míg az érzékenységi értékek az RO esetében az 1. kategória részadathalmazán (0,64), az SL esetében pedig az 1. és 2. kategória részadathalmazán (0,82 és 0,71) alacsonyabbak voltak, mint a teljes adathalmazon (RO: 0,80; SL: 0,90 és 0,78), a 3. és 4. kategória érzékenységi értékei magasabbak voltak (RO: 0,67 és 0,64, szemben a 0,50 és 0,50 értékekkel a teljes adathalmazon; SL: 0,13 és 0,92, szemben a 0,08 és 0,73 értékekkel a teljes adathalmazon), ami azt jelzi, hogy a kockázatos kategóriák jobban előrejelezhetők a kukoricakultúrában. Az RO RF-modelleket a kukorica-részhalmazon és a teljes adathalmazon összehasonlítva a pontosság, a kiegyensúlyozott pontosság, az AUC és az F1-értékek a kukorica-részhalmazon a magasabb kockázatú (3. és 4.) kategóriák esetében jelentősen magasabbak voltak. Ezért arra a következtetésre jutottunk, hogy a kukoricakutúrákra vonatkozó legmagasabb kockázat előrejelzése a talajerózióra nem kevésbé megbízható, mint a lefolyásra.

A tévesen besorolt RO és SL esetek 74%-a, illetve 81%-a került a szomszédos csoportokba, többnyire alacsonyabb kategóriába (RO: 66%; SL: 72%), bár igen jelentős súlyozott Cohen-féle Kappa-értékekkel (RO: 0,71, SL: 0,78, mindkettő $p < 0{,}001$). A talajművelés relatív jelentősége a kukorica esetében volt a legnagyobb az RO minden kategóriájában. Az SL tekintetében a talajművelés volt a legfontosabb jellemző a legkevésbé kockázatos kategóriában. A többi kategóriára vonatkozóan más változók (növénytakaró, hónap) voltak a legfontosabbak az előrejelzésben. A növénytakaró relatív

szerepe a kukoricakultúra esetében nagyobb volt, különösen a nagyobb eróziós események vonatkozásában.

13. Táblázat A lefolyás (RO, mm, balra) és talajveszteség (SL, t/ha, jobbra) függő változókra és a kukorica kultúrnövény adatsorra alkalmazott véletlen erdő (RF) modell osztályozási minőségi mutatói és a különböző változók relatív fontossága TP: helyes pozitív; TN: helyes negatív; FP: hamis pozitív; FN: hamis negatív döntés; N = a megfigyelések teljes száma; AUC: Area Under the ROC curve; ROC: receive operating characteristic curve; művelés (szántás vagy kímélő talajművelés); növénytakaró (1-től 5-ig terjedő skála 0 és 100% között 20%-os lépésekkel), az esemény hónapja (januártól decemberig).

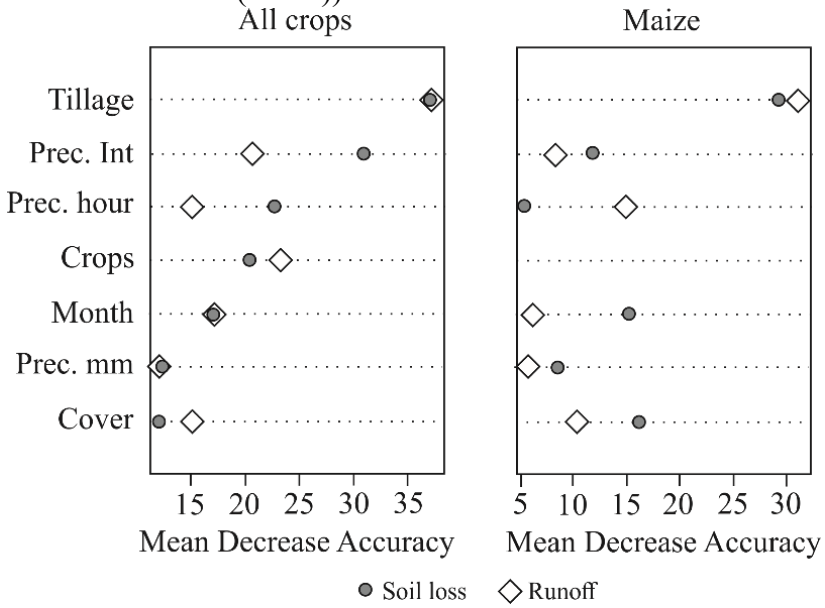
Kukorica kultúrnövény	Elfolyás (RO)				Talajveszteség (SL)			
	Pontosság (Accuracy)	Súlyozott Kappa	OOB hiba	Pontosság arány (Accuracy Ratio) Teszt/Tanuló	Pontosság (Accuracy)	Súlyozott Kappa	OOB hiba	Pontosság arány (Accuracy Ratio) Teszt/Tanuló
	0,73	0,78***	56,52%	0,99	0,71	0,71***	91%	0,96
	Csoportonkénti statisztikák							
Kategóriák	1	2	3	4	1	2	3	4
RO[mm], SL [t/ha]	RO=0	0<RO<=0,3	0,3<RO<1,2	1,2<=RO	SL=0	0<SL<=0,03	0,03<SL<0,15	0,15<=SL
OOB hiba	0,65	0,38	0,78	0,61	0,52	0,35	0,92	0,56
Szenzitivitás ($TP_i/(TP_i+FN_i)$)	0,64	0,87	0,67	0,64	0,82	0,71	0,13	0,92
Specificitás ($TN_i/(TN_i+FP_i)$)	1,00	0,73	0,90	0,98	0,89	0,76	1,00	0,92
Pozitív predikciós érték (precizitás)	1,00	0,67	0,62	0,90	0,74	0,65	1,00	0,73
Negatív Pozitív predikciós érték $NPV_i=(TN_i/(TN_i+FN_i))$	0,92	0,90	0,91	0,90	0,93	0,80	0,88	0,98
Prevalencia $((TP_i+FN_i)/N)$	0,18	0,38	0,20	0,23	0,28	0,39	0,13	0,20
Detekciós szint (TP_i/N)	0,12	0,33	0,13	0,15	0,23	0,28	0,02	0,18
Detekciós Prevalencia $((TP_i+FP_i)/N)$	0,12	0,50	0,22	0,17	0,31	0,43	0,02	0,25
Kiegyensúlyozott pontosság $((Szenzitivitás_i+Specificitás_i)/2)$	0,82	0,80	0,78	0,81	0,85	0,73	0,56	0,92
AUC	0,93	0,84	0,87	0,96	0,90	0,77	0,69	0,95
F1 érték $(2\frac{PPV*Szenzitivitás}{PPV+Szenzitivitás})$	0,78	0,75	0,64	0,75	0,78	0,68	0,22	0,81
	A változók fontossága (*MDA*)							
Kategóriák	1	2	3	4	1	2	3	4
Művelés	**25,49**	**14,92**	**14,35**	**10,31**	**28,49**	2,74	**2,21**	**11,35**
Kultúrnövény	0,00†	7,79	0,00†	**9,70**	5,18	**8,13**	0,00†	**16,99**
Növénytakaró	0,00†	4,57	2,78	2,29	0,23	**10,3**	**4,54**	**14,99**
Hónap	7,33	0,88	1,00	1,90	5,29	5,11	0,00†	4,67
Csapadék (mm)	**8,07**	**12,03**	2,51	2,34	0,84	5,43	1,97	6,08
Csapadékhullás időtartama (óra)	0,00†	8,86	**3,29**	5,26	**10,19**	**6,56**	0,18	4,72
Csapadékintenzitás	**25,49**	**14,92**	**14,35**	**10,31**	**28,49**	2,74	**2,21**	**11,35**

*** szignifikáns $p < 0{,}001$; a három legmagasabb fontosságú változók kategóriánként félkövérrel jelölve
† Egy alacsony prediktív erővel rendelkező változó pontossága, a permutáció során véletlenszerűen kismértékű növekedhet, ami így negatív fontossági pontszámot eredményezhet, Ezeket nullának tekintettük.

Összességében, minden kategóriát figyelembe véve, eredményeink szerint a legfontosabb változó a talajművelés volt (27. ábra). A teljes adatsort tekintve a csapadék

intenzitása volt a legfontosabb elfolyást magyarázó változó, míg a talajveszteség esetében a három fő változó (csapadékintenzitás, -tartam és növénykultúra) együttes hatása volt a második legfontosabb változó. A kukorica részadathalmaz esetében a csapadék időtartama, a növénytakaró és a csapadékintenzitás voltak a legfontosabb talajveszteséget magyarázó változók, míg az elfolyás esetében a növénytakaró, a hónap és a csapadékintenzitás együttes hatása volt a legmeghatározóbb.

27. Ábra A változók általános fontossága, a pontosság átlagos csökkenése (*MDA*) alapján az összes (All crops, balra), illetve a kukorica (Maize) kultúrnövényre vonatkozó adatokra (jobbra), a véletlen erdő (RF) klasszifikációs modell alapján, az elfolyás (Runoff) és a talajveszteség (Soil loss) függő változókkal. Magyarázó változók: Tillage: (művelés: szántás vagy kímélő talajművelés); Crops (kultúrnövény: búza, napraforgó, repce, tavaszi árpa, parlag), Cover (növénytakaró: 1-től 5-ig terjedő skála 0 és 100% között 20%-os lépésekkel), Month (az esemény hónapja: januártól decemberig), Prec.mm (csapadék mennyisége, mm); Prec.hour (csapadék időtartama (óra)), Prec. Int. (30 perces csapadékintenzitás maximuma (mm/h)).



6.2.5 Következtetés

Az RF klasszifikációs módszer képes volt az eróziós kockázat előrejelzésére. A kukorica adatállomány esetében az RF-modell a szélsőséges eseményeket jelezte előre a legjobban, majd a lefolyás nélküli kategóriát. A legnagyobb és a legkisebb kockázatot jelentő csoportok érzékenysége mind meghaladta a 82%-ot a talajveszteség és a 64%-ot a lefolyás esetében. A talajművelés típusa (kímélő vagy szántásos) volt a legfontosabb hatótényező. A hosszú távú vizsgálat azt mutatta, hogy a kímélő talajművelés alkalmazása lehetővé tette a csapadék jelentős részének megtartását a szántóföldeken, és ennek következtében a talajveszteség egy nagyságrenddel alacsonyabb maradt a tolerálható értéknél. Eredményünkkel megmutattuk, hogy a véletlen erdő modell alkalmas a lefolyás és a talajveszteség kockázatának modellezésére. Mivel az adatkészlet bővülésével várhatóan jelentősen tudnánk javítani az előrejelzések pontosságán és precizitásán, ezért eredményünk motiváció lehet a lefolyás és a talajveszteség rendszeres és megbízható monitoringrendszerének kifejlesztésére.

6.3 A relatív terméshozam és a talaj tulajdonságainak összefüggései

E fejezet az alábbi publikáción alapszik:

Juhos, K., Szabó, S., Ladányi, M. 2015. 'Influence of soil properties on crop yield: a multivariate statistical approach'. Int. Agrophys., 29(4), pp. 433-440. https://doi.org/10.1515/intag-2015-0049 **Q2**

6.3.1 Motiváció és célkitűzés

A terméshozam és a talaj közötti kapcsolat nagyon összetett: függ a talaj fizikai és kémiai tulajdonságai és számos természeti tényező közötti összetett kölcsönhatásoktól. A táj- és talajtulajdonságok változékonyságának és a terméshozamra gyakorolt hatásuknak a megértése a régióspecifikus és fenntartható gazdálkodási rendszerek és a földhasználat tervezésének kritikus eleme. A terméshozam előrejelzésére számos statisztikai módszert dolgoztak ki. E módszerek alkalmassága az adatbázis szerkezetétől és méretétől függ, de mindegyik módszernek megvannak a maga korlátai.

A talajmutatók termelékenységre gyakorolt hatásának leírására ezek kapcsolatának összetettsége és nemlineáris tulajdonságai miatt a szakirodalomban gyakran alkalmazott lineáris megközelítések általában nem szerencsések. A lineáris megközelítés a multikollinearitási problémát is magában hordozza (ld. 5.1.5. fejezet). Egyes, erősen korreláló változók eltávolítása a modellekből pedig fontos információk elvesztéséhez vezethet.

A független változók közötti multikollinearitási problémák megoldására több szerző a parciális legkisebb négyzetek regressziót (PLS, ld. 5.2.3. fejezet) alkalmazta (Corwin és mtsai, 2003; Ping és mtsai, 2004). A PLS-módszer eredményeképpen kapott loadingok értelmezésével azonosíthatóvá váltak azok a talajtulajdonságok, amelyek a legnagyobb hatással vannak a terméshozamra. A PLS-módszer által képzett főkomponensek, ha jól is magyarázzák a hozamot, a magyarázó változók közötti bonyolult összefüggéseket elfedhetik, és így az interpretációt nehezítik meg. Egy másik lehetőség az erősen korreláló változók kezelésére a főkomponens-elemzés (PCA, ld. 5.2.2. fejezet), illetőleg a főkomponens-regresszió (PCR, ld. 5.2.3. fejezet), melyek szintén elterjedtek a szakirodalomban (Ayoubi és mtsai, 2009; Cox és mtsai, 2003; Mallarino és mtsai, 1999; Shukla és mtsai, 2004). Stenberg (1998) szerint a dimenziócsökkentő eljárások, amellett, hogy információveszteséggel járnak, ismét nehezen interpretálható új változókat eredményeznek. A klasszifikációs és regressziós fák (CART) alkalmazásával az eredmények a módszer alacsony predikciós hibájáról számolnak be (Ahman és Bhatti, 2015; Tittonell és mtsai, 2007; Zheng és mtsai, 2009), ám ezeknek az eredményeknek sem egyszerű az ok-okozati elemzése.

A kutatásunk célja az volt, hogy világos interpretációt lehetővé tevő, és a talajparaméterek esetenként nemlineáris összefüggéseit is figyelembe vevő módszerrel feltárjuk a talajtulajdonságok és a hozam közötti kapcsolatot egy kelet-magyarországi régióban az időjárási viszonyok függvényében, és bizonyítsuk, hogy ezzel a módszerrel a terméshozamra pontosabb régióspecifikus előrejelzés adható különböző időjárási körülmények esetén.

6.3.2 Az adatok alapjául szolgáló megfigyelések

A 28 parcellából álló, mintegy 225 hektáros kutatási terület Kelet-Magyarországon található, ahol őszi búza (*Triticum aestivum* L.), kukorica (*Zea mays* L.) és napraforgó (*Helianthus annuus* L.) termesztése folyt. Minden egyes parcellán 2004 és 2013 között

feljegyezték a termesztett növény termésadatait. A vizsgált tíz év alatt négy aszályos év volt (2007; 2009; 2012 és 2013), amikor áprilistól augusztusig 200 mm-nél kevesebb csapadék hullott (Pálfai-féle aszályindex, PDI, Pálfai, 2002).

A talaj jellemzésére 11 talajtulajdonság adatait használtuk: pH, $CaCo_3$ (összes karbonát, %), EC (konduktivitás, μS/cm), AL-Na (oldható nátrium, mg/kb), tengerszint feletti magasság (m), leiszapolható rész <0,02 mm százalékos aránya, agyag <0,002 mm százalékos aránya, OC (humusz, %), AL-P_2O_5 (elérhető foszfor, mg/kg), AL-K_2O (elérhető kálium, mg/kg), Hargitai-N (szerves tápanyag, mg/kg).

6.3.3 A kutatás hipotézise

Hipotézisünk szerint az egyszerű talajtani mutatók önmagukban nem alkalmasak arra, hogy feltárják a talaj tulajdonságai és a terméshozam közötti kapcsolatot, és ezért az összefüggéseket mélyebben feltáró módszer szükséges e kapcsolat megértéséhez.

6.3.4 Módszer

A 11 talajparamétert két csoportba osztottuk aszerint, hogy az adott paraméter (ebben a régióban) minél nagyobb, vagy minél kisebb értéke előnyös.

Mivel a talaj potenciális termőképessége nem egyenlő az éppen aktuális hozamokkal, és más külső tényezők is befolyásolják azokat, a nyers hozamadatok elemzése helyett tíz évre vonatkozó relatív termésátlag- és termésváltozékonysági mutatókat képeztem. Az adatokat először normáltam; képeztem az egyes növényeknek a p parcellán vett RY_p relatív hozamát ($0<RY_p \leq 1$), amelyet az alábbiak szerint számoltam ki:

$$RY_p = \frac{Y_p}{Y_{max}},$$

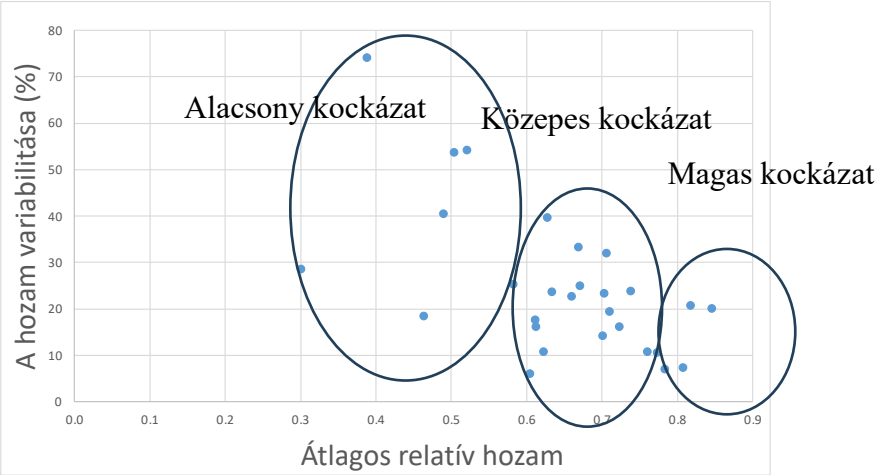
ahol Y_p a p parcella hozama (t/ha), $Y_{\max}$ az adott kultúrnövény maximális hozama a 28 parcellára és a 10 évre nézve (t/ha). Ezután képeztem a parcellák relatív hozamának átlagát ($\overline{RY_p}$), amit kiszámítottam a 10 éves adatsorra, illetve külön a négy száraz évre (2007, 2009, 2012, 2013), amikor az április és augusztus között mért csapadékösszeg nem haladta meg a 200 mm-t, valamint a maradék hat (nem száraz) évre is. A p parcella hozamának variabilitását a variációs koeffifiensével fejeztem ki (%):

$$CV\left(RY_p\right) = 100 * \frac{SD(RY_p)}{\overline{RY_p}}.$$

A talajtulajdonságokat leíró változók adathalmazát főkomponens-regresszióval (PCR, ld. 5.2.3. fejezet) elemeztem. Első lépésben főkomponens-elemzést (PCA, ld. 5.2.2. fejezet) végeztem, a loadingokat varimax rotációnak vetettem alá. Annak érdekében, hogy jól interpretálható főkomponenseket kapjunk, a talajparamétereket először a (0,1) intervallumra képeztem úgy, hogy a magasabb értékek a talajfunkció szempontjából előnyösebb tulajdonságot képviseljenek. Az AL-P_2O_5 és a $CaCO_3$ változók ferdeségét logaritmus (ln) transzformációval kompenzáltam. Az átlagos relatív hozam ($\overline{RY_p}$) és a hozamváltozékonyság ($CV\left(RY_p\right)$) függő változókkal többszörös lineáris regressziót végeztem lépésenkénti (*stepwise,* ld. 5.1.5. fejezet) módszerrel, az 1-nél nagyobb sajátértékkel rendelkező főkomponensek felhasználásával.

6.3.5 *Eredmények*

Ha ábrázoljuk a hozam variabilitását ($CV(RY_p)$) az átlagos relatív hozam ($\overline{RY_p}$) függvényében, akkor a parcellák egyértelműen három termelékenységi osztályba sorolhatók a relatív termésátlag mentén (28. ábra). A legnagyobb terméshozamot és a legkisebb szórást a csernozjom talajokon, a közepes terméshozamot és közepes szórást a szolonyeces talajokon, a legkisebb terméshozamot és a legnagyobb szórást pedig a szürke erdőtalajokon találtuk.



28. Ábra A parcellák átlagos relatív terméshozama és termésváltozékonysága (1. osztály: alacsony kockázat, 2. osztály: mérsékelt kockázat, 3. osztály: magas kockázat). Referenciahozam: relatív terméshozam 1,0 értékkel = a 10 év maximális terméshozama (kukorica, őszi búza és napraforgó esetén: 10,0; 7,1, illetve 4,5 t/ha).

Az 1-nél nagyobb sajátértékek alapján a PCA három főkomponenst eredményezett, amelyek a teljes változókészlet varianciájának összesen 84,93%-át magyarázzák (14. Táblázat). Az 1. főkomponens (PC1) loadingjai alapján az alábbi változókkal állt magas korrelációban: EC, AL-Na, agyag, leiszapolható rész, pH és tengerszint feletti magasság. Az első faktor a teljes variancia 52,50%-át magyarázta, és jól megkülönböztette a szolonyec talajon lévő parcellákat a csernozjom talajon lévő parcelláktól. A PC2 magas korrelációban állt a $CaCO_3$, humusz (OC) és pH változókkal, és jól megkülönböztette az alacsony szervesanyag-tartalmú kalciumos szürke erdőtalajokat a kalciumos csernozjomoktól. A második faktor által magyarázott variancia 19,25% volt. A PC3 az AL-P2O5 és az AL-K2O változókkal állt magas korrelációban a teljes variancia 13,17%-át magyarázva. A modellbe bevont változók kommunalitásai magasak voltak (>0,82), kivéve egy változót (Hargitai-N), amely az elemzésben szereplő többi változóval csupán 40,2% közös varianciát tartalmazott.

A főkomponensekkel mint magyarázó változókkal többszörös lineáris regressziós modellel magyaráztam az összes évre számított átlagos relatív hozamot ($\overline{RY_p}$). A lépésenkénti (*stepwise*) módszer mindhárom főkomponenst bevette a modellbe, a magyarázott variancia $R^2 = 0{,}49$ ($p < 0{,}01$) volt (15. Táblázat). A modellre vonatkozó F-teszt szignifikáns modellt eredményezett, és a PC1, PC2 és PC3 faktorok együtthatói a konstanssal együtt mind szignifikánsak voltak.

Csupán az aszályos, illetve a kevésbé száraz évek esetében külön-külön is illesztettem lineáris modellt az átlagos relatív hozamra a fentiekhez hasonlóan. Az aszályos évek esetében a szignifikáns modellbe az első (PC1) és a harmadik (PC3) főkomponens

került bele. A magyarázott varianciahányad $R^2 = 0{,}30$ ($p < 0{,}05$) volt. A PCR a kevésbé száraz években az átlagos relatív hozamra a PC2 és PC3 főkomponenseket választotta be a lineáris modellbe, így a magyarázott varianciahányad $R^2 = 0{,}40$ ($p < 0{,}01$) volt.

A termés változékonyságára ($CV(RY_p)$) vonatkozó PCR a PC1 és PC2 faktorokat választotta be a lineáris modellbe. A magyarázott varianciahányad $R^2 = 0{,}41$ ($p < 0{,}01$) volt.

14. Táblázat A főkomponens-elemzés eredménye az 1-nél nagyobb sajátértékeik alapján választott három főkomponens esetén: az egyes főkomponensek által magyarázott varianciahányad, a faktorloadingok (az egyes változóknak a főkomponensekkel való korrelációja), illetve a kommunalitások (az egyes változók által hordozott megosztott varianciahányad) a Kaiser-Meyer-Olkin-féle mérőszámmal, amely az adatok PCA-ra való alkalmasságát mutatja, illetve a Bartlett-féle szfericitás (Khi-négyzet) teszttel.

Főkomponensek	PC1	PC2	PC3	
Sajátértékek	5,776	2,118	1,449	
Magyarázott varianciahányad %	52,502	19,253	13,169	
Kumulált magyarázott varianciahányad %	52,502	71,758	84,926	
Faktorloadingok				Kommunalitás
konduktivitás (EC)	**0,971**	-0,074	-0,021	0,945
AL-Na	**0,970**	0,014	-0,062	0,947
agyag	**0,949**	0,122	0,184	0,965
leiszapolható rész	**0,850**	0,323	0,187	0,875
pH	**0,733**	0,591	0,275	0,960
tengerszint feletti magasság	**0,726**	0,370	0,395	0,826
$CaCO_3$	0,036	**-0,907**	0,146	0,904
humusz	0,533	**0,759**	-0,017	0,827
AL-K_2O	-0,043	0,001	**-0,927**	0,837
AL-P_2O_5	-0,119	0,121	**-0,926**	0,859
Hargitai-N	0,428	-0,345	0,315	0,353

Modelldiagnosztika: KMO $= 0{,}69$, $\chi^2(55) = 378{,}132$; $p < 0{,}001$.
A 0,7 feletti abszolútértékű loadingok félkövérrel szedve.

15. Táblázat A PCR modellek eredményei: a becsült paraméterek és a regressziós diagnosztika (a modell F értéke, a paraméterek Student-féle t értékei és a magyarázott variancia, R^2).

	Becsült paraméterek		$\|t\|$(df=26)	F	R^2
$\overline{RY_p}$ aszályos évek	constant	0,66	30,74***	5,36*	0,30*
	PC1	0,40	2,38*		
	PC2	Nem szerepel a modellben			
	PC3	0,38	2,25*		
$\overline{RY_p}$ kevésbé száraz évek	constant	0,64	36,90***	8,47**	0,40**
	PC1	Nem szerepel a modellben			
	PC2	0,52	3,34**		
	PC3	0,37	2,40*		
$\overline{RY_p}$ összes év	constant	0,64	34.30***	7,61**	0,49**
	PC1	0,36	2,45*		
	PC2	0,45	3,04**		
	PC3	0,40	2,75*		
$CV(RY_p)$	constant	24,88	10,46***	8,74*	0,41**
	PC1	-0,45	-2,94**		
	PC2	-0,46	2,98**		
	PC3	Nem szerepel a modellben			

*szignifikáns $p < 0{,}05$; **$p < 0{,}01$; $p < 0{,}001$ szinten

6.3.6 Következtetés

Eredményeink szerint az aszályos években a szikesedés, a talaj textúrája és a tápanyagtartalom határozta meg a terméshozamokat, míg a nedvesebb években az alacsonyabb domborzati helyzet, a talaj szerves anyaga és a tápanyagtartalom volt a fő korlátozó tényező. A relatív terméshozam variációs koefficienssel kifejezett változékonyságát elsősorban a szikesedés, a talaj textúrája, a domborzati helyzet, illetve a talaj szerves anyaga magyarázták.

6.4 Talajminőségi indikátorhalmaz

E fejezet az alábbi publikáción alapszik:

Juhos K., Czigány Sz., Madarász B., **Ladányi M.** 2019. 'Interpretation of soil quality indicators for land suitability assessment – A multivariate approach for Central European arable soils, Ecological Indicators, 99 pp. 261–272. https://doi.org/10.1016/j.ecolind.2018.11.063 **Q1**

6.4.1 Motiváció és célkitűzés

A talajok és azok funkciói kritikus fontosságúak a különböző ökoszisztéma-szolgáltatások biztosítása szempontjából. A talajoknak az ökoszisztéma-szolgáltatásokhoz való hozzájárulásának számszerűsítésére azonban még nemigen léteznek kielégítő módszerek. Ezért nehéz automatizálni és szabványosítani az indikátorok kiválasztására és értékelésére szolgáló matematikai és statisztikai módszereket.

Célunk egyrészt egy új, helyspecifikus talajminőségi és ökológiai indikátorhalmaz választásához és értékeléséhez használható módszer kidolgozása, másrészt a növények növekedését korlátozó tényezők azonosítása volt egy olyan adatbázis alapján, amely a közép-európai szántóföldekre jellemző leggyakoribb magyarországi talajokat reprezentálja.

6.4.2 Az adatok alapjául szolgáló mérések

Az értékeléshez 1045, összesen mintegy 5000 hektárnyi területet képviselő parcella adatait használtuk, ahol 0–30 cm-es rétegében felvételezésre került a talajszerkezet, a talajvízszint mélysége, a talaj szervesanyag-tartalma (SOM), a pH, a kalcium-karbonát-egyenértéke (CCE), a konduktivitás (EC), a Na, az elérhető N, P, K, Mg, S, Cu, Zn és Mn.

6.4.3 Módszer

A mintákat 25 talajtípusba soroltuk, amely a csernozjom, barna, illetve szürke erdőtalaj, réti talaj, agyag, homok, valamint szolonyec talajtípusok további finomításával került meghatározásra.

A változók páronkénti lineáris kapcsolatát a Pearson-féle korrelációs együtthatóval (r) mértem. Az ln-transzformált adatokra az azokban rejlő struktúra feltárása céljából főkomponens-elemzést (PCA, ld. 5.2.2. fejezet)) végeztem. Az 1-nél magasabb sajátértékű főkomponenseket az egyes változók varimax-forgatott loadingjai (a változó és a főkomponens közötti korrelációi) alapján értékeltem.

A talajparamétereknek a talajképzést magyarázó erejét diszkriminanciaelemzéssel (DA, ld. 5.1.4. fejezet) vizsgáltam, ahol az előzőekben kapott főkomponensek független változókként szerepeltek, a válaszváltozó pedig a 25 szintű talajtípus volt. Az adatok normalitását a változók eloszlásának ferdeségével és csúcsosságával ellenőriztem.

A célul kitűzött helyspecifikus indikátorhalmaz választásához és értékeléséhez használható függvények definiálásához a hazai trágyázás és talajjavítás hosszú távú kísérleti eredményeit és talajművelési módszereit vettük alapul. Számba vettük a hazai vetésforgókban leggyakrabban előforduló egyes növénykultúrák (búza, kukorica, napraforgó és repce) szakirodalomban fellelhető ökológiai igényeit, ám nem az egyes növényekre, hanem általánosan, az ilyen vetésforgókra alkalmas módszert kerestünk. A Juhos Katalin által nagy szakértelemmel és részletességgel végzett szakirodalmi áttekintést és adatgyűjtést követően az egyes talajparaméterek kritikus küszöbértékei által

meghatározott intervallumain azok földhasználati értékeit a (0,1) intervallumon lineáris vagy nemlineáris függvénnyel fejeztem ki, ahol a 0 az alkalmatlanságot, míg az 1 az optimális megfelelést jelentette. Mivel az egyes paraméterek nem értékelhetők egymástól függetlenül, az egymást közvetlenül befolyásoló talajtulajdonságokat figyelembe véve a modelleket egyes esetekben talajkategóriák szerinti bontásban definiáltam.

6.4.4 Eredmények

- A főkomponens-elemzés eredménye

Az 1-nél nagyobb sajátértékek alapján a PCA négy főkomponenst (PC1–PC4) eredményezett, amelyek a teljes változókészlet varianciájának összesen 75,66%-át magyarázzák (16. Táblázat). A változók kommunalitása minden változó esetén meghaladta a 0,588-at. A szemcseméret-eloszlás és a textúra által befolyásolt és a talaj termékenységét és termelékenységét kifejező tulajdonságok a PC1-ben fejeződnek ki a textúra, Mg, Cu, EC, SOM, K és Na magas loadingjai alapján. A PC1 a változók varianciájának 33,55 %-át magyarázza. A második főkomponens (PC2) a teljes variancia 22,04%-át magyarázza, és a talaj savasságát-lúgosságát fejezi ki a Mn, a CCE és a pH mutatók magas loadingértékei alapján A PC3, melyben a felvehető P és Zn loadingjai magas értékeket mutatnak, a variancia 10,93%-át magyarázza. A PC4 az elérhető nitrogén és kén magas loadingjaival a teljes variancia 9,13%-át teszi ki.

16. Táblázat A főkomponens-elemzés eredménye az 1-nél nagyobb sajátértékeik alapján választott négy főkomponens esetén: az egyes főkomponensek által magyarázott varianciahányad, a faktorloadingok (az egyes változóknak a főkomponensekkel való korrelációja), illetve a kommunalitások (az egyes változók által hordozott megosztott varianciahányad)

Főkomponensek	PC1	PC2	PC3	PC4
Sajátérték	4,70	3,09	1,53	1,28
Magyarázott varianciahányad %	33,55	22,04	10,93	9,13
Kumulált magyarázott varianciahányad %	33,55	55,94	66,53	75,66
Paraméterek (kommunalitás)	Faktorloadingok			
Textúra (0,88)	**0,879**	0,128	-0,177	-0,234
Mg (0,87)	**0,845**	-0,198	-0,071	-0,336
Cu (0,84)	**0,839**	-0,321	0,080	-0,145
konduktivitás (0,67)	**0,807**	-0,098	-0,053	0,066
K (0,74)	**0,672**	0,241	0,460	-0,124
SOM (0,75)	**0,645**	0,370	-0,002	-0,191
Na (0,68)	**0,638**	0,543	-0,391	0,153
CCE (0,91)	-0,073	**0,943**	-0,117	0,014
Mn (0,77)	0,094	**-0,812**	0,298	-0,092
pH (0,74)	-0,073	**0,812**	0,223	-0,164
P (0,82)	0,023	0,445	**0,717**	0.402
Zn (0,63)	0,425	-0,163	**0,602**	0,321
S (0,73)	0,434	0,086	-0,289	**0,673**
N (0,59)	0,438	-0,178	-0,079	**0,601**

CCE: a kalcium-karbonát-egyenértéke
A 0,5 feletti abszolútértékű loadingok félkövérrel szedve.

Az indikátorokat a következő kategóriákba soroltuk: (PC1a) a tápanyag-visszatartást és a kationcsere-kapacitást jellemző mutatók: textúra, SOM, konduktivitás és Na; (PC1b) a tápanyagok, amelyek viszonylag függetlenek a gazdálkodási gyakorlatoktól: K, Mg, Cu; (PC2) a bázistelítettséget meghatározó mutatók, melyek a tápanyagok elérhetőségét határozzák meg: pH, CCE, Mn; (PC3 és PC4) nagymértékben változó mennyiségben elérhető tápanyagok, melyek igen érzékenyek az éghajlatra és a talajművelés típusára: N, S, P, Zn.

- A diszkriminanciaelemzés eredménye

A lineáris diszkriminanciaelemzés során az előzőekben kapott négy főkomponens (PC1, PC2, PC3 és PC4) szkórjaival mint független változókkal kívántam magyarázni a talajtípus osztályozását. Az első és a második diszkriminanciafüggvény (DF) a független változók varianciájának 70,9%-át, illetve 27,1%-át magyarázta, azaz szinte teljes egészében magyarázta a teljes varianciát. A struktúramátrix értékei szerint a főkomponensek rangsora a DF1-ben PC1 (0,709), PC2 (-0,497), PC4(0,100) és PC3 (0,089), míg a DF2-ben PC2 (0,792), PC1 (0,542), PC3 (0,354) és PC4 (-0,022). A talajtípusok elsősorban a talajok fizikai és kémiai tulajdonságait jelző PC1 és PC2 értékek függvényében differenciálódtak. Ugyanakkor a PC3 és PC4 hatása kevésbé bizonyult fontosnak.

- A talajparaméterek földhasználati értékére vonatkozó elemzés

A pH földhasználati értékét egy olyan bilogisztikus modellel jellemeztem, amelynek a kritikus küszöbértékeivel határolt intervallumának egy belső pontjában van egy telítődési értéke (p_0), ez előtt monoton növekvő, ez után pedig monoton fogyó (meredeksége és inflexiós pontja a növekvő szakaszban p_1, illetve p_2, a csökkenő fázisban p_3, illetve p_4 (17. Táblázat).

Telítődési függvényt követő lebomlási függvényből álló összetett modellt használtam a textúra földhasználati értékének leírására. Ez esetben a talajvízmélység értékeit figyelembe véve optimalizáltam a növekvő és csökkenő fázisok meredekségeit (p_1, p_2), valamint az x tengelyen való eltolás és a csúcspont helyének paramétereit (p_3, p_4).

A konduktivitás (EC) és Na változók földhasználati értékét monoton csökkenő logisztikus modell („kevesebb jobb") segítségével jellemeztem, ahol p_0, p_1, p_2, p_3 ilyen sorrendben a mínusz, illetve plusz végtelenben vett határérték, a meredekség, illetve az inflexiós pont paraméterei.

A rendelkezésre álló K és P földhasználati értékének monoton növekvő logisztikus modelljeit („több jobb") jelentősen befolyásolja a talaj textúrája és pH-ja, ezért paramétereiket ennek megfelelően módosítottam.

A talaj szervesanyag-tartalmára (SOM), valamint a rendelkezésre álló Mg, Zn és Cu változókra telítődési modelleket definiáltam, ahol p_1 a telítődési érték, p_2 a meredekség. Itt is eltérő paraméterekkel különböztettem meg a talaj textúrájától függő földhasználati értéküket.

A rendelkezésre álló Mn telítődési modellje esetében a függvényparamétereket a talaj pH-értéke szerint különböztettem meg.

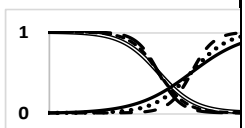
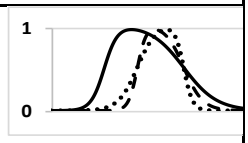
A talaj N- és S-tartalmának földhasználati értékét monoton növekvő lineáris („több a jobb") függvénnyel jellemeztem ($y = x/x_{\max}$, ahol $x_{\max}$ az adathalmazban szereplő maximális érték).

A talajt jellemző paraméterek talajhasználati értékét ezután kiszámoltam a 25 talajtípusra, illetve képeztem ezek átlagát. Ezen értékek alapján négy talajhasználati

kategóriát különböztettünk meg (18. táblázat): 0,8–1,00: nincs korlátozó hatása (üres kör); 0,6–0,81: enyhe korlátozó hatása van (harmadrész színezett kör); 0,4–0,61: erős korlátozó hatása van (kétharmadrészt színezett kör); 0,41 alatt pedig talajhasználati értéke csekély (fekete kör).

Ha eredményünket összevetjük a főkomponens-elemzés eredményeivel, akkor láthatjuk, hogy a talajt legkevésbé meghatározó, értékeiben igen változékony két paraméter, S és N (PC4) talajhasználati értéke csekély. Ez az eloszlásuk nagyon erős ferdeségéből és az átlagértékeikhez képest igen magas szórásaikból is adódik. A főkomponens-elemzés harmadik főkomponense a P és Zn varianciáját magyarázta, ezek talajhasználati értékei mutattak a legtöbb talajtípuson alacsonyabb értékeket.

17. Táblázat A talajt jellemző paraméterek talajhasználati értékét meghatározó görbék, azok paramétereinek jelentése, a talajhasználati értéket jellemző görbe értelmezési tartománya (azok kritikus küszöbértékei által meghatározott intervallumok), a paraméterek meghatározásánál figyelembe vett egyéb talajtulajdonság, valamint a függvények jelleggörbéi

Modell	Formula (Y az X földhasználati értéke)	Talajt jellemző paraméter (X)	X kritikus küszöbértékei által meghatározott intervallumok	figyelembe vett egyéb talaj-tulajdonság	Jelleggörbe
Telítődési	$Y = p_1 * (1 - exp(-p_2 * X))$ p_1 telítődési érték p_2 meredekségi tényező	SOM Mg Cu Mn Zn	0-4,5 m/m% 0-200 ppm 0-1,4 ppm 0-120 ppm 0-4 ppm	textura* textura** textura*** pH† textura***	
	textura	<30	30-37	38-42	43-50	>50
	SOM	$p_1 = 1{,}04$ $p_2 = 1{,}18$	$p_1 = 1{,}09$ $p_2 = 0{,}77$	$p_1 = 1{,}20$ $p_2 = 0{,}45$	$p_1 = 1{,}98$ $p_2 = 0{,}17$	$p_1 = 4{,}12$ $p_2 = 0{,}06$
	textura	<30	30-42	>42		
	Mg	$p_1 = 1{,}03;$ $p_2 = 0{,}04$	$p_1 = 1{,}07;$ $p_2 = 0{,}02$	$p_1 = 1{,}22;$ $p_2 = 0{,}01$		
	textura	<38	38-50	>50		
	Cu	$p_1 = 1{,}01; p_2 = 6{,}002$	$p_1 = 1{,}08; p_2 = 2{,}28$	$p_1 = 2{,}63; p_2 = 0{,}34$		
	Mn	$p_1 = 1{,}09; p_2 = 0{,}03$	$p_1 = 1{,}03; p_2 = 0{,}14$	$p_1 = 1{,}00; p_2 = 5{,}87$		
	Zn	$p_1 = 1{,}02; p_2 = 1{,}65$	$p_1 = 1{,}30; p_2 = 0{,}41$	$p_1 = 2{,}64; p_2 = 0{,}12$		
Logisztikus görbe	$Y = p_0 + \frac{(p_1 - p_0)}{1 + exp(-p_2 * (X - p_3))}$ p_0 és p_1 határérték $-\infty$-ben, $+\infty$-ben p_2 meredekségi tényező; p_3 inflexiós pont	AL-K2O AL-P2O5 konduktivitás Al-Na	0-260 ppm 0-350 ppm 0,2-2,0 dS/m 0-600 ppm	Textura* CCE††	
	textura	<30	30-37	38-42	43-50	>50
	AL-K2O	$p_0 = 0;$ $p_1 = 1{,}02$ $p_2 = 0{,}04;$ $p_3 = 90{,}47$	$p_0 = 0;$ $p_1 = 1{,}02$ $p_2 = 0{,}04;$ $p_3 = 124{,}19$	$p_0 = 0;$ $p_1 = 1{,}04$ $p_2 = 0{,}04;$ $p_3 = 151{,}27$	$p_0 = 0;$ $p_1 = 1{,}02$ $p_2 = 0{,}04;$ $p_3 = 161{,}54$	$p_0 = 0;$ $p_1 = 1{,}01$ $p_2 = 0{,}04;$ $p_3 = 171{,}39$
	CCE	<0,1	0,1-1,0	1,1-5,0	5,1-10	>10
	AL-P2O5	$p_0 = 0;$ $p_1 = 1{,}00$ $p_2 = 0{,}03;$ $p_3 = 66{,}45$	$p_0 = 0;$ $p_1 = 1{,}01$ $p_2 = 0{,}03;$ $p_3 = 85{,}05$	$p_0 = 0;$ $p_1 = 1{,}02$ $p_2 = 0{,}03;$ $p_3 = 108{,}09$	$p_0 = 0;$ $p_1 = 1{,}00$ $p_2 = 0{,}03;$ $p_3 = 126{,}95$	$p_0 = 0;$ $p_1 = 0{,}99$ $p_2 = 0{,}02;$ $p_3 = 153{,}82$
	Konduktivitás Al-Na	$p_0 = 1{,}15; p_1 = 0{,}00\ p_2 = 3{,}94; p_3 = 0{,}78$ $p_0 = 1{,}11; p_1 = 0{,}09\ p_2 = 0{,}01; p_3 = 173{,}22$				
Bilogisztikus görbe	$Y = p_1/(1 + exp(-p_2 * (X - p_3))) - p_1/(1 + exp(-p_4 * (X - p5)))$ p_0 telítődési érték p_1, p_3 meredekségi tényező; illetve p_2, p_4 inflexiós pont a növekvő/csökkenő fázisban	pH	3-11		
	pH	$p_0 = 1{,}09$; $p_1 = 1{,}47; p_2 = 4{,}42; p_3 = 2{,}91; p_4 = 7{,}99$				
Összetett (telítődési + lebomlási) görbe	$Y = \left(1 - exp(-p_1 * (X - p_3))\right) - (1 - exp(-p_2 * (X - p_4)^2))$ p_1, p_2 meredekségi tényező a növekvő/csökkenő fázisban; p_3 x tengely menti eltolás p_4 inflexiós pont	textura	20-70 cm	talajvízszint††††	

* textura kategóriák: <30 (homok); 30–37 (homokos vályog); 38–42 (vályog and iszapos vályog); 43–50 (agyagos vályog and iszapos agyag); >50 (agyag)

** textura kategóriák: <30 (homok); 30–42 (homokos vályog, vályog, iszapos vályog,); >42 (agyag vályog, iszapos agyag, agyag)

***textura kategóriák: <38 (homok and homokos vályog); 38–50 (vályog, iszapos vályog, agyag vályog, iszapos agyag); >50 (agyag)

† pH kategóriák <6; 6–8; >8; †† CCE kategóriák: <0,1; 0,1–1; 1,1–5; 5–10; >10 %;

††† talajvízszint-kategóriák: <50; 50–85; 86–120;121–180; >180

18. Táblázat A talajt jellemző 13 paraméternek a 17. Táblázatbeli módon definiált talajhasználati értékének átlagai a 25 talajtípusra, négy talajhasználati kategóriába sorolva: 0,81–1,00: nincs korlátozó hatása (üres kör); 0,61–0,81: enyhe korlátozó hatása van (harmadrész színezett kör); 0,41–0,61: erős korlátozó hatása van (kétharmadrészt színezett kör); 0,41 alatt pedig talajhasználati értéke csekély (fekete kör).

	Talajtípusok																								
	csernozjom			réti talaj	barna erdő-talaj	agyag	szürke erdőtalajok										barna erdőtalaj		homok		szolonyec		szürke erdőtalajok		
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
pH	○	○	○	○	○	◔	◔	○	○	○	◔	◔	○	◔	○	◔	○	◔	○	◔	○	○	◔	○	○
textura	○	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	◕	◕	◔	◔	○	○	○	○	○
EC	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
SOM	○	◔	○	◔	◔	◔	◔	◕	○	◕	◕	◔	◔	◔	◕	◔	◔	◔	◔	◔	◔	◕	◕	◔	◔
P	◕	○	◕	◕	●	◕	◔	◕	◕	◕	●	◔	◕	●	◔	◕	◕	○	◔	◔	○	◔	◔	◔	●
K	○	○	○	○	◔	◔	○	○	○	○	◔	○	○	○	○	◔	○	◔	○	◔	○	○	○	○	○
Mg	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Na	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	●	◔	◔	◕
Zn	◕	◔	◔	◔	◔	◔	◕	●	◔	◕	●	◔	◔	◔	●	◕	○	○	○	◔	◔	◕	◕	◕	○
Cu	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Mn	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
S	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
N	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

6.4.5 Következtetés

Kutatásunk célja egyrészt egy új, helyspecifikus talajminőségi és ökológiai indikátorhalmaz választásához és értékeléséhez használható módszer kidolgozása, valamint a vizsgált talajtípusokon a növénytermesztést korlátozó tényezők azonosítása volt. A talajhasználati értékek nemlineáris függvényekkel való kifejezése alapján a Magyarországon legjellemzőbb 25 vizsgált talajtípuson a leggyakoribb korlátozó tényezők a következők voltak: textúra, talajvízszint mélysége, SOM, pH, Na, rendelkezésre álló K, P és Zn. Ezek a mutatók a talajminőség értékeléséhez szükséges minimális adatkészletet indikálják. A talaj tulajdonságai azonban nem önállóan, hanem összetett és kombinált módon befolyásolják a termékenységet és a talaj termőképességét. Úgy véljük, hogy a talaj egy alkalmassági indexét alapozhatnánk ezekre a pontszámokra, ám óvakodnánk egy egyszerű additív integrációs módszer használatától. A talajminőség-értékelési modell továbbfejlesztése és egy komplex talajminőségi index kidolgozása és módszertanának javítása érdekében a következő céljaink a talajparaméterek hierarchikus rangsorolásának és csoportosításának meghatározására irányulnak.

6.5 A kanyargós szillevéldarázs fagytűrése és hőmérséklettől függő fejlődése

E fejezet az alábbi publikációkon alapszik:

Papp V., **Ladányi M.**, Vétek G. 2016. 'Temperature–dependent development of *Aproceros leucopoda* (Hymenoptera: Argidae), an invasive pest of elms in Europe' Journal of Applied Entomology, 42(6) pp. 589–597. https://onlinelibrary.wiley.com/doi/10.1111/jen.12503 **Q1**

Vétek G., Fekete V., **Ladányi M.**, Cargnus E., Zandigiacomo P., Oláh R., Schebeck M., Schopf A. 2020. 'Cold tolerance strategy and cold hardiness of the invasive zigzag elm sawfly *Aproceros leucopoda* (Hymenoptera: Argidae)'. Agricultural And Forest Entomology 22(3) pp. 231–237. https://doi.org/10.1111/afe.12376 **Q1**

Ezt a fejezetet egy gyors lefolyású covidfertőzés következtében fiatalon elhunyt kiváló szakembernek, kedves kollégámnak, Vétek Gábornak ajánlom. Vele dolgozni mindig inspiráló élmény volt, és sokat tanulhattam tőle, amiért mindig hálás leszek. A vele együtt elért eredmények bemutatásával szeretnék emléket állítani a növényvédelem szolgálatában végzett odaadó munkájának.

6.5.1 Motiváció és célkitűzés

Az Európában nem őshonos invazív kanyargós szillevéldarázs (*Aproceros leucopoda* Takeuchi, 1939, Hymenoptera: Argidae) az Európában való 2000-es évek eleji megjelenésétől kezdve igen gyorsan elterjedt. Azóta különböző szilfajokon jelentős károkról számoltak be mind erdőállományokban, mind városi környezetben.

A rovarok hideg hőmérsékletre adott válaszának három alapvető típusa van: (1) a fagyérzékeny fajok az alacsony hőmérséklet közvetlen hatására elpusztulnak; (2) a fagyintoleráns rovarok képesek túlélni a hideget, amíg nem képződik jég a testükben; (3) a fagytoleráns rovarok túlélik a sejten kívüli jégképződést. Számos rovarban fejlődött ki alkalmazkodási válasz a fagypont alatti hőmérsékletekre. Például egyes fajok fagyálló anyagok termelésével képesek csökkenteni a testnedvek fagyáspontját, és így szuperhűtött állapotban maradni. Azt a hőmérsékletet, amikor a spontán jégképződés végül bekövetkezik, szuperhűlési pontnak (Super Cooling Point, SCP) nevezik. A fagyintoleráns fajoknál a szuperhűlési pont az alsó letális hőmérsékletnek felel meg. A rovarok hidegtűrési stratégiájának megismeréséhez az első lépés a fajra jellemző SCP meghatározása.

A rovarok fejlődését meghatározó egyik legfontosabb környezeti tényező a hőmérséklet. Az egyes fejlődési szakaszok fejlődési sebességfüggvényét a rovarok állandó hőmérsékleti tartományban történő nevelésével lehet meghatározni a megfigyelt adatokra való megfelelő modell illesztésével.

A kanyargós szillevéldarázs telelésének biológiai aspektusait, illetve fejlődésük hőmérséklettől való függését korábban nem vizsgálták, márpedig a rovarok hidegtűrése az adaptációjuk egyik kulcskérdése. A rovarok fenológiájának előrejelzéséhez pedig feltétlenül szükséges a hőmérséklet és a fejlődési sebesség közötti kapcsolat ismerete.

Célkitűzéseink tehát a következők voltak: (1) a különböző európai országokban és években telelő kanyargós szillevéldarázs lárváira vonatkozó SCP-értékeinek becslése és statisztikai elemzése (2) a faj hidegtűrési stratégiájának azonosítása (3) annak meghatározása, hogy az áttelelő lárvák fagyérzékenyek, illetve fagyintoleránsak, esetleg fagytoleránsak-e, valamint azt, hogy (4) laboratóriumi kísérletek alapján meghatározzuk fejlődésük és túlélésük hőmérséklettől való függését.

6.5.2 Az adatok alapjául szolgáló megfigyelések, illetve kísérletek

2013/2014 őszén/telén négy alkalommal (október, november, január és február), valamint 2017 márciusában és áprilisában kerültek begyűjtésre az *A. leucopoda* vastag falú gubói egy Kecskemét közelében található erdős ültetvényből. 2015 márciusában két további helyszínről (Udine és Gonars, Olaszország) is kaptunk gubókat. Az SCP értékek mérése a BOKU intézet (Universität für Bodenkultur, Bécs, Ausztria) laboratóriumában történt.

A kanyargós szillevéldarázs hidegtűrését három hőmérsékleten és az azokhoz tartozó időintervallumban, 7 kísérleti beállításban mérték: -1,6 °C (10, 20, 30 nap), -4 °C (10, 20, 30 nap), -10,5 °C (9 nap), túlélésüket a kísérleteket követő első és harmadik napon rögzítették.

A lárvák hőmérsékletfüggő fejlődését szibériai szilfán (*Ulmus pumila*) nevelve vizsgálták hat állandó hőmérsékleten (11 °C, 15 °C, 18,5 °C, 23 °C, 25 °C, 27 °C) és 16 óra világos + 8 óra sötét (16L:8D) fotoperiódus mellett. A faj lárvái 4–7 stádiumon keresztül fejlődtek.

6.5.3 Módszer

Az SCP-adatok tisztításának első lépéseként a 312-ből 18 extrém magas SCP-értéket kizártam, melyeket valószínűleg a minták előkészítése során a lárva gubójának véletlen sérülése okozott (ld. 4. Példa). Az SCP-értékeket egytényezős robusztus ANOVA-val (Brown-Forsythe), valamint Games-Howell post-hoc teszttel hasonlítottam össze (ld. 3.1. fejezet). A hibatagok normalitását a ferdeségük és csúcsosságuk alapján ellenőriztem.

A kanyargós szillevéldarázs lárváinak túlélési arányát Marascuilo-eljárással hasonlítottam össze (Marascuilo, 1966, ld. 1. Példa).

A hőmérséklet és a peték, lárvák, első és második bábstádiumok, valamint egy teljes generáció fejlődési sebessége közötti kapcsolatot lineáris fok-nap-modellel (*day degree model*, DD, egyszerűsége miatt az eredményt itt nem közlöm) és nemlineáris modellekkel írtam le (ld. 5.1.5. fejezet).

A DD-modell szerint a fejlődési sebesség $r(T) =$ (1/fejlődési időnapokban) és a hőmérséklet közötti kapcsolat lineáris egyenlettel írható le: $r(T) = a + bT$, ahol T a kezdő hőmérséklet, a a tengelymetszet, és b a lineáris függvény meredeksége. A T_{min} alsó hőmérsékleti küszöbértéket ($T_{\min} = -a/b$) és a fejlődés befejezéséhez szükséges alsó küszöbérték feletti foknapok számát, a K állandót ($K = 1/b$) a lineáris egyenlet alapján számolhatjuk ki (Honěk, 1996; Li, 1998; Mathieu és mtsai, 2014).

A kanyargós szillevéldarázs fejlődésének alsó és felső hőmérsékleti küszöbértékének (amely az a hőmérsékleti érték, amelyeknél a fejlődés megáll) és optimális hőmérsékletének (amely az a hőmérséklet, amelyen a fejlődési idő várhatóan a legrövidebb) becslésére az alábbi formában definiált Lactin-2 modellt (Lactin és mtsai, 1995) használtam (29.(b) Ábra):

$$r(T) = e^{\rho T} - e^{\rho T_{\mathrm{L}} - (T_{\mathrm{L}} - T)/\Delta T} + \lambda$$

A függvényben ρ az optimális hőmérsékleten bekövetkező növekedés sebességét, T_{L} azt a hőmérsékletet jelöli, amelyen az életfolyamatok már nem tarthatók fenn hosszabb ideig, ΔT pedig a T_{L} és a T_{opt} optimális hőmérséklet különbsége. A képletben szereplő λ egy additív paraméter, amely lehetővé teszi a fejlődési küszöbérték becslését azáltal, hogy Logan és mtsai (1976) hasonló alakban

$$r(T) = \psi \cdot (e^{\rho T} - e^{\rho T_{\mathrm{L}} - \frac{T_{\mathrm{L}} - T}{\Delta T}})$$

felírt modelljéből a ψ (maximális fejlődési sebesség) konstans szorzót elhagyva és λ-t additív tagként a modellbe illesztve a függvény az abszcisszát metszi (Pachu és mtsai, 2018). A Lactin-2-függvény zérushelye ($T_{\min}$) numerikusan becsülhető, maximuma egyszerű deriválással határozható meg:

$$max\big(r(T)\big) = \frac{\Delta T ln(\rho \Delta T)}{1 - \rho \Delta T} + T_{\mathrm{L}}.$$

A K konstanst a Laktin-2 modell által meghatározott $T_{\min}$ érték alapján becsültem a pete-, és lárvastádiumra, az első és a második báb stádiumokra, valamint egy teljes generációra vonatkozóan.

A Lactin-2 modellt összehasonlítottam az alábbi modellekkel:

(1) Brière-2-modell (Brière és mtsai, 1999):

$$r(T) = aT(T - T_0)(T_{\mathrm{L}} - T)^{1/m},$$

ahol a $(0 < a < 1, T_{\min} < T < T_{\max})$ empirikus paraméter, és amelynek segítségével a T_{opt} optimális hőmérséklet is meghatározható (Régnier és mtsai, 2022):

$$T_{\mathrm{opt}} = \frac{(2mT_{\max} + (m+1)T_{\min}) + \sqrt{4m^2T_{\max}^2 + (m+1)^2T_{\min}^2 - 4m^2T_{\min}T_{\max}}}{4m+2},$$

ahol $T_{\min}$ és $T_{\max}$ az alsó és a felső hőmérsékleti küszöbérték, amelynél a fejlődés megkezdődik, illetve megszűnik.

(2) Brière-1-modell, a Brière-2-modell rögzített $m = 2$ esetben (Brière és mtsai, 1998, 29.(b) Ábra)

(3) Bieri-modell (Bieri és mtsai, 1983, 29.(a) ábra):

$$r(T) = \alpha(T - T_{\min}) - \beta exp(T - T_{\max}),$$

ahol α empirikus paraméter, $T_{\min}$ és $T_{\max}$ mint előbb $(0 < \alpha < 1,\ T_{\min} < T < T_{\max})$, és amely alapján a T_{opt} optimális hőmérséklet is meghatározható (Kondakis és mtsai, 2022):

$$T_{\mathrm{opt}} = T_{\max} + \frac{log(\alpha) - log\,(log(\beta))}{log\,(\beta)}.$$

(4) Analytis-modell (Analytis, 1981, 29.(c) ábra):

$$r(T) = \alpha(T - T_{\min})^m(T_{\max} - T)^n,$$

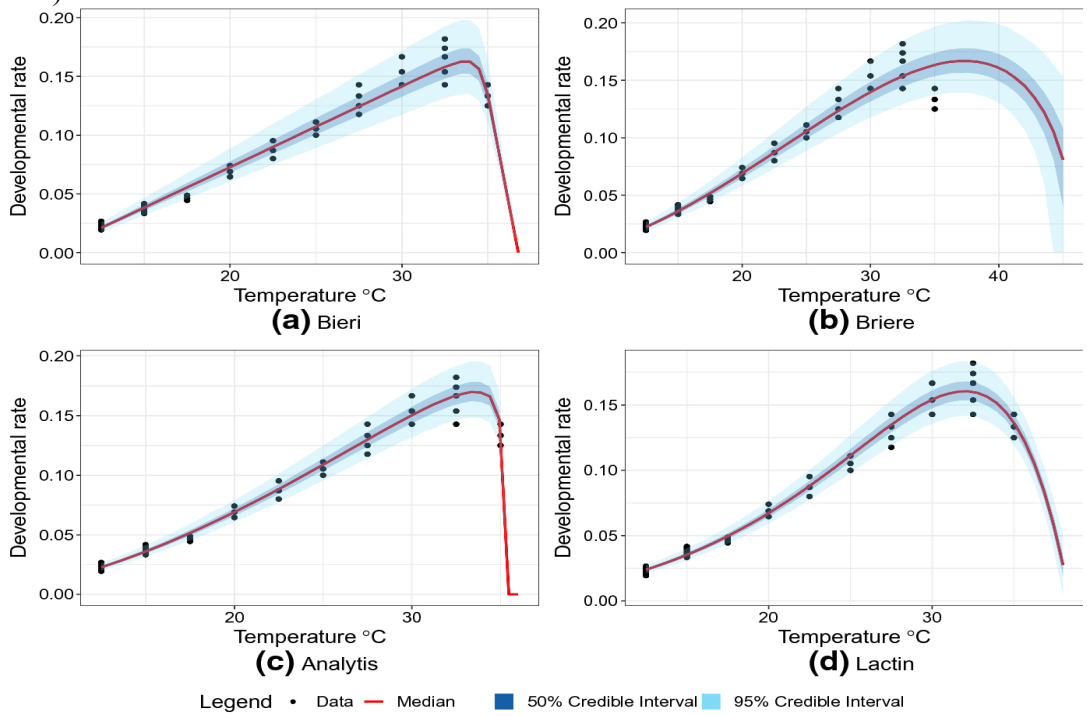
ahol α, m és n empirikus paraméterek, $T_{\min}$ és $T_{\max}$ mint előbb $(0 < \alpha < 1,$ $T_{\min} < T < T_{\max})$, és amely alapján a T_{opt} optimális hőmérséklet:

$$T_{\mathrm{opt}} = \frac{nT_{\max} + mT_{\min}}{n+m}.$$

A modellek kritikai összehasonlításáról szóló források megerősítik, hogy a szakirodalomban szereplő számos más modellből ez az öt modell (Lactin-2, Brière-1, Brière-2, Bierei, Analytis) a legígéretesebb (Roy és mtsai, 2002; Lardeux és mtsai, 2008; Jalai, 2010; Shi, 2011a,b; Moallem és mtsai, 2017; Rebaudo és Rabhi, 2018; Santis, 2021).

A modellekben szereplő paramétereket iterációval becsültem, miközben minimalizáltam a hibatagok átlagos négyzetösszegének négyzetgyökerét (RMSE). Értékeltem az R^2 értékeket, a modell paramétereit, valamint a modellek és paramétereik szignifikanciáját.

A legjobb modelleket (Lactin-2, Brière-1, Bieri) az Akaike (AIC, Akaike, 1973) és a Bayes-i információs kritérium (BIC, Schwarz, 1978) alapján választottam ki (ld. 2.4. fejezet).



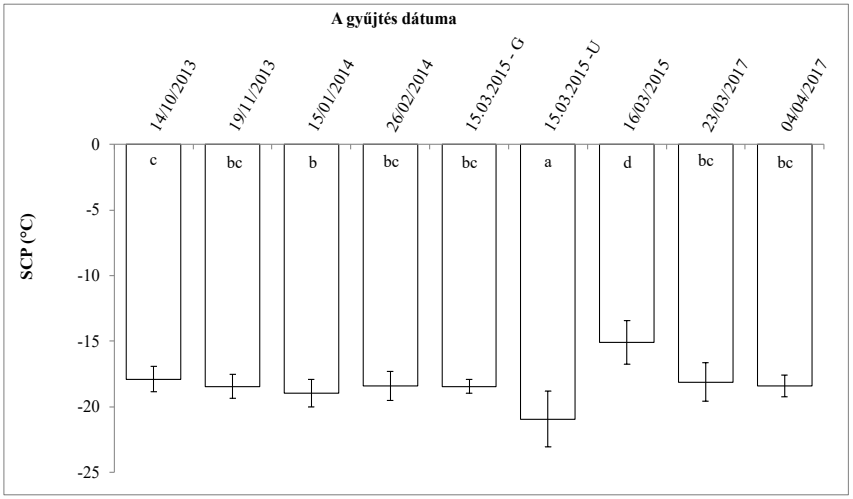
29. Ábra Négy hőmérséklettől függő fejlődési modell sematikus ábrája: (a) Briéri (Briéri és mtsai, 1983); (b) Briére (Briére és mtsai, 1998, 1999); (c) Analytis (Analytis, 1981); (d) Lactin (Lactin és mtsai, 1995) (Forrás: Kondakis és mtsai, 2022)

6.5.4 Eredmények

A kanyargós szillevéldarázs telelő bábjai a hideg hőmérsékletre kiterjedt szuperhűléssel reagáltak. A különböző megfigyelések között szignifikáns különbségek voltak az SCP-értékek között (Brown-Forsythe: $F(8; 54{,}33) = 25{,}51;\ \mathrm{p} < 0{,}001$). A Games-Howell post-hoc teszt eredményei a 30. ábrán láthatók.

Az SCP-méréseket követő napon egyik példány sem maradt életben, ami igazolja, hogy a kanyargós szillevéldarázs fagyintoleráns.

A szillevéldarazsak bábjainak hétféle beállításban végzett hidegkezelése után egy, illetve három nappal túlélési arányuk nem különbözött szignifikánsan egymástól (Marascuilo's $\chi 2(6) < 4{,}48;\ p > 0{,}61$).



30. Ábra A kanyargós szillevéldarázs telelő bábjainak 2013 és 2017 között, különböző helyeken (G - Gonars, U - Udine, Olaszország; illetve Kecskemét) gyűjtött átlagos szuperhűlési pontjai és szórásai. A különböző betűk szignifikánsan különböző csoportokat jelölnek (Games-Howell, $p < 0{,}05$).

A kanyargós szillevéldarázs petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődését leíró, az Akaike (AIC, Akaike, 1973) és a Bayes-i információs kritérium (BIC, Schwarz, 1978) alapján a legjobb modellek a Lactin-2, a Brière-1, valamint a Bieri-modell volt. A Brière-2 modell m együtthatójának becslésével nem kaptam jobb modellt, mint ha azt $m = 2$-nek rögzítettem (Brière-1). Az Analytis-modell nemcsak az információs kritériumok alapján bizonyult gyengébbnek, de becslései sok esetben nem voltak szignifikánsak, így e két modell eredményeit nem közlöm.

A kanyargós szillevéldarázs petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődését leíró legjobb modellek együtthatóinak becsült értékei a standard hibáikkal, a becsült paraméterekből származtatott értékek, valamint a modellek értékelése és az összehasonlításukhoz szükséges AIC és BIC értékek a 19. (Lactin-2), a 20. (Bière-1), illetve a 21. (Bieri) Táblázatban találhatóak.

19. Táblázat A kanyargós szillevéldarázs petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődését leíró Lactin-2-modellegyütthatók becsült értékei (zárójelben a standard hibái): T_{L} (°C) az a hőmérséklet, amelyen az életfolyamatok már nem tarthatók fenn hosszabb ideig, ρ az optimális hőmérsékleten bekövetkező növekedési sebesség, λ empirikus paraméter; a modellek értékelése: magyarázott variancia (R^2), illetve a modellre vonatkozó ANOVA F értéke; valamint a származtatott paraméterek: T_{min} (°C) és T_{max} (°C) alsó és a felső hőmérsékleti küszöbértékek, amelynél a fejlődés megkezdődik, illetve megszűnik, T_{opt} (°C) az optimális hőmérséklet, K pedig a fejlődés befejezéséhez szükséges alsó küszöbérték feletti foknapok száma.

	Pete (N=584)	Lárva (N=324)	Báb 1 (N=315)	Báb 2 (N=306)	Egy nemzedék (N=306)
Becsült modellbeli paraméterek					
ρ	0,010*** (<0,001)	0,004* (<0,001)	0,018*** (0,417)	0,014*** (0,008)	0,002*** (<0,001)
T_{L} (°C)	28,8*** (<0,001)	29,0* (<0,001)	27,7*** (0,417)	27,1*** (0,008)	27,3*** (<0,001)
ΔT (°C)	0,032*** (<0,001)	0,003* (<0,001)	0,627^{+} (0,361)	0,087*** (<0,001)	0,117*** (<0,001)
λ	-1,094*** (0,004)	-1,026* (0,003)	-1,098*** (0,023)	-1,122*** (0,010)	-1,016*** (0,001)
A modell értékelése					
R^2	0,86***	0,71***	0,60***	0,77***	0,90***
F	9228,1***	3227,6***	1895,4***	3075,2***	10380,0***
AIC	23,45	26,96	18,72	21,62	31,87
BIC	215,80	221,07	164,41	183,14	252,07
Származtatott paraméterek					
T_{min}[a] (°C)	9,0	6,6	5,1	8,1	7,1
T_{max}[b] (°C)	28,8	29,0	27,0	27,0	27,0
T_{opt}[b] (°C)	28,6	28,9	24,9	26,5	26,4
K[b]	84,0	245,9	42,9	57,2	432,7

szignifikancia szint ***$p < 0{,}001$; **$p < 0{,}01$; *$p < 0{,}05$; + $p < 0{,}1$; ns: nem szignifikáns
a. A T_{min} és K paramétereinek standard hibái mind 0,001 alatt voltak (*bootstrapping*, 1000, ld. 5.3. fejezet)
b. A magas hőmérsékleten mért kis mintaelemszám miatt nem megbízható, újabb kísérletekkel javítandó becslések.

A Lactin-2-modell nagy előnye volt, hogy paraméterbecslései (egy kivétellel) igen erősen szignifikánsak voltak ($p < 0{,}001$), ám a petefázis kivételével valamivel nagyobbak voltak az AIC és BIC értékeik, mint az eggyel kevesebb paraméterbecslést igénylő Brière-1 modellnek, bár ezek a különbségek csekélyek voltak. Ugyanakkor a magyarázott varianciahányaduk ugyanezekben a fázisokban magasabbak voltak a Brière-1-modell R^2 értékeinél. A Brière-1-modell a petefázis kivételével mindig a legkisebb AIC és BIC értékeket adta, de nagy hátránya, hogy a T_{min} értékek becslése nem olyan erősen szignifikáns, mint a Lactin-2-, vagy a Bieri-modellé, melyek becsléseihez viszonyítva jóval alacsonyabb közelítést ad. Tegyük hozzá, hogy míg a Lactin-2-modell a T_{min} értéket nem modellbeli paraméterként becsüli, hanem a modellbeli paraméterekből numerikusan állítható elő, a Bieri-féle becslés, amely hasonló T_{min} értékeket eredményezett, modellparaméter.

20. Táblázat A kanyargós szillevéldarázs petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődését leíró Brière-1-modellegyüttható becsült értékei (zárójelben a standard hibái): $T_{\min}$ (°C) és $T_{\max}$ (°C) alsó és a felső hőmérsékleti küszöbértékek, amelynél a fejlődés megkezdődik, illetve megszűnik; a empirikus paraméter; a modellek értékelése: magyarázott variancia (R^2), illetve a modellre vonatkozó ANOVA F értéke; valamint a származtatott paraméterek: T_{opt} (°C) az optimális hőmérséklet, K pedig a fejlődés befejezéséhez szükséges alsó küszöbérték feletti foknapok száma.

	Pete (N=584)	Lárva (N=324)	Báb 1 (N=315)	Báb 2 (N=306)	Egy nemzedék (N=306)
Becsült modellbeli paraméterek					
$T_{\min}$ (°C)	6,84*** (0,76)	4,00* (1,77)	4,00^{+}(2,29)	4,7* (2,20)	3,5* (1,56)
$T_{\max}$ (°C)	40,0*** (1,98)	37,1*** (3,55)	32,8*** (2,66)	40,0*** (7,32)	35,7*** (2,68)
$a * 10^5$	12*** (1,3)	4*** (0,9)	31*** (7,6)	15** (5,0)	2,3*** (0,42)
A modell értékelése					
R^2	0,88***	0.70***	0.52***	0,68***	0,80***
F	6848,0***	4237,5***	2090,8***	2273,0***	7000,0***
AIC	25,71	24,94	16,37	19,02	28,49
BIC	274,99	186,23	127,60	144,79	208,45
Származtatott paraméterek					
T_{opt} (°C)	32,76	30,12	26,7	32,5	28,9
K	97,0	304,0	46,2	74,0	561,6

szignifikancia szint ***$p < 0{,}001$; **$p < 0{,}01$; *$p < 0{,}05$; + $p < 0{,}1$; ns: nem szignifikáns

A $T_{\max}$ értékek becslése a kísérletben a magas hőmérsékleten mért kis mintaelemszám miatt egyik modell esetében sem tekinthető megbízhatónak. A Lactin-2 modell ezeket az értékeket jóval alacsonyabbra becsüli, mint a másik két modell. Ezt a paramétert újabb kísérletekkel javítani szükséges.

A Bieri-modell paraméterbecslései igen erősen szignifikánsak, és a petefázis esetében ez a modell eredményezte a legalacsonyabb AIC és BIC értékeket igen magas magyarázott varianciahányadokkal. E modell hátránya, hogy két empirikus paramétere is van, amelyeknek az interpretációja nehézkes.

21. Táblázat A kanyargós szillevéldarázs petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődését leíró Bieri-modellegyüttható becsült értékei (zárójelben a standard hibái): $T_{\min}$ (°C) és $T_{\max}$ (°C) alsó és a felső hőmérsékleti küszöbértékek, amelynél a fejlődés megkezdődik, illetve megszűnik; a, b empirikus paraméterek; a modellek értékelése: magyarázott variancia (R^2), illetve a modellre vonatkozó ANOVA F értéke; valamint a származtatott paraméterek: T_{opt} (°C) az optimális hőmérséklet, K pedig a fejlődés befejezéséhez szükséges alsó küszöbérték feletti foknapok száma.

	Pete (N=584)	Lárva (N=324)	Báb 1 (N=315)	Báb 2 (N=306)	Egy nemzedék (N=306)
Becsült modellbeli paraméterek					
$T_{\min}$ (°C)	9,2*** (<0,001)	4,0*** (1,77)	6,4***(0,66)	8,7*** (0,38)	7,3*** (0,26)
$T_{\max}$ (°C)	40,0*** (<0,001)	37,1*** (3,55)	32,8*** (2,66)	27,0*** (<0,001)	27,0***(<0,001)
a	0,013*** (<0,001)	0,004*** (<0,001)	0,026*** (0,001)	0,019** (0,001)	0,002***(<0,001)
b	6,13***(<0,001)	6,13***(<0,001)	6,13***(<0,001)	6,13***(<0,001)	6,13***(<0,001)
A modell értékelése					
R^2	0,85***	0.71***	0.71***	0,76***	0,90***
F	8660,2***	3216,1***	3220,0***	3436,0***	10298,5***
AIC	23,33	26,94	26,94	21,54	31,88
BIC	214,89	220,91	220,91	182,62	252,11
Származtatott paraméterek					
T_{opt} (°C)	39,00	35,80	26,1	26,1	25,7
K	83,1	264,0	44,0	64,8	428,1

szignifikancia szint ***$p < 0{,}001$; **$p < 0{,}01$; *$p < 0{,}05$; + $p < 0{,}1$; ns: nem szignifikáns

Mivel az optimális hőmérsékletet minden modell esetén származtatott, a $T_{\min}$ -től függő érték, érthető, hogy ezek is erősen különböznek az egyes modellek esetén. A Brière-1- és Bieri-modellek becslései túl magasnak tűnnek, a Lactin-2-modell becslése sokkal hihetőbb. A fejlődés befejezéséhez szükséges alsó küszöbérték feletti foknapok K számát pedig közvetlenül a nagyon eltérő $T_{\min}$ -ből kapjuk, ezért ez a paraméter nem alkalmas a modellek összehasonlítására.

6.5.5 Következtetés

A különböző fagypont alatti hőmérsékleteknek (-1,6 °C és -4,0 °C -on 10, 20 és 30 napig, valamint -10,5 °C-on 9 napig) kitett, telelő egyedek túlélési aránya 89,2% és 100% között mozgott, ami arra utal, hogy az *A. leucopoda* nem fagyérzékeny faj. Eredményeink arra utalnak, hogy az alacsony téli hőmérséklet várhatóan nem lesz fontos korlátozó tényező az *A. leucopoda* telelési sikeressége szempontjából. A diapauza időszakát is figyelembe véve a becslések szerint az *A. leucopoda* potenciálisan akár négy-öt generáción keresztül is fejlődhet évente Magyarországon.

A kanyargós szillevéldarázs petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődését leíró legjobb modellek eredményeit tekintve a legjobban teljesítő modellnek a Lactin-2-modellt választottuk, és ennek eredményeit ismertettük az idézett publikációban.

Bár a faj fejlődése a hőmérséklet növekedésével felgyorsult, feltételezzük, hogy az optimális hőmérséklet valószínűleg nem 24,9 °C és 28,9 °C körül van, ahogyan azt a Lactin-2-T_{opt} paraméterek becslései alapján megállapítottuk, és még inkább nem ennél magasabban. Ehelyett a túlélési és termékenységi arányok alapján a faj optimális hőmérsékleti tartománya inkább 15,0 °C és 19,5°C között van.

A három legjobb modell becsléseinek különbözősége rámutat a kísérletben szereplő hőmérsékleti beállítások módosításának szükségességére. A hat kísérletben használt hőmérsékleti érték (11 °C, 15 °C, 18,5 °C, 23 °C, 25 °C, 27 °C) helyett 10–12 értékre lenne szükség: legalább két 27 °C feletti, de ahhoz közeli, legalább két 18,5 °C alatti (pl. 10,5 °C és 5,0 °C), és lehetőleg a 18,5 °C és 27 °C közötti tartomány sűrűbb felosztására lenne szükség, hogy a modellektől megbízhatóbb becsléseket várhassunk. Az is fontos lenne, hogy a magas hőmérsékleti tartományokon megfelelően magas egyedszámmal is rendelkezzünk, ami a mi esetünkben a 27 °C-on nem valósult meg.

Eredményeink alapvető ismereteket adnak hozzá a kanyargós szillevéldarázs élettörténetéhez, melyek mind az alap-, mind az alkalmazott tudományok számára fontosak. Ezek az eredmények hozzájárulhatnak egy Európában invazív idegen faj biológiájának jobb megértéséhez.

6.6 A származás hatása a bor beltartalmára gépi tanulási módszerek alapján

E fejezet az alábbi publikáción alapszik:

Nyitrainé Á. D., **Ladányi M.**, Varga Z., Szövényi Á.P., Matolcsi R. 2022. 'The Effect of Grapevine Variety and Wine Region on the Primer Parameters of Wine Based on ^{1}H NMR-Spectroscopy and Machine Learning Methods'. Diversity. 14(2):74. https://doi.org/10.3390/d14020074 **Q1**

6.6.1 Motiváció és célkitűzés

A nukleáris mágneses rezonancia (NMR) spektroszkópia az 1940-es évek óta alkalmazott szerkezetvizsgálati módszer. Az NMR-technika számos előnnyel jár, többek között a minta egyszerű előkészítésével és rövid futási idejével (Simmler és mtsai, 2014). Egyszerű és gyors módszer a kémiai összetételek meghatározására, továbbá alkalmas a szerves vegyületek (metabolitprofilok), például aminosavak (Consonni és Cagliani, 2019), szerves savak, alkoholok (Du és mtsai, 2007, Consonni és mtsai, 2017), cukrok és fenolos vegyületek azonosítására is (Anastasiadi és mtsai, 2009). Az NMR-spektroszkópia így az utóbbi években egyre elterjedtebb alkalmazássá vált az élettudományok egyes területein, beleértve a termékek minőségellenőrzését (Minoja és Napoli, 2014), a hitelesítést (Consonni és Cagliani, 2010), a legkülönfélébb élelmiszerek elemzését (Spyros és Dais, 2012, Spyros, 2016, Consonni és mtsai, 2016, Lovatti és mtsai, 2020), de a biológiai tudományokban (Wenck és mtsai, 2023) és humán orvoslástudományban is (Zhao és mtsai, 2019).

A legtöbb nukleáris mágneses rezonancia-spektroszkópia (NMR) alapú vizsgálat proton ^{1}H NMR spektroszkópiát használ, mivel az ^{1}H-atomok szinte minden szerves vegyületben és ismert metabolitban előfordulnak (Abdul-Hamid és mtsai, 2019). Az egydimenziós ^{1}H NMR-spektrumok különösen hasznosak a metabolizmusvizsgálatokban, mivel a technika jól automatizálható, rendkívül megbízható és nagyon gyors.

Minden szőlőfajta egyedi szerkezeti képlettel rendelkezik, amely sajátosság megjelenik a végtermék (bor) összetételében. Egy borminta teljes NMR-spektruma a bor molekuláris ujjlenyomatának tekinthető, mivel értékeit számos borkészítési tényező módosítja, például a szőlőművelési technika, a talaj klimatikus viszonyai, a szőlőfajta, az erjesztési technika, illetve a földrajzi eredet (Godelmann és mtsai, 2013, Mazzei és mtsai, 2010, Monkahova és mtsai, 2011, McClure, 2017., Liu és mtsai, 1996, Alsante és mtsai, 2011). Ezért e módszer közvetlenül felhasználható a borok összehasonlítására és azonosítására, így az a borok elemzésének innovatív modern módszerévé vált.

Tanulmányunkban négy magyar fajtából (Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir) készült borok ^{1}H NMR-spektroszkópiával nyert adatainak vizsgálatát tűztük ki célul 861 bormintáját figyelembe véve, amelyek két borvidékről (Villány, Eger) származtak 2015-ből és 2016-ból.

A fajták és régiók modellalapú osztályozását kívántuk előállítani a borok 22 tulajdonságának (alkoholok, cukrok, savak, bomlástermékek, biogén aminok, polifenolok, erjedési vegyületek stb.) ismeretében, valamint meghatározni az osztályozásban szereplő legfontosabb paramétereket.

6.6.2 Az adatok alapjául szolgáló mérések

A ^{1}H NMR-spektroszkópiamódszerrel 53 borkomponens egyidejűleg és gyorsan mérhető, és olyan komplex adathalmaz jön létre, amely alkalmas a bor földrajzi eredetének és/vagy fajtájának meghatározása céljából végzett többváltozós statisztikai elemzésre

(Lindon és mtsai, 2000). A vizsgált 53 komponens közül a borelemzés legfontosabb nyolc paramétere a következő: alkohol, glükóz, fruktóz, glicerin, borkősav, tejsav, valamint az összes és a cukormentes extrakt. Ezeken az alapparamétereken kívül 7 bomlástermék (ecetsav, acetoin, biogén amin (etil-acetát) és polifenolok (kaftársav, galluszsav, szikimisav és trigonellin)), továbbá az erjedésmarker 7 vegyülete (2,3-butándiol, 2-feniletanol, 3-metil-butanol, acetaldehid, galakturonsav, metanol és borostyánkősav) szerepelt az adathalmazban, azaz 22 ismérvvel dolgoztunk. A rendelkezésre álló adatok régió, évjárat és fajta szerinti mintaelemszáma a 22. Táblázatban található.

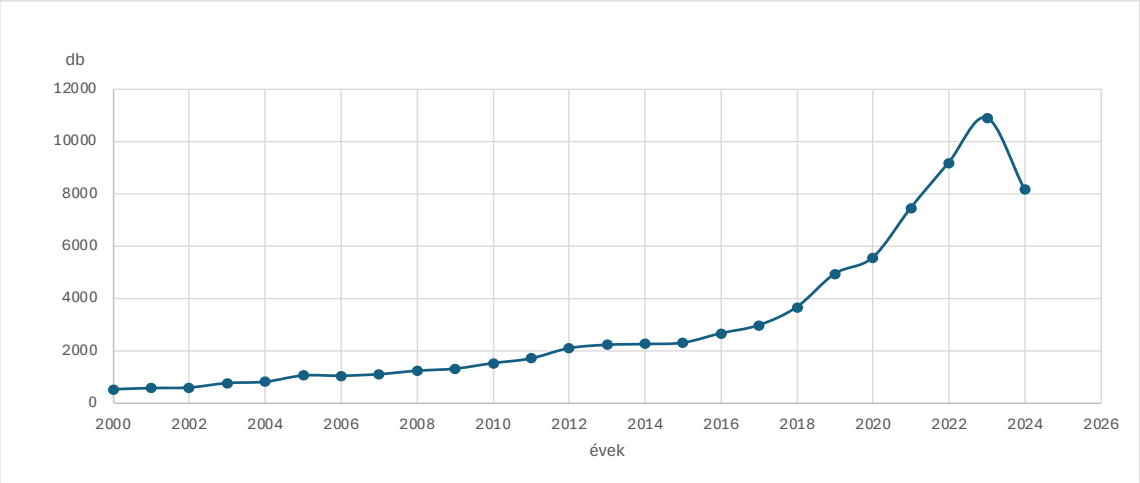
22. Táblázat A vizsgálatba bevont Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir fajták két magyarországi borvidékekről, Villányból és Egerből, valamint a 2015-ös és 2016-os évjáratokból származó mintáinak száma. A statisztikai mintaelemszám 861 volt.

Régió	Évjárat	Cabernet Sauvignon	Kékfrankos	Merlot	Pinot Noir	Összesen
Villány	2015	65	65	127	56	313
	2016	61	66	9	60	196
	Összes	126	131	136	116	509
Eger	2015	104	9	34	35	182
	2016	44	17	57	52	170
	Összes	148	26	91	87	352

6.6.3 Módszer

A közelmúltban Amargianitaki és mtsai (2017), Kalogiouri és Samanidou (2021), valamint Masetti és mtsai (2021) többféle többváltozós elemzési módszert (PCA, ld. 5.2.2. fejezet), parciális legkisebb négyzetek módszere (PLS, ld. 5.2.3. fejezet), diszkriminanciaelemzés (DA, ld. 5.1.4. fejezet), *k*-legközelebbi szomszéd módszer (KNN, ld. 5.1.1. fejezet), mesterséges neurális hálózatok (ANN, ld. 5.3.1. fejezet) és támogatóvektor-gépek (SVM, ld. 5.1.7. fejezet) mutattak be, illetve hasonlítottak össze szakirodalmi áttekintő munkáikban, hogy vizsgálják azok osztályozási sikerét a fajta, az évjárat, a földrajzi eredet, illetve a szezonalitás függvényében, NMR-adatok figyelembevételével; az utóbbi szerzőcsoport nem csupán bormintákra fókuszálva.

A hagyományos, vagy hagyományosan alkalmazott kemometriai technikák (PCA. Ld. 5.2.2. fejezet; DA, ld. 5.1.4. fejezet; PLS, ld. 5.2.3. fejezet; KNN, ld. 5.1.1. fejezet) azonban nem teljesítenek jól nagyszámú változó és ahhoz képest kevés mintaszámot tartalmazó adatok osztályozásában, különösen magas osztályszámok esetén, illetve ha a feltárni kívánt összefüggések nem lineárisak. Ezzel szemben az olyan gépi tanulási módszerek, mint a véletlen erdő (RF, ld. 5.3.2. fejezet), illetve a támogatóvektor-gépek (SVM, ld. 5.1.7. fejezet) különösen is hatékony módszerek ilyen esetben is (Breiman, 2001, Mascellani és mtsai, 2021, Martelo-Vidal és Vázquez, 2016). Nem véletlen, hogy az NMR-adatokon alapuló gépi tanulási osztályozási módszerek használata az utóbbi években igen elterjedt, és számos élettudományi alkalmazási területen találkozunk velük az agrártudományokon túl a humán orvostudományt is beleértve (31. Ábra).



31. Ábra A Google-tudós „NMR, machine learning, classification” keresésre adott találatok száma 2000–2004 augusztus 31-ig. Az itt bemutatott publikációnk 2021-ben került leadásra és 2022-es megjelenésű.

Mivel a 22 ismérvet tartalmazó adathalmazról rendelkeztünk korábbi osztályozási ismeretekkel (származás és fajta), ezért felügyelt gépi tanulási osztályozási módszereket teszteltem a négy fajta és két származás (összesen nyolc csoport) elkülönítésére: lineáris diszkriminanciaelemzés (LDA, ld. 5.1.4. fejezet), mesterséges neurális hálózatok (ANN, ld. 5.3.1. fejezet), támogatóvektor-gépek (SVM, ld. 5.1.7. fejezet) és véletlen erdő (RF, ld. 5.3.2. fejezet).

Adat-előkészítésként standardizáltam az elemzésbe bevont 22 ismérvet. A többváltozós kiugró értékeket a Mahalanobis-távolsággal szűrtem (Mahalanobis, 1936, Li és mtsai, 2019, ld. 5. Példa).

Az adathalmazunkat véletlenszerűen két részre, a tanuló és a tesztelő adathalmazra osztottam 3:1 arányban. A nyolc csoport igen eltérő mintaelemszámmal rendelkezett (22. Táblázat), ezért a két részhalmazt úgy hoztam létre, hogy azok megtartsák az eredeti mintaméret-arányokat (*stratified sampling*, ld. 4.4. fejezet). Így elkerülhettem, hogy olyan tanuló/tesztelő halmazaink legyenek, amelyek nem (vagy túl kevés) mintaelemet tartalmaznak a legkisebb csoportokból (Merlot Villányból, 2016 vagy Kékfrankos Egerből, 2015). Mind a négy módszer esetében 1000 futtatást végeztem különböző véletlenszerű felosztással. A tanuló adathalmaz által előállított osztályozást a fennmaradó tesztadathalmazon teszteltem.

A mesterséges neurális hálózatok (ANN)-modellünk hiperparamétereinek hangolásához 100 iterációval, 25 ismétlésből álló bootstrap mintavételezést (ld. 5.3. fejezet) végeztem. Szigmoid aktiváló függvényt használtam, a rejtett rétegben a csomópontok száma 5 volt, a túlillesztés elkerülésére szolgáló regularizációs paraméter pedig 0,1.

Az RF hiperparamétereit hangolással optimalizáltam rácskereséssel (*tuning*, *grid search,* ld. 5.3.2. fejezet) az OOB hiba becslése, a pontosság (*accuracy*), az F1 érték és az AUC figyelembevételével (*ntree* = 500 és *mtry* = 4, ld. 4.6. fejezet).

A 6.2. fejezetben már említettem, hogy ha a magyarázó változókban a faktortípusú magyarázó változók szintjeinek száma eltér (ami esetünkben fennállt), akkor a pusztán a Gini-index (ld. 5.1.3. fejezet) alapján végzett felosztási szabály (*splitting rule*) torzítást eredményezhet (Breiman, 1984, Strobl és mtsai, 2005), ezért a pontosság (*accuracy*) átlagos csökkentésének és a Gini-index átlagos csökkenésének elvét egyszerre figyelembe

vevő módosított (kombinatorikus) kiválasztási szabályt alkalmaztam (Strobl és mtsai, 2006).

A változók fontosságát a pontosság átlagos csökkenése (MDA, ld. 5.3.2. fejezet és 6.2.3. fejezet) szerint a teljes modellre és a kategóriák szerint is kiszámoltam. A pontosság kiszámításához az OOB mintakészletet használtam.

Az SVM modell esetében a γ-paraméterű sugáralapú bázisfüggvényt választottam (ld. 5.1.7. fejezet). A paraméterek optimalizálását ismét egy hangolási folyamatnak vetettem alá, amelynek eredményeként a γ = 0,05, a költségérték pedig C=10 lett. A változók fontosságát a ROC-görbe AUC értéke alapján számoltam ki (ld. 4.6. fejezet).

A fajták és borvidékek közötti különbségek vizualizálására pókhálóábrát készítettem. Ehhez az előzőekben kapott eredmények alapján a legnagyobb osztályozó erővel rendelkező változókat használtam azok $(X - min(X))/(max(X) - min(X))$ transzformációit követően (ld. 6. Példa). Ezzel a transzformációval a legmagasabb értékek 1, a legalacsonyabbak pedig nulla értéket vettek fel, azaz az összes komponens összehasonlíthatóvá vált azok nagyságrendjétől függetlenül. A pókháló ábrán felismerhetjük, hogy melyik fajta/borvidék eredményezte a leginkább megkülönböztető komponensek magas vagy alacsony értékeit.

6.6.4 Eredmények

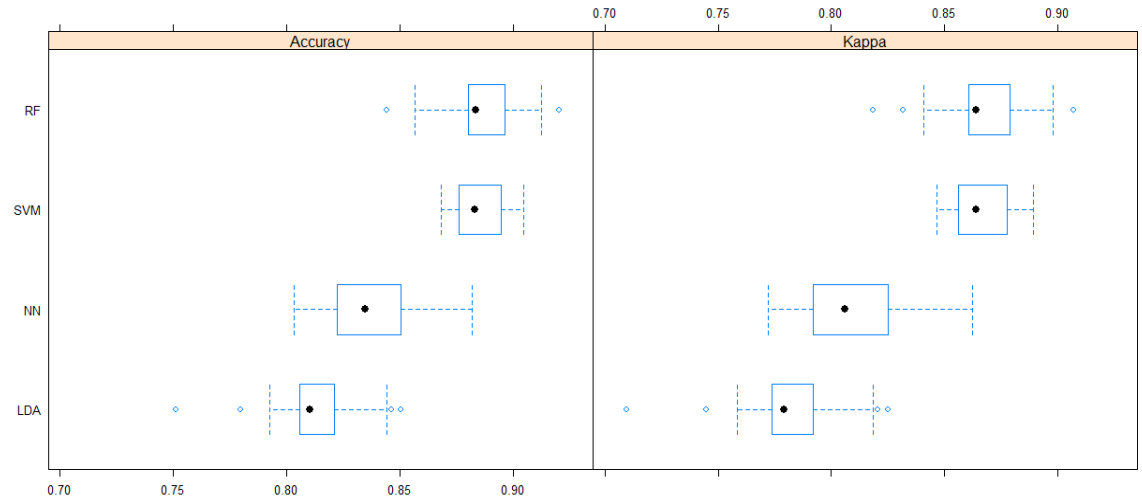
A tesztelt négy osztályozási módszer közül (LDA, ANN, SVM és RF) pontosságuk alapján az utóbbi kettő volt a legsikeresebb. Mindegyiknél 1000 futtatás alapján mutatom be az osztályozási eredményeket, melyek megerősítik, hogy az SVM- vagy az RF-módszerek alkalmazása sikeres megközelítés lehet a borok NMR-adatokra épülő azonosítására.

A modellek klasszifikációs teljesítményét a 4.6. fejezetben ismertetett mutatók szerint értékeltem.

- Lineáris diszkriminanciaelemzés és mesterséges neurális hálózatok

Az LDA meglehetősen jól teljesített: az első három diszkriminanciafüggvény (DV_1, DV_2, DV_3) a magyarázó változók megkülönböztető képességének rendre a 43,2, 24,2 és 17,7%-át tették ki, a kanonikus korrelációk mindegyike 0,76 fölött volt, és mind a három diszkriminanciafüggvény Wilk-féle lambda-értéke szignifikáns volt. A helyes besorolás aránya 84,9%-os volt, keresztvalidációval is 83,7% (ld. 2.2. fejezet). Mégis, a többi módszerrel összehasonlítva az LDA volt a leggyengébb modell a nyolc csoport (négy fajta és két borvidék) elkülönítésében. Ez érthető, hiszen ne feledjük, hogy az LDA meglehetősen érzékeny a kiugró értékekre és az egyenlőtlen mintanagyságra, ami a mi esetünkben fennállt. Emellett megköveteli a többváltozós normalitást, illetve a szóráshomogenitást, amely feltételek a mi esetünkben enyhén sérültek, bár ezekkel a feltételsérülésekkel szemben meglehetősen robusztus, ha az összes többi feltétel többé-kevésbé igaz. A multikollinearitás szintén torzított eredményeket okozhat az erősen redundáns információt hordozó tulajdonságok miatt (Rao, 1948, Nisbet és mtsai, 2018).

Az ANN modellek általában valamivel jobban működtek, mint az LDA modellek, bár jelentősen rosszabbul, mint az RF vagy SVM modellek (32. ábra). Figyelembe véve azt is, hogy az ANN modellek a legkevésbé interpretálhatóak (Molnar, 2022), az egyes módszerek 1000 futtatása, valamint a modell pontossága és a Cohen-féle Kappa alapján az RF és SVM modelleket választottam ki, hogy részletesen értékeljem teljesítményüket.



32. Ábra A lineáris diszkriminanciaelemzés (LDA), a neurális hálózatok (NN), a támogatóvektor-gépek (SVM) és a véletlen erdő (RF) modellek osztályozási teljesítményének összehasonlítása, a pontosság (*accuracy*) és a Cohen-féle Kappa-értékek alapján, az egyes módszerek 1000 futtatása alapján. A cél a fajták (Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir) és a borvidékek (Eger, Villány) elkülönítése volt, figyelembe véve 22 tulajdonságukat, az alkoholok, cukrok, savak, bomlástermékek, biogén aminok, polifenolok és erjedési vegyületek tekintetében, 861 mintaelemszámú 2015-ös és 2016-os bormintákból álló adatsor alapján. A fekete pontok a mediánértékeket jelölik.

- Véletlen erdő (RF) és támogatóvektor-gépek (SVM)

A tesztelő és a tanuló adathalmazra számított pontossági arány az RF és az SVM esetében 0,99, illetve 0,94 volt, ami megerősítette, hogy a túlillesztés nem torzította az eredményeinket (23. és 24. Táblázat).

A tesztelő adathalmazra alkalmazott RF- és SVM-modellek teljes pontossága 0,95, illetve 0,94 volt. Az érzékenység, a pontosság, a kiegyensúlyozott pontosság és az AUC értékek alapján mind az RF-, mind az SVM-modellek a legsikeresebben az egri és a villányi Pinot Noir, valamint a villányi Kékfrankos kategóriákat jósolták meg. Ezen csoportok érzékenységi értékei az RF esetében 100%-os, az SVM esetében pedig 91%-nál magasabbak voltak. Az RF és az SVM esetében mindhárom csoport meghaladta a 0,92, illetve 0,87 pontosságot, a 0,98, illetve 0,94 kiegyensúlyozott pontosságot, a 0,99, illetve 0,98 AUC-értéket, valamint a 0,96, illetve 0,89 F1-pontszámot.

Az egri Kékfrankos előrejelzése volt a leggyengébb, ami nem meglepő, mivel ebben a csoportban kevés megfigyelésünk volt (9 és 17 a két évben). Ennek a kategóriának az érzékenysége az RF és az SVM esetében mindössze 0,50 és 0,67 volt, alacsony pontossággal (0,60 és 0,57) és F1-értékekkel (0,55 és 0,62). A kiegyensúlyozott pontosság (0,73 és 0,81) és az AUC (0,93 és 0,73) értékek szintén ennél a csoportnál voltak a legalacsonyabbak.

Ebből arra következtethetünk, hogy a leghatékonyabb osztályozási módszer az RF volt, amelyet az SVM szorosan követett, mindkettő igen szignifikáns súlyozott Kappa-értékkel (0,95; $p < 0{,}001$), de még az RF is gyengén teljesített, amikor a megfigyelések száma nagyon alacsony volt.

23. Táblázat A fajta és származás válaszváltozókra épített véletlen erdő (RF) modell osztályozási minősége (CE: Cabernet Sauvignon, Eger; CV: Cabernet Sauvignon, Villány; BE: Kékfrankos, Eger; BV: Kékfrankos, Villány; ME: Merlot, Eger; MV: Merlot, Villány; PE: Pinot Noir, Eger; PV: Pinot Noir, Villány; CI: 95%-os konfidenciaintervallum)

Pontosság 0,95 CI (0,91; 0,97)	Súlyozott Kappa 0,95 ***		OOB teljes hiba 12,14%			Pontosság arány (Teszt/Tanuló) 0,99		
	Csoportonkénti eredmények							
Kategóriák	CE	CV	BE	BV	ME	MV	PE	PV
OOB hiba	0,07	0,12	0,20	0,05	0,18	0,09	0,15	0,03
Pontosság	0,98	0,99	0,98	1,00	0,98	0,98	0,99	1,00
Szenzitivitás	0,95	0,97	0,50	1,00	0,91	0,91	1,00	1,00
Specificitás	0,99	0,97	0,97	0,96	0,97	0,99	0,99	0,99
Precizitás	0,95	0,94	0,60	1,00	0,91	0,97	0,92	1,00
Negatív predikciós érték	0,99	0,99	0,99	1,00	0,99	0,98	1,00	1,00
Prevalencia	0,17	0,16	0,03	0,15	0,11	0,16	0,10	0,13
Detekciós érték	0,16	0,15	0,01	0,15	0,10	0,14	0,10	0,13
Detekciós prevalencia	0,17	0,16	0,02	0,15	0,11	0,15	0,11	0,13
Kiegyensúlyozott pontosság	0,97	0,97	0,73	0,98	0,94	0,95	0,99	0,99
AUC	0,99	0,99	0,93	0,99	0,99	0,99	0,99	0,99
F1 érték	0,95	0,96	0,55	1,00	0,91	0,94	0,96	1,00

***: szignifikáns $p < 0{,}001$ szinten

24. Táblázat A fajta és származás válaszváltozóra épített támogatóvektor-gép (SVM) modell osztályozási minősége (CE: Cabernet Sauvignon, Eger; CV: Cabernet Sauvignon, Villány; BE: Kékfrankos, Eger; BV: Kékfrankos, Villány; ME: Merlot, Eger; MV: Merlot, Villány; PE: Pinot Noir, Eger; PV: Pinot Noir, Villány; CI: 95%-os konfidenciaintervallum)

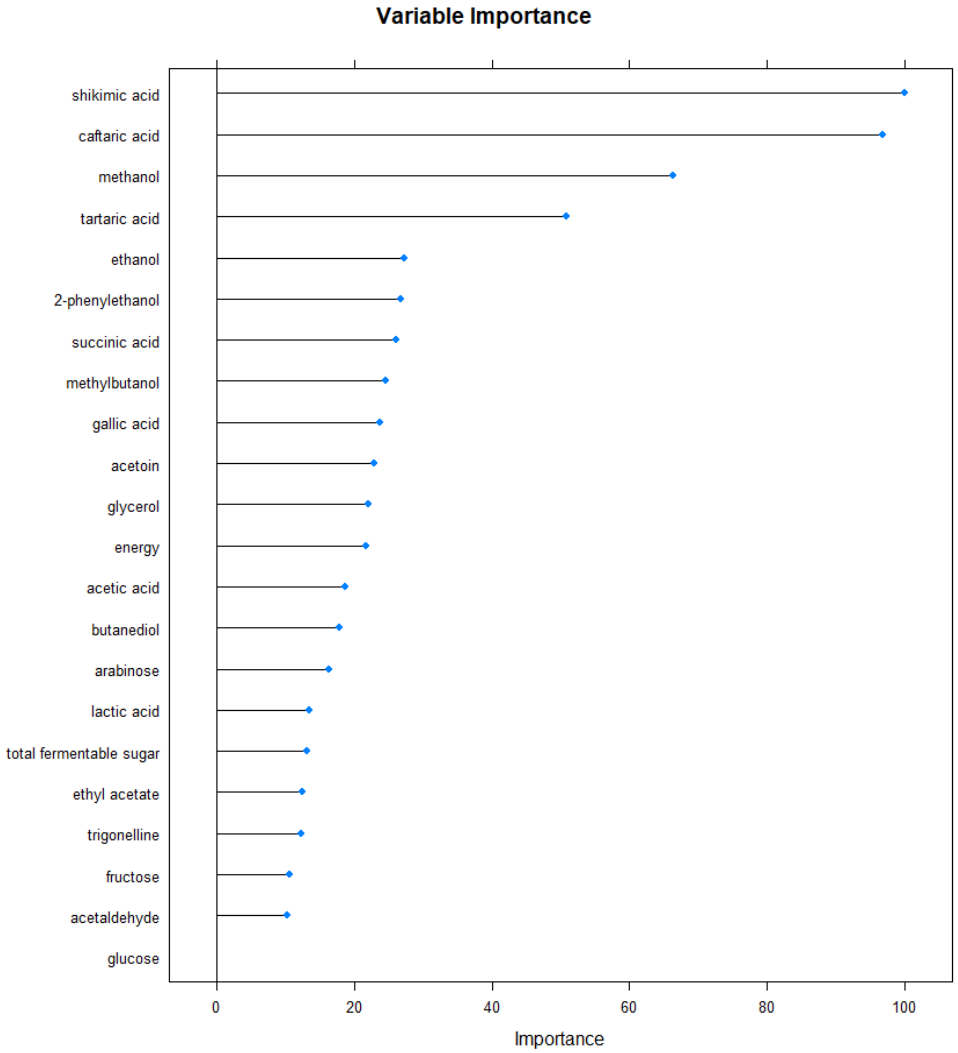
Pontosság 0,95 CI (0,91; 0,97)	Súlyozott Kappa 0,95 ***		OOB teljes hiba 12,14%			Pontosság arány (Teszt/Tanuló) 0,99		
	Csoportonkénti eredmények							
Kategóriák	CE	CV	BE	BV	ME	MV	PE	PV
OOB hiba	0,98	0,98	0,98	0,99	0,98	0,99	0,98	1,00
Pontosság	0,97	0,91	0,67	0,94	0,91	0,94	0,91	1,00
Szenzitivitás	0,98	0,97	0,96	0,96	0,96	0,97	0,98	0,98
Specificitás	0,92	0,97	0,57	1,00	0,91	0,97	0,87	0,97
Precizitás	0,99	0,98	0,99	0,99	0,99	0,99	0,99	1,00
Negatív predikciós érték	0,17	0,16	0,03	0,15	0,11	0,16	0,10	0,13
Prevalencia	0,17	0,14	0,02	0,14	0,10	0,15	0,09	0,13
Detekciós érték	0,18	0,15	0,03	0,14	0,11	0,15	0,11	0,14
Detekciós prevalencia	0,98	0,94	0,81	0,95	0,94	0,95	0,94	0,99
Kiegyensúlyozott pontosság	0,97	0,97	0,73	0,98	0,94	0,95	0,99	0,99
AUC	0,95	0,94	0,62	0,97	0,91	0,96	0,89	0,98
F1 érték	0,98	0,98	0,98	0,99	0,98	0,99	0,98	1,00

***: szignifikáns $p < 0{,}001$ szinten

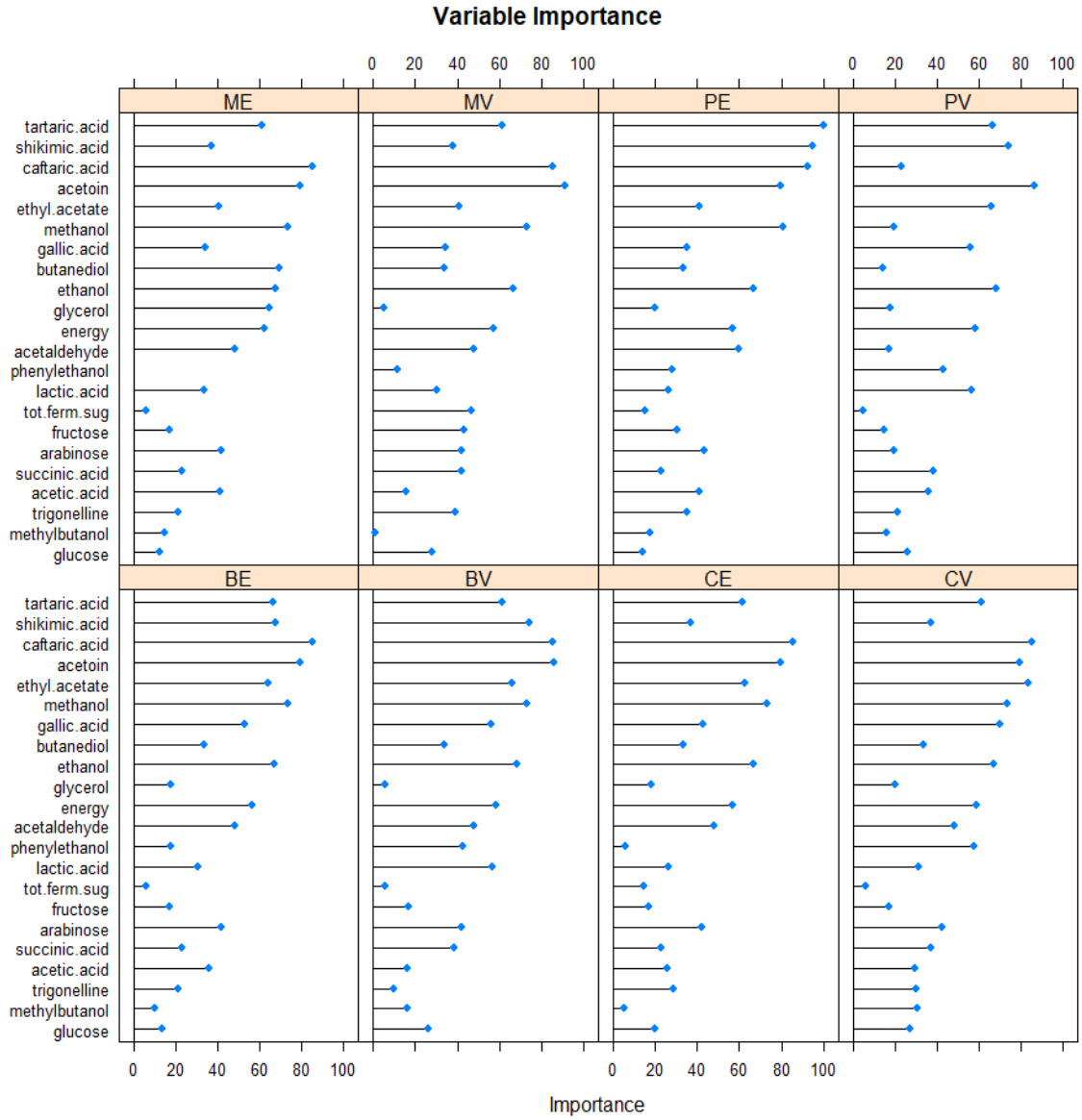
A változók fontosságát az SVM és RF modell kimenetei alapján számoltam ki a négy fajta (Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir), valamint a két borvidék (Eger, Villány) származási helyre vonatkozóan 2015-ös és 2016-os, összesen 861 elemű minta alapján 22 tulajdonságuk (alkoholok, cukrok, savak, bomlástermékek, biogén aminok, polifenolok és erjedési vegyületek) szerint. Az első hét legfontosabb változó mind az SVM, mind az RF modellekben megegyezett (33. Ábra).

A változó fontossági értékeit csoportonként is kiszámítottam (34. ábra).

Az RF- és az SVM-modellek megegyeztek abban, hogy a hét legfontosabb paraméter a szikimisav, a kaftársav, a metanol, a borkősav, az etanol, a 2-feniletanol és a borostyánkősav volt. Az etanol, a metanol és a borostyánkősav az alkoholos erjedésre utal, míg a kaftársav, a szikimisav és a borkősav a szőlőre jellemző összetevők. A 2-feniletanol paraméter az erjedés intenzitását és a szőlő állapotát is jelzi a szőlőhéj ízében.



33. Ábra A fajták (Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir) és borvidékek (Eger, Villány) osztályozása során a támogatóvektor-gépek (SVM) modellen alapuló változófontosságok, figyelembe véve a bormintákon mért 22 tulajdonságot (alkoholok, cukrok, savak, bomlástermékek, biogén aminok, polifenolok és erjedési vegyületek) 2015-ös és 2016-os, összesen 861 elemű minta alapján.

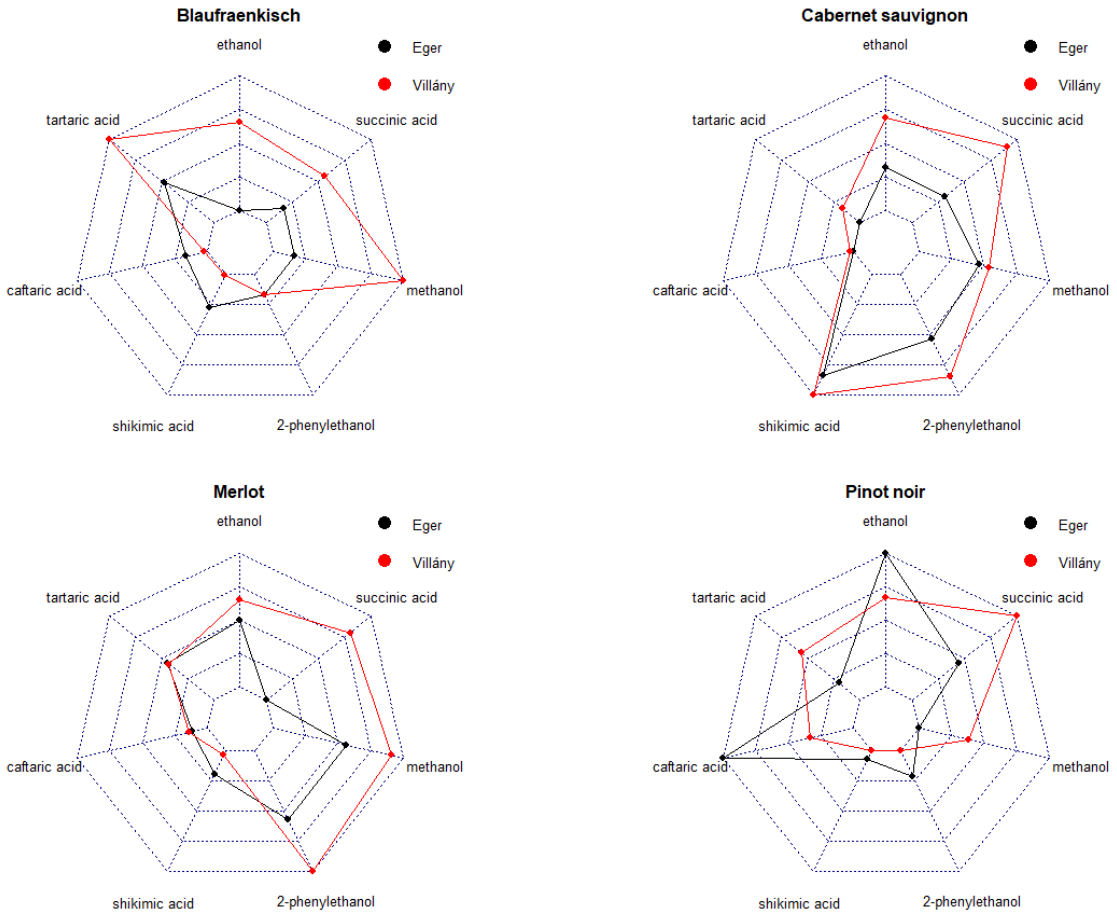


34. Ábra A fajták (Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir) és borvidékek (Eger, Villány) osztályozása a támogatóvektor-gépek (SVM) modellen alapuló változófontosságok az összes fajta/borvidék csoportra vonatkozóan, figyelembe véve a bormintákon mért 22 tulajdonságot (alkoholok, cukrok, savak, bomlástermékek, biogén aminok, polifenolok és erjedési vegyületek) 2015-ös és 2016-os, összesen 861 elemű minta alapján (CE: Cabernet Sauvignon, Eger; CV: Cabernet Sauvignon, Villány; BE: Kékfrankos, Eger; BV: Kékfrankos, Villány; ME: Merlot, Eger; MV: Merlot, Villány; PE: Pinot Noir, Eger; PV: Pinot Noir, Villány).

Végül a hét legfontosabb változó (szikimisav, kaftársav, metanol, borkősav, etanol, 2-feniletanol és szukcininsav) alapján szemléltettem a fajta/borvidék jellemzőit.

A 34. és a 35. Ábra szerint a kiválasztott hét paraméter közül a magas szikimisav és 2-feniletanol a Cabernet Sauvignonra, a magas etanol a Pinot Noirra, a magas 2-feniletanol pedig a Merlot-ra volt jellemző. A 2-feniletanol, a metanol és a borostyánkősav alapján különböztethető meg a két borvidék (Eger és Villány) a Cabernet Sauvignon és a Merlot esetében. A villányi mintákban ezek a paraméterek egyértelműen magasabbak

voltak, mint az egri mintákban. A borostyánkősav, a metanol és a borkősav a Kékfrankos és a Pinot Noir esetében erős osztályozó erővel bírtak, amelyek a villányi mintákban mért magas értékeikkel jellemezték a borvidéket. A magasabb etanol és borkősav értékek inkább az egri borvidékről származó Pinot Noir-ra voltak jellemzőek.



35. Ábra A fajták (Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir) és borvidékek (Eger, Villány) 2015-ből és 2016-ból származó bormintáinak jellemzői, figyelembe véve azokat a komponenseiket, amelyek mind a véletlen erdő (RF), mind pedig a támogatóvektor-gépek (SVM) modelljei szerint a hét legnagyobb elválasztási hatékonysággal rendelkeztek (etanol, borkősav, metanol, 2-feniletanol, borostyánkősav, kaftársav és szikimisav). Az eredeti értékeket a (0,1) intervallumra transzformáltuk, megtartva arányukat, hogy összehasonlíthatóvá tegyük a különböző nagyságrendű változókat.

A borminták fajta, földrajzi régió, évjárat, vagy akár szezonalitás szerinti osztályozása ^{1}H NMR adatok alapján napjainkban igen intenzíven vizsgált téma. Az adatok elemzésére számos többváltozós módszer használata terjedt el, mint például a MANOVA (Alsante és mtsai, 2011, Geana és mtsai, 2016, Gougeon és mtsai, 2018, Mascellani és mtsai, 2021), a klaszterelemzés (Liu és mtsai, 1996), a főkomponens-elemzés (PCA, Liu és mtsai, 1996, Pereira és mtsai, 2005, Alsante és mtsai, 2011, Caruso és mtsai, 2012, Amargianitaki és Spyros, 2017, Gougeon és mtsai, 2018, Filho és mtsai, 2019, Mascellani és mtsai, 2021), a parciális legkisebb négyzetek módszere és a diszkriminanciaelemzés (PLS, DA, Sing és Sander, 2005, Du és mtsai, 2007, Viggiani és mtsai, 2008, Son és mtsai,

2008, Caruso és mtsai, 2012, Papotti és mtsai, 2013, Monakhova és mtsai, 2015, Geana és mtsai, 2016, Filho és mtsai, 2019), a látens struktúrákra történő ortogonális vetítés (OPLS, Ali és mtsai, 2011), a lineáris vagy kvadratikus diszkriminanciaelemzés (LDA/QDA, Alsante és mtsai, 2011, Geana és mtsai, 2016, Martelo-Vidal, 2016, Amargianitaki és Spyros, 2017) és az osztályanalógák közvetett modellezése (*Soft Independent Modelling of Class Analogies*, SIMCA, Martelo-Vidal és Vázquez, 2016).

Monakhova és munkatársai (2015) német borok NMR-ujjlenyomatainak elemzése során úgy javították diszkriminanciaelemzés-eredményeiket, hogy előzetesen PCA-módszert alkalmaztak és a kapott független főkomponenseket alkalmazták a DA magyarázó változóiként.

Az élelmiszer-tudományokban széles körben használt gépi tanulási osztályozási módszerek, mint például a mesterséges neurális hálózatok (ANN), a véletlen erdő (RF) és a támogató vektor gépek (SVM) a borminták megkülönböztetésére érdekes módon csak késlekedve, az utóbbi években terjedtek el széleskörűen, pedig ezek igen hatékony modellek. Az RF és az SVM egyik előnye más osztályozási módszerekkel szemben az, hogy kevésbé hajlamosak a túlillesztésre. Az SVM a mért változók számához képest mérsékelten alacsony mintanagyság esetén is jól működik, bár nagyon nagy adathalmazok esetén kevésbé sikeres.

Az LDA-, az ANN-, az RF- és a SVM-módszerek összehasonlítására egy kiegyensúlyozatlan és nem magas mintaelemszámú adatsor birtokában a fenti előnyök voltak a fő motivációm. A két régióból származó négy fajta ^{1}H NMR-adatai alapján az RF- és SVM-modelleket találtam a leghatékonyabbnak, ami összhangban van Martelo-Vidal és Vázquez (2016) eredményével, akik három módszert (SIMCA, LDA és SVM) hasonlítottak össze 39 polifenolos profilokat tartalmazó borminta alapján, melyek két spanyolországi régióból származtak; az osztályozásban az SVM-módszert találták a leghatékonyabbnak.

Mascellani és munkatársai (2021) különböző típusú, fajtájú és földrajzi eredetű cseh bormintákat vizsgáltak. PCA, hierarchikus klaszterezés, PLS-, DA- és RF-módszereket alkalmaztak különböző osztályozási célokkal. A legsikeresebb RF-modelleket 13 borszőlőfajta osztályozására fejlesztették három lépésben, ^{1}H NMR spektroszkópia alapján. Az első lépésben különböző bortípusokat osztályoztak, a második lépésben különböző fajtákat különböztettek meg, a harmadik lépésben pedig a kifejezetten a Chardonnay és a Pinot Noir borok megkülönböztetésével foglalkoztak. Ami a fajtaosztályozás sikerességét illeti (földrajzi régió szerinti besorolás nélkül), a Pinot Noir (96%), Kékfrankos (96%) és Cabernet Sauvignon (77%) esetében olyan eredményeket kaptak, amelyek összhangban vannak a mi RF-eredményeinkkel (Pinot Noir (0,99% és 100%), Kékfrankos (98%, 100%), Cabernet Sauvignon (98% és 99%) Eger és Villány esetében), de mi származást is figyelembe véve kaptuk ezeket az eredményeket. A legfontosabb jellemzőkként a prolint, a fenilalanint, a metanolt, a katekint, a tirozint és az epikatekint jelölték meg, amelyek közül a metanol és a katekin esetében modelljeink egyetértésben voltak. Viggiani és munkatársai (2008) a bormintákban szintén a borostyánkősavat, a prolint és a 2,3-butándiolt találták a legmegkülönböztetőbb összetevőként, amikor évjárat és földrajzi régió szerint különítették el azokat.

Godelmann és munkatársai (2013) ^{1}H NMR-spektroszkópiával elemeztek Németországban termelt különböző borfajtákat. Többváltozós adatelemzést (PCA, LDA és MANOVA) alkalmaztak. A Pinot Noir ismét a sikeresen osztályozott fajták között volt, aminek eredményeképpen a szikimisav, a kaftársav és a 2,3-butándiol rendelkezett a legerősebb osztályozó erővel. Az osztályozást a szőlőfajta, az évjárat és a földrajzi eredet tekintetében végezték. Papotti és munkatársai (2013) eredményei is megfelelnek ezeknek

az eredményeknek, mivel ők is a 2,3-butándiolt, a tej- és a borostyánkősavat, a treonin és az almasavat találták a legfontosabb vegyületeknek a PCA- és PLS- és DA-alapú fajtaosztályozásukban.

Geana és munkatársai (2016) a Romániában termelt többek között Cabernet Sauvignon, Merlot és Pinot Noir borokat szintén sikeresen megkülönböztették egymástól NMR-alapú metabolizmus segítségével, MANOVA és LDA alkalmazásával. A fajták osztályozásában a szikimisav, a tejsav, az ecetsav, a citromsav és a borostyánkősav bizonyult a legjelentősebb változóknak. Arról számoltak be, hogy a rendelkezésükre álló hat évjáratból származó adatok alapján az évjáratokat is sikeresen meg tudták különböztetni. Magdas és munkatársai (2019) azonban nem jártak sikerrel, amikor öt évjáratot próbáltak megkülönböztetni PCA és LDA modellekkel. Ahogy mi, úgy Caruso és munkatársai (2012) is két évjáratot elemeztek PCA-, LSD-, PLS-, illetve DA-alapú megközelítéssel, az évjáratok elválasztása esetükben sem volt kivitelezhető.

6.6.5 Következtetés

Az ^{1}H NMR-alapú metabolizmusadatokat véletlen erdő, illetve támogatóvektorgépek-módszerekkel elemezve eredményesnek bizonyult az Egerből és Villányból 2005-ből és 2006-ból származó Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir borok fajták és régiók szerinti osztályozás. A kifejezetten magas helyes besoroláson túl ezeknek a módszereknek megvan az az előnye, hogy kiválaszthatjuk a legfontosabb, az elkülönítésre jelentős hatást gyakorló tulajdonságokat. Esettanulmányunkban a szikimisav, a kaftársav, a metanol, a borkősav, az etanol, a 2-feniletanol és a szukcininsav voltak a legfontosabb jellemzők. A nagyon kevés mintaelemszámú csoportok esetén e mégoly sikeres módszerek sem eredményeztek elfogadható megkülönböztetést. A lineáris diszkriminanciaelemzés és a mesterséges neurális hálózatok kevésbé bizonyultak sikeresnek osztályozási eredményeiket illetően.

6.7 A rózsagyökérben (*Rhodiola rosea*) való metabolit-felhalmozódás és génexpresszió időbeli változásának hasonlósági struktúrája

E fejezet az alábbi publikáción alapszik:

Mirmazloum I., **Ladányi M.**, György Zs. 2015. 'Changes in the Content of the Glycosides, Aglycons and their Possible Precursors of Rhodiola rosea during the Vegetation Period'. Nat Prod Commun. 10(8) pp. 1413–1416. https://doi.org/10.1177/1934578X150100082 **Q2**

6.7.1 Motiváció

A *Rhodiola rosea* L. (Crassulaceae) más néven rózsagyökér a népi gyógyászatban évszázadok óta alkalmazott gyógynövény, amelynek immunstimuláns tulajdonságai régóta ismertek. Az Európa és Ázsia magas sarkvidéki részein honos, vadon termő egyedek begyűjtése igen körülményes.

6.7.2 Az adatok alapjául szolgáló mérések és célkitűzés

Négy rózsagyökér (*Rhodiola rosea* L.) egyednek (L, M, U, G) leveléből és gyökértörzséből vett mintákból négy gén (PAL, 4CL, CCR, CAD) relatív génexpresszióját mérték három időpontban (vegetációs időszak elején, közepén és végén, 36. Ábra). A génexpresszió mennyisége a három megfigyelési időpontban igen heterogén volt. Azonban a fő kérdés nem magára a génexpresszió mennyiségére vonatkozott, hanem a változás dinamikája érdekelt bennünket. Tehát arra kerestük a választ, hogyan tudnánk leírni a növények növekedés-csökkenés-stagnálás mintázatát a vegetáció folyamán. Arra kerestük a választ, hogy mennyire hasonlítanak egymásra az egyes gének a négy növény két szervén mért megfigyelések alapján.

6.7.3 Módszer

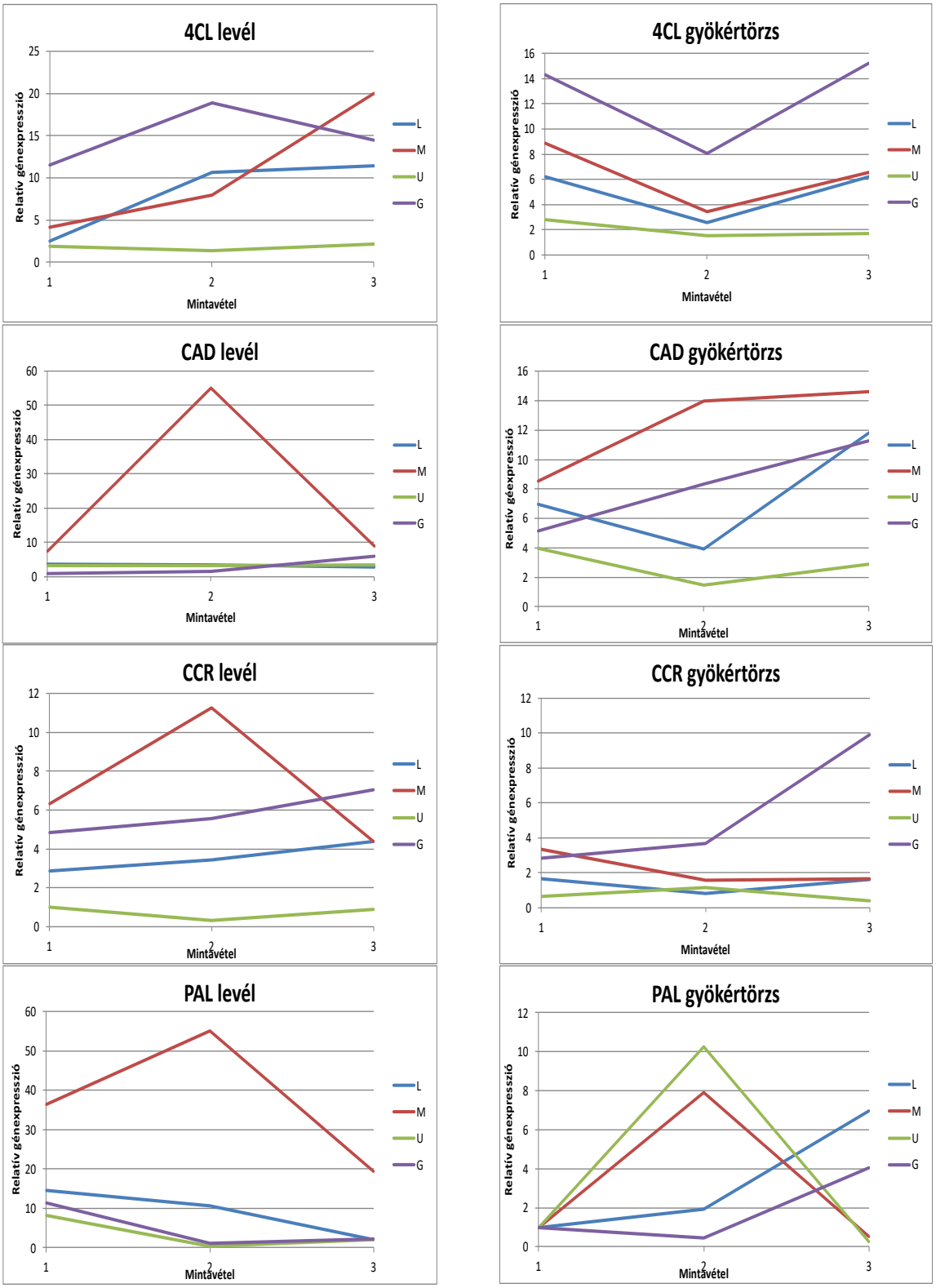
Annak érdekében tehát, hogy jellemezzük a génexpresszió folyamatának időbeli hasonlóságát, illetve különbözőségét, mind a négy növény leveléből és gyökértörzséből származó négy gén mindegyikére egy-egy háromdimenziós kódot vezettünk be, azaz 4*2*4 háromdimenziós kódot képeztünk:

$$C_k^l(i) = (c_k^l(i)_{12}, c_k^l(i)_{13}, c_k^l(i)_{23}),$$

ahol $k = 1,2, \dots 4$ jelenti az egyes géneket, $l = 1,2$, hogy a minta a levélből vagy a gyökértörzséből származott, és $i = 1,2, \dots 4$ a négy egyedre vonatkozik.

A $c_k^l(i)_{ts}$ $(ts = 12, 13, 23)$ értéke +1 vagy -1, ha a génexpresszió növekszik, illetve csökken a t-edik és az s-edik mintavételi időpontok között, és a növekedés/csökkenés mértéke meghaladja a három mintavételi alkalommal mért mennyiség átlagának 10%-át. A kód értéke 0-val egyenlő, ha a vegyület mennyiségének abszolút változása az átlag 10%-a alatt van (37. Ábra).

$C_k^l(i)$ alkalmas a folyamat leírására egy meghatározott gén, növény és mintavételi hely esetén; továbbá összehasonlíthatjuk a növényeket, a géneket és a levél-gyökér párokat az azokat jellemző kódok távolsága szerint.



36. Ábra Négy gén (PAL, 4CL, CCR, CAD) relatív génexpressziója három időpontban (vegetációs időszak eleje, közepe és vége) négy rózsatövis egyed (L, M, U, G) leveleiből és gyökértörzséből vett mintákból

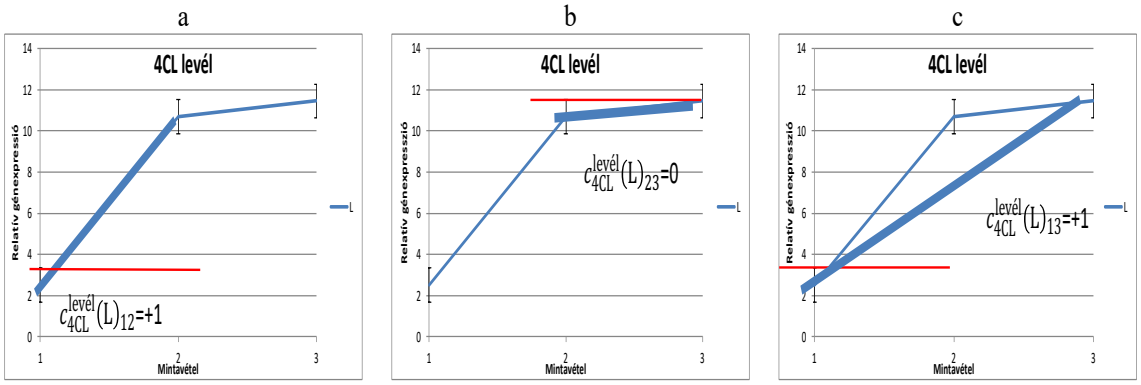
Két egyed euklideszi távolságát (rögzített k gén és l mintavételi hely esetén), egy dimenziómentes mennyiségként definiáltuk:

$$D_k^l(j,i) = D_k^l(i,j) = \sqrt{\sum_{t=1}^{2}\sum_{s=2}^{3}(c_k^l(i)_{ts} - c_k^l(j)_{ts})^2}.$$

Egy egyednek a három másik egyedtől való távolságainak összegét pedig így számoltuk:

$$I_k^l(i) = \sum_{i\neq j} D_k^l(i,j).$$

Az $I_k^l(i)$ értéket az i növény k génre és l mintavételi helyre vonatkoztatott különbözőségi indexének neveztük. Az $\overline{I_k^l} = \sum_{i=1}^{4} I_k^l(i)\,/4$ átlagot pedig a k génre és l mintavételi helyre vonatkozó átlagos különbözőségi indexnek.

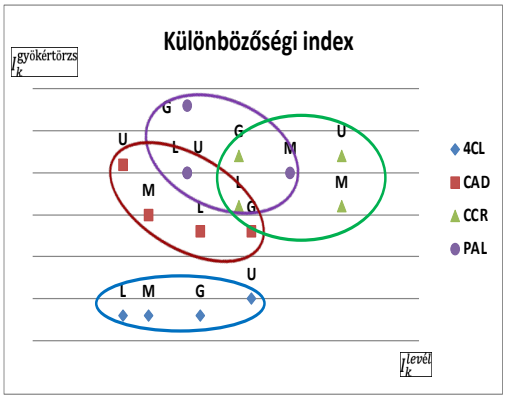


37. Ábra Példa a $C_k^l(i) = (c_k^l(i)_{12}, c_k^l(i)_{13}, c_k^l(i)_{23})$ egyes elemeinek előállítására az i = L növényminta leveléből (l = levél) vett k = 4CL gén expressziójának a vegetációs időszakban az első és második (a), a második és harmadik (b), valamint az első és harmadik (c) időpontok közötti változására

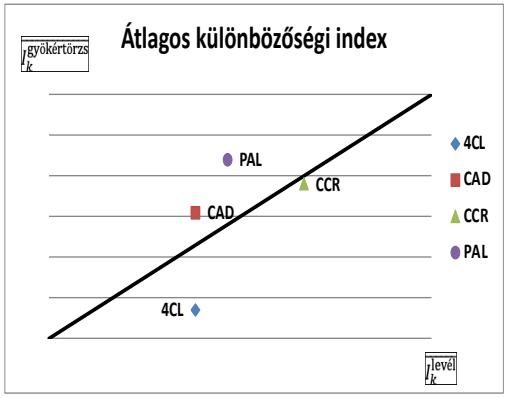
6.7.4 Eredmények

Ha most ábrázoljuk a kétdimenziós síkon a tizenhat darab $P_k(i) = (I_k^{\text{levél}}(i), I_k^{\text{gyökértörzs}}(i))$, illetve a négy darab $Q_k = \left(\overline{I_k^{\text{levél}}}, \overline{I_k^{\text{gyökértörzs}}}\right)$ pontot a négy génre vonatkozóan, akkor minél közelebb van egy pont az origóhoz, annál hasonlóbb görbecsoportot képvisel (38. Ábra). Továbbá, ha egy pont az $(y = x)$ identitás függvény alatt van ($I_k^{\text{gyökértörzs}}(i) < I_k^{\text{levél}}(i)$, illetve $\overline{I_k^{\text{gyökértörzs}}} < \overline{I_k^{\text{levél}}}$), akkor a gyökértörzsre vonatkozó görbék hasonlóbbak, míg ha az identitás függvény felett helyezkedik el ($I_k^{\text{gyökértörzs}}(i) > I_k^{\text{levél}}(i)$, illetve $\overline{I_k^{\text{gyökértörzs}}} > \overline{I_k^{\text{levél}}}$), akkor a hasonlóság a levélben fejeződik ki jobban.

a



b

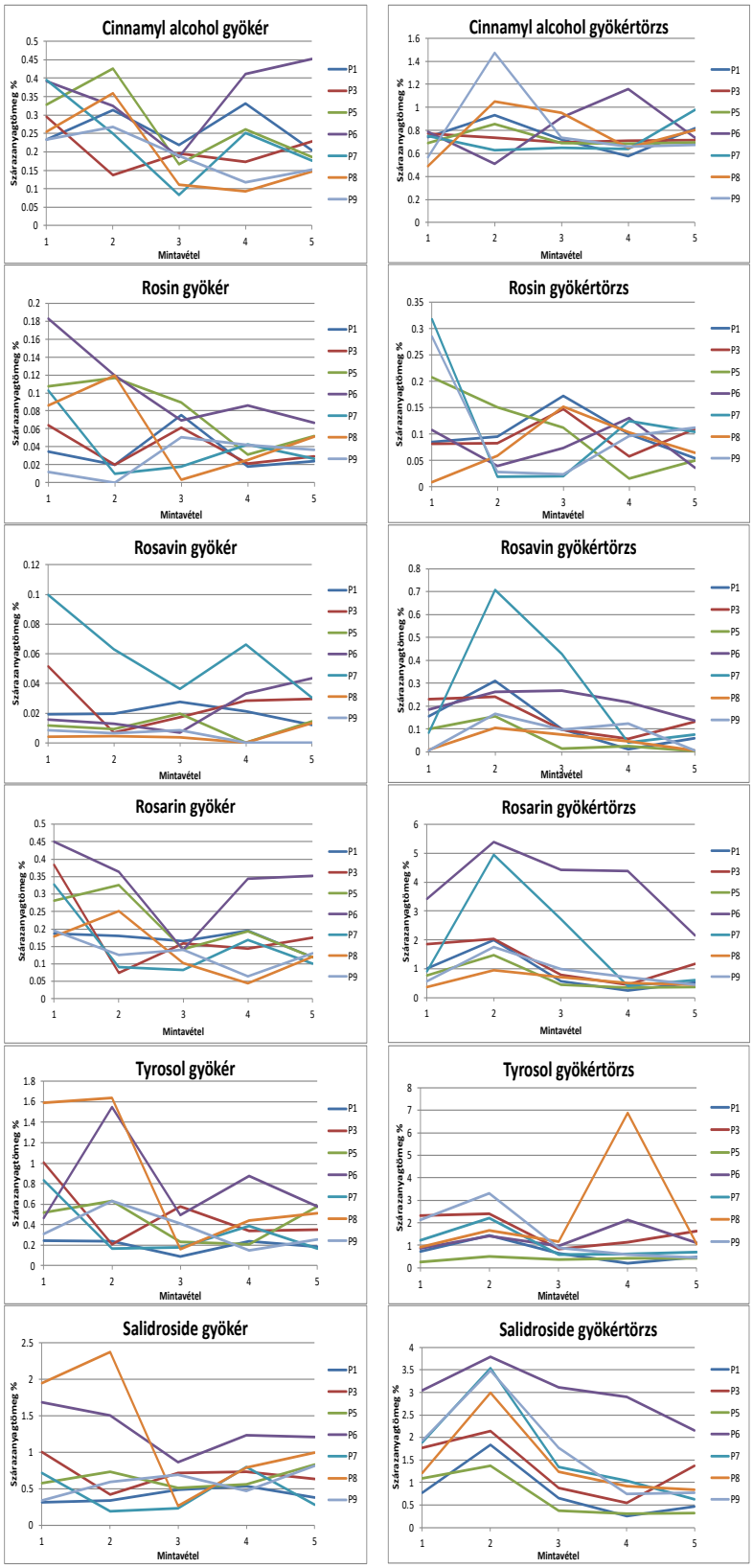


38. Ábra A $P_k(i) = (I_k^{\text{levél}}(i), I_k^{\text{gyökértörzs}}(i))$ különbözőségi indexek a négy rózsagyökér (*Rhodiola rosea* L.) egyedre ($i = \text{L, M, U, G}$) és a levélen, illetve gyökéren ($l = $ levél, illetve gyökér) mért négy gén expressziójára ($k = \text{PAL, 4CL, CCR, CAD}$) vonatkozóan (a), illetve a $Q_k = \left(\overline{I_k^{\text{levél}}}, \overline{I_k^{\text{gyökértörzs}}}\right)$ átlagos különbözőségi indexek a négy gén expressziójára vonatkozóan. A pontok a gének expressziójának a vegetációs időszakban tapasztalható változásának mintázatát jellemzik.

Jól megfigyelhető, hogy a növényekre jellemző leghasonlóbb mintázatot a gyökértörzsből származó 4CL gén expressziójának a vegetációs időszakban tapasztalható változása mutatott, e tekintetben pedig a legheterogénebb a gyökértörzsből származó PAL gén volt. A CAD és PAL gének expressziójának változása a levélből vett mintákban a növények között hasonlóbb mintázatot mutatott, mint a gyökértörzsekből vett mintákban.

6.7.5 Az adatok alapjául szolgáló mérések

Következő lépésként hét rózsagyökér (*Rhodiola rosea* L.) egyednek a vegetáció időszak alatt öt alkalommal vett gyökér-, illetve és gyökértörzsmintájából HPLC-vel hatféle vegyület (rozin, rozavin, rozarin, fahéjalkohol, szalidrozid és tirozol) mennyiségének változását kívántuk jellemezni. A metabolitok mennyisége az öt megfigyelési időpontban ismét igen heterogén volt (39. Ábra). Azonban ezek időbeli változását mintázatként értelmezve, ezen mintázatok hasonlóságát, illetve különbözőségét az előzőekben ismertetett módszer kis módosításával jellemeztük.



39. Ábra Hét rózsagyökér (*Rhodiola rosea* L.) egyednek a vegetáció időszak alatt öt alkalommal vett gyökér-, illetve és gyökértörzsmintájából HPLC-vel mért hatféle vegyület (rozin, rozavin, rozarin, fahéjalkohol, szalidrozid és tirozol) mennyiségének változása (szárazanyagtartalom %)

6.7.6 Módszer

Mind a hét növény gyökerében és gyökértörzsében lévő hat vegyület mindegyikére egy-egy tízdimenziós kódot vezettünk be, azaz 6*2*7 tízdimenziós kódot képeztünk:

$$C_k^l(i) = (c_k^l(i)_{12}, c_k^l(i)_{13}, \dots, c_k^l(i)_{15}, \dots, c_k^l(i)_{45}),$$

ahol $k = 1,2, \dots 6$ jelenti az egyes vegyületeket, $l = 1,2$, hogy a minta a gyökérből vagy a gyökértörzsből származott, és $i = 1,2, \dots 7$ a hét egyedre vonatkozik.

A $c_k^l(i)_{ts}$ $(ts = 12, 13, 14, 15, 23, 24, 25, 34, 35, 45)$ értéke +1 vagy -1, ha a vegyület tartalma növekszik, illetve csökken a t-edik és az s-edik mintavételi időpontok között, és a növekedés/csökkenés mértéke meghaladja az öt mintavételi alkalommal mért mennyiség átlagának 10%-át. A kód értéke 0-val egyenlő, ha a vegyület mennyiségének abszolút változása az átlag 10%-a alatt van.

Két egyed euklideszi távolságát (rögzített k vegyület és l mintavételi hely esetén) egy dimenziómentes mennyiségként definiáltunk:

$$D_k^l(j,i) = D_k^l(i,j) = \sqrt{\sum_{t=1}^{4} \sum_{s=2}^{5} (c_k^l(i)_{ts} - c_k^l(j)_{ts})^2},$$

egy egyednek a hat másik egyedtől való távolságainak összegét pedig így számoltuk:

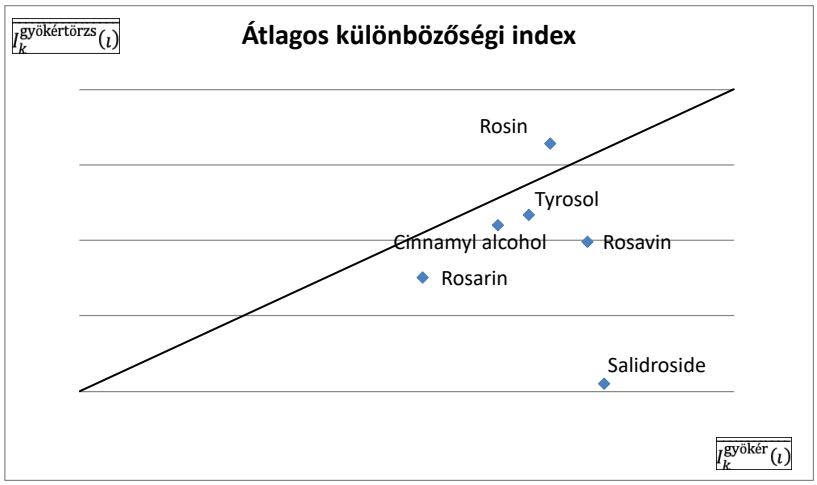
$$I_k^l(i) = \sum_{i \neq j} D_k^l(i,j).$$

Az $I_k^l(i)$ értéket az i növény k vegyületre és l mintavételi helyre vonatkoztatott különbözőségi indexének neveztük, az $\overline{I_k^l} = \sum_{i=1}^{7} I_k^l(i)\ /7$ átlagot pedig a k vegyületre és l mintavételi helyre vonatkozó átlagos különbözőségi indexnek.

6.7.7 Eredmények

Ha most ábrázoljuk a kétdimenziós síkon a hat darab $P_k = \left(\overline{I_k^{\text{gyökér}}}, \overline{I_k^{\text{gyökértörzs}}}\right)$ pontot a hat vegyületre vonatkozóan, akkor ismét igaz, hogy minél közelebb van egy pont az origóhoz, annál hasonlóbb görbecsoportot képvisel (40. Ábra). Továbbá, ha egy pont az $(y = x)$ identitás függvény alatt van $(\overline{I_k^{\text{gyökértörzs}}} < \overline{I_k^{\text{gyökér}}})$, akkor a gyökértörzsre vonatkozó görbék hasonlóbbak, míg ha az identitás függvény felett helyezkedik el $(\overline{I_k^{\text{gyökértörzs}}}(i) > \overline{I_k^{\text{gyökér}}})$, akkor a hasonlóság a gyökérben fejeződik ki jobban.

A rosin kivételével minden metabolit a gyökértörzsből való mintákban mutatott hasonlóbb mintázatot a növények között, ez a tulajdonság a szalidrosid esetében volt a legkifejezettebb.



40. Ábra A $P_k = \left(\overline{I_k^{\text{gyökér}}}, \overline{I_k^{\text{gyökértörzs}}}\right)$ átlagos különbözőségi indexek a hat vegyületre (rozin, rozavin, rozarin, fahéjalkohol, szalidrozid és tirozol) vonatkozóan. A pontok a metabolitok mennyiségének a vegetációs időszakban tapasztalható változásának mintázatát jellemzik.

6.7.8 Következtetés

A kidolgozott egyszerű dimenziómentes távolságon alapuló módszer alkalmas arra, hogy összehasonlítsuk a növények különböző szerveiben kimutatható génexpressziók, illetve metabolitok mennyiségi változásának dinamikáját.

7 KIEMELT EREDMÉNYEK

A kiemelt eredmények esetén a munkám az adatelemzés/modellezés/statisztikai elemzés volt, ezeket tekintem az én eredményeimnek. Az elemzésekből levont szakmai következtetések a társszerzőim érdemei.

1. Bemutattam és 28 saját (társszerzős) publikációból (D1:3; Q1:15; Q2:8; Q3:2) vett 22 példával szemléltettem az agrártudományi területeken folyó kutatások adatelemzésen alapuló háttérmunkájának egyes részleteit, amelyek a publikációkban már nem minden esetben láthatóak, de elengedhetetlenül szükségesek voltak ahhoz, hogy az adatokban rejlő információt megtaláljuk, megértsük, és így választ kaphassunk a kutatások célkitűzéseiben szereplő tudományos kérdésekre, illetve azokat publikációra alkalmas módon közölni tudjuk.

2. Az aszálynak a talajban élő mikroízeltlábúakra vonatkozó hatásának elemzésére kétféle megközelítést használtam, melyek alkalmasak voltak a populációknak az aszály súlyosságától és időbeli elhúzódásától függő alkalmazkodásának jellemzésére. A kétféle megközelítés (1) a relatív aktivitássűrűségen (RAD) alapuló, dinamikusan változó faktorú MANOVA alkalmazása, illetve (2) a speciális ordinális skálázás az aktivitássűrűség-különbségre (ADD) vonatkozott (Flórián és mtsai, 2019, D1).

3. 16 éves nagyparcellás szántóműveléses kísérlet adatait (kultúrnövény, talajművelés, növénytakaró), valamint csapadékindikátorokat felhasználva a teljes adathalmazra és a kukoricakultúra részadathalmazra vonatkozóan a lefolyás- és talajveszteség-események súlyosságát négy különálló véletlen erdő (RF) klasszifikációs modellel becsültem. A kukorica adatállomány esetében az RF-modell a szélsőséges eseményeket jelezte előre a legjobban. A legnagyobb és a legkisebb kockázatot jelentő csoportok érzékenysége mind meghaladta a 82%-ot a talajveszteség és a 64%-ot a lefolyás esetében. A modell segítségével azonosítani lehetett a lefolyást és a talajveszteséget befolyásoló legfontosabb tényezőket. Ezekkel az eredményekkel igazoltam, hogy a véletlen erdő modell alkalmas a lefolyás és a talajveszteség modellezésére és a kockázat súlyosságának a fenti adatokból való becslésére (Madarász és mtsai, 2021, D1).

4. Egy kelet-magyarországi, 28 parcellából álló, mintegy 225 hektáros kutatási területen termesztett őszi búza (*Triticum aestivum* L.), kukorica (*Zea mays* L.) és napraforgó (*Helianthus annuus* L.) 10 éves (2004–2013) terméshozam-adatain, valamint e parcellákról nyert 11 talajtulajdonságot leíró paramétereken és meteorológiai adatokon alapuló, a relatív terméshozamra és annak változékonyságára vonatkozó főkomponens-regressziót végeztem. Az adatok előzetes speciális transzformációjával azok összefüggéseit és hatásaik irányát jól interpretálni képes modellt kaptam. Ennek segítségével beazonosíthatóvá váltak a relatív terméshozam változékonyságát, valamint a relatív terméshozamokat leginkább meghatározó/korlátozó paraméterek az aszályos és a kevésbé száraz években (Juhos és mtsai, 2015, Q1).

5. Kidolgoztam egy, a talajminőségi indikátorhalmaz választásához és értékeléséhez használható, a talajparaméterek földhasználati értékét lineáris és nemlineáris modellekkel leíró új módszert, mely segítségével lehetővé vált Magyarország

legjellemzőbb talajtípusain a növénytermesztést korlátozó tényezők helyspecifikus azonosítása (Juhos és mtsai, 2019, Q1).

6. Összehasonlítottam öt modellt az invazív kanyargós szillevéldarázs (*Aproceros leucopoda*, Hymenoptera: Argidae) petéjének, lárvájának, bábfázisainak, valamint egy teljes nemzedékének a hőmérséklettől függő fejlődésének leírására, illetve paraméterbecsléseik alapján a minimális, maximális, optimális, illetve a fejlődéshez szükséges alsó küszöbérték feletti foknapok számának meghatározására. Igazoltam, hogy a Lactin-2 modell teljesítménye a legjobb. A modellek kritikai értékelése alapján javaslatot fogalmaztam meg a további kísérletek beállítására (Papp és mtsai, 2016, Q1, Vétek és mtsai, 2020, Q1).

7. Megmutattam, hogy a véletlen erdő, illetve a támogatóvektor-gépek alkalmas módszerek a borok ^{1}H NMR-alapú metabolizmusadatai alapján a fajták és régiók szerinti magas szintű megkülönböztetésre. A kidolgozott modelleket Egerből és Villányból 2005-ből és 2006-ból származó Cabernet Sauvignon, Kékfrankos, Merlot és Pinot Noir borokra alkalmaztam. A módszerek segítségével azonosíthatóvá vált a borok megkülönböztetése szempontjából legfontosabb hét összetevő (Nyitrainé Sárdy és mtsai, 2022, Q1).

8. Kidolgoztam egy módszert, amellyel *Rhodiola rosea* L. (Crassulaceae) növényegyedek leveléből és gyökértörzséből származó gének relatív expressziójának a vegetációs időszakban való változását jellemezhetjük a változás dinamikáját mutató mintázat leírásával. A módszer kis módosításával a növények gyökeréből és gyökértörzséből származó mintákban mért metabolitjainak a vegetációs időszakban való változását is jellemezni lehetett. A módszerrel a növények különböző szerveiben kimutatható génexpressziók, illetve metabolitok mennyiségi változásának dinamikája összehasonlíthatóvá vált (Mirmazloum és mtsai, 2015, Q2). Megfelelő módosítással a módszer más folyamatok időbeli lefutásának hasonlósági jellemzésére is alkalmazható diszkrét időpontokban mért megfigyelések alapján.

8 A PÉLDÁKBAN IDÉZETT SAJÁT PUBLIKÁCIÓK

1) Bakos J.L., **Ladányi M.** Szalay L. 2024. 'Frost hardiness of flower buds of 16 apricot cultivars during dormancy'. Folia Horticulturae, 36(1) pp. 81–93. https://doi.org/10.2478/fhort-2024-0005 Q2
2) Basky Zs. **Ladányi M.** Simončič A. 2017. 'Efficient reduction of biomass, seed and season long pollen production of common ragweed (*Ambrosia artemisiifolia* L.)'. Urban Forestry & Urban Greening 24 pp. 134–140. http://dx.doi.org/10.1016/j.ufug.2017.03.028 **D1**
3) Bodor(-Pesti) P., Baranyai L., **Ladányi M.**, Bálo B., Strever A.E., Bisztray Gy.D., Hunter J.J. 2013. 'Stability of Ampelometric Characteristics of *Vitis vinifera* L. cv. 'Syrah' and 'Sauvignon blanc' Leaves: Impact of Within-vineyard Variability and Pruning Method/Bud Load'. South African Journal of Enology and Viticulture 34(1) pp. 129–137. https://doi.org/10.21548/34-1-1088 **Q1**
4) Boziné-Pullai K., Csambalik L., Drexler D., Reiter D., Tóth F., Tóthné Bogdányi F., **Ladányi M.** 2021. 'Tomato Landraces Are Competitive with Commercial Varieties in Terms of Tolerance to Plant Pathogens – A Case Study of Hungarian Gene Bank Accessions on Organic Farms. Diversity. 13(5):195. https://www.mdpi.com/1424-2818/13/5/195 **Q1**
5) Cao T., Kovács Z., **Ladányi M.** 2023. 'Cyclic Production of Galacto-Oligosaccharides through Ultrafiltration-Assisted Enzyme Recovery'. Processes. 11(1):225. https://doi.org/10.3390/pr11010225 **Q2**
6) Divéky-Ertsey A., **Ladányi M.**, Biró B., Máté M., Drexler D., Tóth F., Boziné Pullai K., Gere A., Pusztai P., Csambalik L. 2022. 'Tomato Landraces May Benefit from Protected Production – Evaluation on Phytochemicals'. Horticulturae. 8(10):937. https://doi.org/10.3390/horticulturae8100937 **Q1**
7) Fejzullahu F., **Ladányi M.**, Kun-Farkas G. Kun S. 2024. 'Aroma Profile of Apple Spirits Fermented With non-Saccharomyces Yeasts and Their Dynamic Changes During Maturation'. Journal of the American Society of Brewing Chemists, 1–9. https://doi.org/10.1080/03610470.2024.2327911 **Q2**
8) Forgács I., Suller B., **Ladányi M.**, Zok A., Deák T., Horváth-Kupi T., Szegedi E. Oláh R. 2017. 'An improved method for embryogenic suspension cultures of 'Richter 110' rootstock'. Vitis 56 pp. 49–51. https://ojs.openagrar.de/index.php/VITIS/article/view/6751 https://doi.org/10.5073/vitis.2017.56.49-51 **Q2**
9) Hajnal V., Omid Z., **Ladányi M.**, Toth M., Szalay L. 2013. 'Microsporogenesis of Apricot Cultivars in Hungary'. Not. Bot. Horti. Agrobot. Cluj Napoca 41, pp. 434–439. https://doi.org/10.15835/nbha4129162 **Q3**
10) Király I., **Ladányi M.**, Nagyistván O., Tóth M. 2015. 'Assessment of diversity in a Hungarian apple gene bank using morphological markers'. Org. Agr. 5, pp. 143–151. https://doi.org/10.1007/s13165-015-0100-z **Q2**
11) Király K.D., **Ladányi M.**, Fail J. 2022. 'Reproductive Isolation in the Cryptic Species Complex of a Key Pest: Analysis of Mating and Rejection Behaviour of Onion Thrips (*Thrips tabaci* Lindeman). Biology. 11(3):396. https://doi.org/10.3390/biology11030396 **Q1**
12) Kiss A., Grünvald P., **Ladányi M.**, Papp V., Papp I., Némedi E., Mirmazloum I. 2021. 'Heat Treatment of Reishi Medicinal Mushroom (*Ganoderma lingzhi*) Basidiocarp Enhanced Its β-glucan Solubility, Antioxidant Capacity and Lactogenic Properties'. Foods. 10(9):2015. https://doi.org/10.3390/foods10092015 **Q1**

13) László A.M., **Ladányi M.**, Boda K., Csicsman J., Bari F., Serester A., Molnár Zs., Sepp K., Gálfi M., Radács M. 2018. 'Effects of extremely low frequency electromagnetic fields on turkeys'. Poultry Science, 97(2) pp. 634–642, https://doi.org/10.3382/ps/pex304 **D1**
14) Magyar D., Strażyński P., Grewling Ł., Pashley C.H., Satchwell J., Bobvos J., **Ladányi M.** 2023. 'The contribution of aphids (Aphidoidea) to atmospheric concentrations of *Alternaria* and *Cladosporium* spores'. Aerobiologia 39, 345–361 (2023). https://doi.org/10.1007/s10453-023-09797-4 **Q2**
15) Mirmazloum I., **Ladányi M.**, Omran M., Papp V., Ronkainen V-P., Pónya Zs., Papp I., Némedi E., Kiss A. 2021. 'Co-encapsulation of probiotic *Lactobacillus acidophilus* and Reishi medicinal mushroom (*Ganoderma lingzhi*) extract in moist calcium alginate beads'. International Journal of Biological Macromolecules, 192 pp 461–470. https://doi.org/10.1016/j.ijbiomac.2021.09.177. **Q1**
16) Musa S., **Ladányi M.**, Fail J. 2022. 'There Is No Influence of Egg Size on Sex Allocation in Arrhenotokous Lineages of *Thrips tabaci* Lindeman'. Insects. 13(5):408. https://doi.org/10.3390/insects13050408 **Q1**
17) Musa S., **Ladányi M.**, Loredo Varela R.C., Fail J. 2023. 'A morphometric analysis of *Thrips tabaci* Lindeman species complex (Thysanoptera: Thripidae)'. Arthropod Struct Dev. 72:101228. doi: 10.1016/j.asd.2022.101228 **Q1**
18) Nagy G.G., **Ladányi M.**, Arany I., Aszalós R., Czúcz B. 2017 'Birds and plants: Comparing biodiversity indicators in eight lowland agricultural mosaic landscapes in Hungary' Ecological Indicators, 73 pp. 566–573 https://doi.org/10.1016/j.ecolind.2016.09.053 **Q1**
19) Nyitrainé Sárdy Á.D., **Ladányi M.**, Varga Z., Szövényi Á.P., Matolcsi R. 2022. 'The Effect of Grapevine Variety and Wine Region on the Primer Parameters of Wine Based on 1H NMR-Spectroscopy and Machine Learning Methods'. Diversity. 14(2):74. https://doi.org/10.3390/d14020074 **Q1**
20) Palla B., **Ladányi M.**, Cseke K., Buczkó K., Höhn M. 2021. 'Wood Anatomical Traits Reveal Different Structure of Peat Bog and Lowland Populations of *Pinus sylvestris* L. in the Carpathian Region. Forests. 2021; 12(4):494. https://doi.org/10.3390/f12040494 **Q1**
21) Pázmándi M., Maráz A., **Ladányi M.**, Kovács Z. 2017. 'The impact of membrane pretreatment on the enzymatic production of whey-derived galacto-oligosaccharides' Journal of Food Process Engineering 41(2) e12649 https://doi.org/10.1111/jfpe.12649 **Q2**
22) Szalay L., **Ladányi M.**, Hajnal V., Pedryc A., Tóth M. 2016. 'Changing of the flower bud frost hardiness in three Hungarian apricot cultivars'. Horticultural Science (Prague), 43(3), pp. 134–141, https://doi.org/10.17221/161/2015-HORTSCI **Q3**
23) Szalay L., Froemel-Hajnal V., Bakos J., **Ladányi M.** 2019. 'Changes of the microsporogenesis process and blooming time of three apricot genotypes (*Prunus armeniaca* L.) in Central Hungary based on long-term observation (1994–2018)'. Scientia Horticulturae 246, pp. 279–288. https://doi.org/10.1016/j.scienta.2018.09.069 **D1**
24) Szűgyi-Reiczigel Zs., **Ladányi M.**, Bisztray G.D., Varga Zs., Bodor-Pesti P. 2022. 'Morphological Traits Evaluated with Random Forest Method Explains Natural Classification of Grapevine (Vitis vinifera L.) Cultivars'. Plants. 11(24):3428. https://doi.org/10.3390/plants11243428 **Q1**

25) Varga Á., Bihari-Lucena E., **Ladányi M.**, Szabó-Nótin B., Galambos I., Koris A. 2023. 'Experimental Study and Modeling of Beer Dealcoholization via Reverse Osmosis'. Membranes. 13(3):329. https://doi.org/10.3390/membranes13030329 **Q2**
26) Vétek G., Csávás K., Fail J., **Ladányi M.** 2022. 'Host plant range of *Aproceros leucopoda* is limited within Ulmaceae'. Agricultural and Forest Entomology. 24(1) pp 1–7. https://doi.org/10.1111/afe.12463 **Q1**
27) Vétek G., Fekete V., **Ladányi M.**, Cargnus E., Zandigiacomo P., Oláh R., Schebeck M., Schopf A. 2020. 'Cold tolerance strategy and cold hardiness of the invasive zigzag elm sawfly *Aproceros leucopoda* (Hymenoptera: Argidae)'. Agricultural And Forest Entomology 22(3) pp. 231–237. https://doi.org/10.1111/afe.12376 **Q1**
28) Woldemelak W.A., **Ladányi M.**, Fail J. 2021. 'Effect of temperature on the sex ratio and life table parameters of the leek-(L1) and tobacco-associated (T) *Thrips tabaci* lineages (Thysanoptera: Thripidae)'. 63(3) pp. 230–237. https://doi.org/10.1002/1438-390X.12082 **Q1**

9 AZ EREDMÉNYEKHEZ FELHASZNÁLT SAJÁT PUBLIKÁCIÓK

1) Flórián N., **Ladányi M.**, Ittzés A., Kröel-Dulay G., Ónodi G., Mucsi M., Szili-Kovács T., Gergócs V., Dányi L., Dombos, M. 2019. Effects of single and repeated drought on soil microarthropods in a semi-arid ecosystem depend more on timing and duration than drought severity. PLoS ONE 14(7): e0219975. https://doi.org/10.1371/journal.pone.0219975 **D1**
2) Madarász B., Jakab G., Szalai Z., Juhos K., Kotroczó Zs., Tóth A., **Ladányi M.** 2021 Long-term effects of conservation tillage on soil erosion in Central Europe: A random forest-based approach, Soil and Tillage Research, 209(104959), ISSN 0167-1987, https://doi.org/10.1016/j.still.2021.104959 **D1**
3) Juhos, K., Szabó S., **Ladányi M.** 2015. 'Influence of soil properties on crop yield: a multivariate statistical approach'. Int. Agrophys., 29(4), pp. 433–440. https://doi.org/10.1515/intag-2015-0049 **Q2**
4) Juhos K., Czigány Sz., Madarász B., **Ladányi M.** 2019. 'Interpretation of soil quality indicators for land suitability assessment – A multivariate approach for Central European arable soils, Ecological Indicators, 99, pp. 261–272. https://doi.org/10.1016/j.ecolind.2018.11.063 **Q1**
5) Papp V., **Ladányi M.**, Vétek G. 2016. 'Temperature-dependent development of *Aproceros leucopoda* (Hymenoptera: Argidae), an invasive pest of elms in Europe' Journal of Applied Entomology, 42(6) pp. 589–597. https://onlinelibrary.wiley.com/doi/10.1111/jen.12503 **Q1**
6) Vétek G., Fekete V., **Ladányi M.**, Cargnus E., Zandigiacomo P., Oláh R., Schebeck M., Schopf A. 2020. 'Cold tolerance strategy and cold hardiness of the invasive zigzag elm sawfly *Aproceros leucopoda* (Hymenoptera: Argidae)'. Agricultural And Forest Entomology 22(3) pp. 231–237. https://doi.org/10.1111/afe.12376 **Q1**
7) Nyitrainé Á. D., **Ladányi M.**, Varga Z., Szövényi Á.P., Matolcsi R. 2022. 'The Effect of Grapevine Variety and Wine Region on the Primer Parameters of Wine Based on 1H NMR-Spectroscopy and Machine Learning Methods'. Diversity. 14(2):74. https://doi.org/10.3390/d14020074 **Q1**
8) Mirmazloum I., **Ladányi M.**, György Zs. 2015. 'Changes in the Content of the Glycosides, Aglycons and their Possible Precursors of *Rhodiola rosea* during the Vegetation Period'. Nat Prod Commun. 10(8) pp. 1413–1416. https://doi.org/10.1177/1934578X1501000825 **Q2**

10 FELHASZNÁLT IRODALOM

Abdul-Hamid N.A., Abas F., Ismail I.S., Tham C.L., Maulidiani M., Mediani A., Swarup S., Umashankar S., Zolkeflee N.K. 2019. 'Metabolites and biological activities of Phoenix dactylifera L. pulp and seeds: A comparative MS and NMR based metabolomics approach'. Phytochem. Lett. 31, pp. 20–32. https://doi.org/10.1016/j.phytol.2019.03.004

Agresti A. 2002. 'Categorical Data Analysis'. Wiley. ISBN:9780471360933 |Online ISBN:9780471249689 |DOI:10.1002/0471249688 p. 710. 3. kiadás 2012: ISBN 978-0-470-46363-5

Ahman M.N., Bhatti A.U. 2015. 'Comparison of regression models to predict potential yield of wheat from some measured soil properties.' Pakistan J. Agric. Sci., 52(1) pp. 239-257.

Akaike, H. 1973. 'Information theory and an extension of the maximum likelihood principle', Second International Symposium on Information Theory, pp. 267–281. https://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15

Ali K., Maltese F., Toepfer R., Choi Y.H., Verpoorte R. 2011. 'Metabolic characterization of Palatinate German white wines according to sensory attributes, varieties, and vintages using NMR spectroscopy and multivariate data analyses'. J. Biomol. NMR, 49, pp. 255–266. https://doi.org/mer10.1007/s10858-011-9487-3.

Allaire J., Chollet F. 2024. 'keras: R Interface to Keras'. R package version 2.15.0, <https://CRAN.R-project.org/package=keras>.

Alsante K.M., Baertschi S.W., Brian M.C., Marquez L., Sharp T.R.; Zelesky T.C. 2011. 'Degradation and Impurity Analysis for Pharmaceutical Drug Candidates'. Sep. Sci. Technol. 10, pp. 59–169. https://doi.org/10.1016/B978-0-12-375680-0.00003-6

Amargianitaki M., Spyros A. 2017. 'NMR-based metabolomics in wine quality control and authentication'. Chem. Biol. Technol. Agric. 4, 9. https://doi.org/10.1186/s40538-017-0092-x.

Analytis S. 1981. 'Relationship between temperature and development times in phytopathogenic fungus and in plant pests: a mathematical model'. Agric Res Athens 5: pp. 133–159.

Anastasiadi M., Zir A., Magiatis P., Haroutounian S.A., Skaltsounis A.L., Mikros, E. 2009. '^{1}H NMR-based metabolomics for the classification of Greek wines according to variety, region, and vintage. Comparison with HPLC data'. J. Agric. Food Chem. 57, pp. 11067–11074. https://doi.org/10.1021/jf902137e

Arnqvist G. 2020. 'Mixed Models Offer No Freedom from Degrees of Freedom'. Trends in Ecology and Evolution, 35(4) pp. 329–335. https://doi.org/10.1016/j.tree.2019.12.004.
https://www.sciencedirect.com/science/article/pii/S0169534719303465

Ayoubi S., Khormali F., Sahrawat K.L. 2009. 'Relationship of barly biomass and grain yields to soil properties within a field in the arid region: Use of factor analysis.' Acta. Agric. Scand. Section B-Soil Plant. Sci., 59(2) pp. 107–117.

Bates D., Maechler M., Bolker B. Walker S. 2015. 'Fitting Linear Mixed-Effects Models Using lme4'. Journal of Statistical Software, 67(1) pp. 1–48. doi:10.18637/jss.v067.i01.

Beaton D., Fatt C.R. C., Abdi H. 2014. 'An ExPosition of multivariate analysis with the singular value decomposition in R'. Computational Statistics & Data Analysis, 72(0), pp. 76–189.

Beyeler M., Beyeler E., Rounds E., Rounds K.D., Carlson S., Krichmar J. 2019. 'Neural correlates of sparse coding and dimensionality reduction'. PLOS Computational Biology 15(6):e1006908 DOI: 10.1371/journal.pcbi.1006908

Bieri M., Baumgartner J., Bianchi G., Delucchi V., Arx vA. 1983. 'Development and fecundity of pea aphid (*Acyrthosiphon Pisum* Harris) as affected by constant temperatures and by pea varieties'. Mitteilungen der Schweizerischen Entomologischen Gesellschaft 56(1/2) pp. 163–171. https://doi.org/10.5169/seals-402070

Bishop, C. 1995. 'Neural Networks for Pattern Recognition'. Oxford University Press: Oxford, UK, 1995. Újranyomva 1996-ban kétszer, 1997-ben kétszer, 1998-ban, 1999-ben, 2000-ben, 2002.ben, 2003-ban kétszer, 2004-ben kétszer, 2005-ben ISBN: 019853864 2

Boehmke B., Greenwell B. 2020. 'Hands-on Machine Learning with R'. CRC Press, Taylor and Francis Group, Chapman and Hall. ISBN 978-1-138-49568-5 p. 484 https://bradleyboehmke.github.io/HOML/

Bolker B.M., Brooks M.E., Clark C.J., Geange S.W., Poulsen J.R., Stevens M.H.H., White J-S.S. 2009. 'Generalized linear mixed models: a practical guide for ecology and evolution'. Trends in Ecology & Evolution, 24(3) pp. 127–135. https://doi.org/10.1016/j.tree.2008.10.008. url

Bougeard S., Dray S. 2018. 'Supervised Multiblock Analysis in R with the ade4 Package.' Journal of Statistical Software, 86(1), pp. 1–17. doi:10.18637/jss.v086.i01 <https://doi.org/ 10.18637/jss.v086.i01>.

Box, G.E., Cox, D.R. 1964. 'An analysis of transformations'. Journal of the Royal Statistical Society. Series B (Methodological), pp. 211–252.

Breiman L. 1996a. 'Bagging Predictors'. Machine Learning 26(2) pp. 123–40.

Breiman L., Cutler A., Liaw A., Wiener M. 2018. 'Breiman and Cutler's Random Forests for Classification and Regression, R package Version 4.6–14'. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Breiman L., Friedman J., Olshen R., Stone, C. 1984. 'Classification and Regression Trees'. Wadsworth International Group. Imprint Chapman and Hall/CRC, New York, 2017. ISBN 9781315139470 p 368. https://doi.org/10.1201/9781315139470

Breiman L., Friedman, J., Charles J., Stone, C.J., Olshen, R.A. 1984. 'Classification and Regression Trees'. Routledge. eBook 2017. Imprint: Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315139470 eBook ISBN: 9781315139470 ISBN 9780412048418 p. 368.

Breiman, L. 1996b. 'Out-of-bag estimation', Semantic Scholar, url

Breiman, L. 2001. 'Random Forests'. Machine Learning 45(1) pp. 5–32. doi: 10.1023/A:1010933404324.

Brière J.F., Pracros P. 1998. 'Comparison of temperature-dependent growth models with the development of *Lobesia botrana* (Lepidoptera: Tortricidae)'. Environmental Entomology 27: pp. 94-101. https://doi.org/10.1093/ee/27.1.94

Brière J.F., Pracros P., le Roux A.Y., Pierrže J.S. 1999. 'A novel rate model of temperature dependent development for arthropods'. Environmental Entomology 28. pp. 22–29. https://doi.org/10.1093/ee/28.1.22

Buuren van S., Groothuis-Oudshoorn, K 2011. 'mice: Multivariate Imputation by Chained Equations in R'. Journal of Statistical Software, 45(3) pp. 1–67. DOI 10.18637/jss.v045.i03.

Caruso M., Galgano F., Castiglione Morelli M.A., Viggiani L., Lencioni L., Giussani B., Favati F. 2012. 'Chemical profile of white wines produced from 'Greco bianco' grape variety indifferent Italian areas by Nuclear Magnetic Resonance (NMR) and

conventional physico chemical analyses'. J. Agric. Food Chem. 60, pp. 7–15. https://doi.org/10.1021/jf204289u

Chen T, He T, Benesty M, Khotilovich V., Tang Y., Cho H., Chen K., Mitchell R., Cano I., Zhou T., Li M., Xie J., Lin M., Geng Y., Li Y., Yuan J. 2024. 'xgboost: Extreme Gradient Boosting'. R package version 1.7.7.1, <https://CRAN.R-project.org/package=xgboost>.

Chessel D., Dufour A., Thioulouse J. 2004. 'The ade4 Package - I: One-Table Methods.' R News, 4(1), 5-10. <https://cran.r-project.org/doc/Rnews/>.

Chollet F., Allaire J.J. 2018. 'Deep Learning with R'. Manning Publications Company. ISBN 9781617295546 p. 360

Cohen J. 1960. 'A Coefficient of Agreement for Nominal Scales'. Educational and Psychological Measurement 20(1) pp. 37–46. doi: 10.1177/001316446002000104.

Cohen J. 1968. 'Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit'. Psychological Bull. 70, pp. 213–220. doi: 10.1177/001316446002000104

Consonni R., Astraka K., Cagliani L.R., Nenadis N., Petrakis E., Polissiou M. 2016. 'Authenticity of food'. In Encyclopedia of Food and Health; Caballero, B., Finglas, P., Toldrá, F., Eds.; Oxford Academic Press: Oxford, UK, pp. 285–293. DOI:10.1016/B978-0-12-384947-2.00048-9

Consonni R., Cagliani L.R. 2010. 'Chapter 4 - Nuclear magnetic resonance and chemometrics to assess geographical origin and quality of traditional food products. In Advances in Food and Nutrition Research', Steve L.T. Ed. Cambridge Academic Press: Cambridge, UK, 59, pp. 87–165. https://doi.org/10.1016/S1043-4526(10)59004-1

Consonni R., Cagliani L.R. 2019. 'The potentiality of NMR-based metabolomics in food science and food authentication assessment'. Magn. Reson. Chem. MRC 57, pp. 558–578. https://doi.org/10.1002/mrc.4807

Consonni R., Cagliani L.R., Guantieri V., Simonato B. 2017. 'Identification of metabolic content of selected Amarone wine'. Food Chem. 129, pp. 693–699. https://doi.org/10.1016/j.foodchem.2011.05.008

Corwin D.L., Lesch S.M., Shouse P.J., Soppe R., Ayars J.E. 2003. 'Identifying soil properties that influence cotton yield using soil sampling directed by apparent soilelectrical con-ductivity.' Agron. J., 95(2) pp. 352–364.

Cox M.S., Gerard D.P., Wardlaw M.C., Abshire M.J., 2003. 'Variability of selected soil properties and their relationships with soybean yield.' Soil Sci. Soc. Am. J., 67. pp.1296–1302.

CRAN 2021. 'The Comprehensive R Archive Network Https://Cran.r-Project.Org/'.

Cristianini, N., Shawe-Taylor J. 2000. 'An Introduction to Support Vector Machines and Other Kernel-Based Leaning Methods'; Cambridge University Press: Cambridge, UK, ISBN 0-521-78019-5.

Cunningham P., Delany S.J. 2007. 'K-nearest neighbour classifiers'. Multiple Classifier Systems, 34(8) pp. 1–17.

D'Agostino R.B. 1971. 'An omnibus test of normality for moderate and large sample size'. Biometrika, 58, pp. 341–348.

D'Agostino R.B. és Pearson E.S. 1973. 'Testing for departures from normality'. Biometrika, 60(3) pp. 613–622. doi:10.2307/2335012

Dray S., Dufour A. 2007. 'The ade4 Package: Implementing the Duality Diagram for Ecologists.' Journal of Statistical Software, 22(4) pp. 1–20. doi:10.18637/jss.v022.i04 <https://doi.org/10.18637/jss.v022.i04>.

Dray S., Dufour A., Chessel D. 2007. 'The ade4 Package - II: Two-Table and K-Table Methods.' R News, 7(2) pp. 47–52. <https://cran.r-project.org/doc/Rnews/>.

Du Y.Y., Bai G.Y., Zhang X., Liu M.L. 2007. 'Classification of wines based on combination of ^{1}H NMR spectroscopy and principal component analysis'. Chin. J. Chem. 25, pp. 930–936. https://doi.org/10.1002/cjoc.200790181

Dunn J.C. 1974. 'Well-separated clusters and optimal fuzzy partitions'. Journal of Cybernetics 4(1) pp. 95–104.

Efron B. 1979. 'Bootstrap Methods: Another Look at the Jackknife.' Ann. Statist. 7 (1) pp. 1–26, https://doi.org/10.1214/aos/1176344552

Efron B. és Tibshirani R.J. 1997. 'Improvements on cross-validation: the 632+ bootstrap method'. Journal of the American Statistical Association, 92(438) pp. 548–560.

Efron B. és Tibshirani R.J., 1993. 'An Introduction to the Bootstrap', Chapman and Hall. ISBN 9780429246593 p. 456.

Eye von A., Wiedermann W. 2023. 'The General Linear Model: A Primer'. Cambridge University Press, ISBN: 1009322176,9781009322171 p. 175.

Faraway, J.J. 2015. 'Linear Models with R'. Chapman and Hall/CRC. ISBN: 13: 978–1-4398-8734-9 (eBook - PDF) p. 255. url

Farkas P., György Z., Tóth A., Sojnóczki A., Fail J. 2020. 'A simple molecular identification method of the *Thrips tabaci* (Thysanoptera: Thripidae) cryptic species complex'. Bulletin of Entomological Research. 110(3) pp. 397–405. doi:10.1017/S0007485319000762

Fausett L. 1994. 'Fundamentals of Neural Networks'; Prentice Hall: New York, NY, USA, ISBN 0133341860, 9780133341867. p 461.

Fawcett T. 2006. 'An introduction to ROC analysis'. Pattern Recognit. Lett. 27, 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Field A. 2024a. 'Discovering statistics using IBM SPSS statistics'. 6. kiadás, Sage. p. 1144. ISBN 9781529630008

Field A. 2024b. 'Discovering statistics using R and RStudio'. 2. kiadás, Sage. p. 992. ISBN 1526461366

Field A., Miles, J., Field, Z. 2012. 'Discovering Statistics Using R' Sage ISBN 978-1-4462-0045-2, 978-1-4462-0046-9 p. 1426

Filho E.J.A., Silva L.M.A., Ribeiro P.R.V., de Brito E.S., Zocolo G.J., Souza-Leao P.C., Marquez A.T.B., Quintela A.L., Larsen F.H., Canuto K.M. 2019. '^{1}H NMR and LC-MS-based metabolomic approach for evaluation for the seasonality and viticultural practices in wines from Sao Francisco River Valley, a Brazilian semi-arid region'. Food Chem. 289, pp.558–567. https://doi.org/10.1016/j.foodchem.2019.03.103

Fisher R.A. 1922. 'On the Mathematical Foundations of Theoretical Statistics'. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 222(594–604), pp. 309–368. https://doi.org/10.1098/rsta.1922.0009

Fishman, G.S. 1996. 'Monte Carlo: Concepts, Algorithms, and Applications'. Springer. ISBN 978-1-4419-2847-4 ISBN 978-1-4757-2553-7 (eBook) DOI 10.1007/978-1-4757-2553-7 https://link.springer.com/book/10.1007/978-1-4757-2553-7

Fleiss J.L., Cohen J. 1973. 'The Equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability'. Educ. Psychol. Meas. 33, pp. 613–619. doi: https://doi.org/10.1177/001316447303300309

Fleiss J.L., Levin B., Paik M.C. 2003. 'Statistical Methods for Rates and Proportions', Harmadik kiadás. John Wiley & Sons, Hoboken. p. 768. |DOI:10.1002/0471445428. Print ISBN:9780471526292 |Online ISBN:9780471445425

Freeberg T.M., Lucas J.R. 2009. 'Pseudoreplication is (still) a problem'. Journal of Comparative Psychology, 123(4) pp. 450–451. https://doi.org/10.1037/a0017031

Frick H., Chow F., Kuhn M., Mahoney M., Silge J., Wickham H. 2024. 'rsample: General Resampling Infrastructure'. R package version 1.2.1, <https://CRAN.R-project.org/ package=rsample>.

Fritsch S., Guenther F., Wright M. 2019. 'neuralnet: Training of Neural Networks'. R package version 1.44.2, <https://CRAN.R-project.org/package=neuralnet>.

Fryda T., LeDell E., Gill N., Aiello S., Fu A., Candel A., Click C., Kraljevic T., Nykodym T., Aboyoun P., Kurka M., Malohlava M., Poirier S., Wong W. 2024. 'h2o: R Interface for the 'H2O' Scalable Machine Learning Platform'. R package version. 3.44.0.3, <https://CRAN.R-project.org/package=h2o>.

Geana E.I., Popescu R., Costinel D., Dinca O.R., Ionete R.E., Stefanescu I., Artem V., Balaet, C. 2016. 'Classification of red wines using suitable markers coupled with multivariate statistic analysis'. Food Chem. 192, pp. 1015–1024. https://doi.org/10.1016/j.foodchem.2015.07.112

Géron A. 2017. 'Hands-on Machine Learning with Scikit-Learn and Tensor-Flow: Concepts, Tools, and Techniques to Build Intelligent Systems'. O'Reilly Media, Inc.

Ghatak A. 2017. 'Machine Learning with R'. Springer Singapore. p. 210. ISBN 978-981-10-6807-2, e-ISBN 978-981-10-6808-9 DOI 10.1007/978-981-10-6808-9

Gini C. 1912. 'Variabilità e mutabilità [Variability and Mutability]'. Bologna: Tipografia di P. Cuppini.

Godelmann R., Fang F., Humpfer E., Schütz B., Bansbach M., Schäfer H.; Spraul M. 2013. 'Targeted and Nontargeted Wine Analysis by ^{1}H NMR Spectroscopy Combined with Multivariate Statistical Analysis. Differentiation of Important Parameters: Grape Variety, Geographical Origin, Year of Vintage'. Agric. Food Chem. 61, pp. 5610–5619. https://doi.org/10.1021/jf400800d

Goodman L.A., Kruskal W.H. 1954. 'Measures of association for cross classifications'. Part I. J. Am. Stat. Assoc. 49, pp. 732–764. online megjelenés: 2012. https://doi.org/10.1080/01621459.1954.10501231

Goodman L.A., Kruskal W.H. 1979. 'Measures of Association for Cross Classifications'. 1. fejezet In: Measures of Association for Cross Classifications. Springer Series in Statistics. Springer, New York, NY. pp 2–34. https://doi.org/10.1007/978-1-4612-9995-0_1

Gougeon L., da Costa G., Le Mao I., Ma W., Teissedre P.L., Guyon F., Richard T. 2018. 'Wine analysis and authenticity using ^{1}H NMR metabolomics data: Application to Chinese wines'. Food Anal. Methods, 11, pp. 3425–3434. https://doi.org/10.1007/s12161-018-1310-2

Gower J.C. 1971. 'A general coefficient of similarity and some of its properties'. Biometrics. 27 (4) pp. 857–871. doi:10.2307/2528823. JSTOR 2528823. Retrieved 2024-06-03.

Gregorutti B., Michel B., Saint Pierre P. 2015. 'Grouped variable importance with random forests and application to multiple functional data analysis'. Computational Statistics and Data Anal. 90, pp. 15–35. https://doi.org/10.1016/j.csda.2015.04.002

Gu S. 1999. 'Lethal temperature coefficient – a new parameter for interpretation of cold hardiness'. The Journal of Horticultural Science and Biotechnology, 74 pp. 53–59. https://doi.org/10.1080/14620316.1999.11511071

Hajian-Tilaki K. 'Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation'. Caspian J Intern Med. 2013. Spring;4(2) pp. 627–35. PMID: 24009950; PMCID: PMC3755824.

Halkidi M., Batistakis Y., Vazirgiannis M. 2001. 'On Clustering Validation Techniques'. Journal of Intelligent Information Systems 17, pp. 107–145. https://doi.org/10.1023/ A:1012801612483

Hastie T., Friedman J., Tibshirani R. 'The Elements of Statistical Learning: data mining, inference, and prediction', Springer 2009, 2013: 10th printing with corrections. ISBN: 978-0-387-84857-0. p. 745.

Honěk A. 1996. ’Geographical variation in thermal requirements for insect development’. European Journal of Entomology, 93(3): pp. 303–312. url.

Hosmer Jr., D.W., Sturdivant R.X., Lemeshow S. 2013. 'Applied Logistic Regression', Wiley, Harmadik kiadás. ISBN: 978-0-470-58247-3 p. 528

Hutcheson G., Sofroniou N. 1999. 'The multivariate social scientist'. London: Sage. DOI: https://doi.org/10.4135/9780857028075

IBM Corp. Released 2023. IBM SPSS Statistics for Windows, Version 29.0.2.0 Armonk, NY: IBM Corp

Jalali M.A., Tirry L., Arbab A., De Clercq P. 2010. ‘Temperature-dependent development of the two-spotted ladybeetle *Adalia bipunctata,* on the green peach aphid, *Myzus persicae,* and a factitious food under constant temperatures’. Journal of Insect Scienc*e*, 10(1) 124, https://doi.org/10.1673/031.010.12401

James G., Witten D., Hastie T., Tibshirani R. 2013. 'An Introduction to Statistical Learning: with Applications in R'. Springer Texts in Statistics 103. Springer ISBN 1461471389, 9781461471387 p. 426.

Joanes D.N, Gill, C.A. 1998. 'Comparing Measures of Sample Skewness and Kurtosis'. The Statistician 47(1) pp.183–189. https://www.jstor.org/stable/2988433

Kaiser H.F. 1970. 'A second-generation little jiffy'. Psychometrika, 35, pp. 401415.

Kaiser H.F. 1974. 'An index of factorial simplicity'. Psychometrika, 39, pp. 31–36.

Kalogiouri N.P., Samanidou V.F. 2021. 'Liquid chromatographic methods coupled to chemometrics: A short review to present the key workflow for the investigation of wine phenolic composition as it is affected by environmental factors'. Environ. Sci. Pollut. Res. 28, pp. 59150–59164. https://doi.org/10.1007/s11356-020-09681-5.

Kang Z., Catal C., Tekinerdogan B. 2020. ‘Machine Learning Applications in Production Lines: A Systematic Literature Review’. Computers and Industrial Engineering 149:106773.

Karatzoglou A., Smola A., Hornik K. 2023. 'kernlab: Kernel-Based Machine Learning Lab'. R package version 0.9-32, <https://CRAN.R-project.org/package=kernlab>.

Karatzoglou A., Smola A., Hornik K., Zeileis A. 2004. 'kernlab - An S4 Package for Kernel Methods in R.' Journal of Statistical Software, 11(9) pp. 1–20. doi:10.18637/jss.v011.i09 <https://doi.org/10.18637/jss.v011.i09>.

Kassambara A. 2023. 'rstatix: Pipe-Friendly Framework for Basic Statistical Tests'. R package version 0.7.2, <https://CRAN.R-project.org/package=rstatix>.

Kassambara A., Mundt F. 2020. 'factoextra: Extract and Visualize the Results of Multivariate Data Analyses'. R package version 1.0.7, <https://CRAN.R-project.org/package= factoextra>.

Kendall M. 1938. 'A New Measure of Rank Correlation'. Biometrika. 30 (1–2) pp. 81–89. doi:10.1093/biomet/30.1-2.81. JSTOR 2332226.

Kim H.Y. 2013. 'Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis'. Restorative dentistry & endodontics, 38(1) pp. 52–54. doi:10.5395/rde.2013.38.1.52

Kobayashi K., Hasegawa E. 2012. 'Discrimination of Reproductive Forms of *Thrips tabaci* (Thysanoptera: Thripidae) by PCR With Sequence Specific Primers', Journal of Economic Entomology, 105(2) pp. 555–559, https://doi.org/10.1603/EC11320

Kondakis M., Demiris N., Ntzoufras I., Papanikolaou N.E. 2022. 'Inference and model determination for temperature-driven non-linear ecological models'. Environ Ecol Stat **29**, pp. 509–555, https://doi.org/10.1007/s10651-022-00531-w

KSH 2008. 'A Fenntartható Fejlődés Indikátorai Magyarországon'. Központi Statisztikai Hivatal, Budapest. https://www.ksh.hu/docs/hun/xftp/idoszaki/fenntartfejl/fenntartfejl06.pdf

Kuhn M. 2008. 'Building Predictive Models in R Using the caret Package'. Journal of Statistical Software, 28(5)pp. 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn M. 2020. 'caret: Classification and Regression Training. R Package Version 6.0-86'. https://CRAN.R-project.org/package=caret.

Kuhn M., Johnson K. 2013. 'Applied Predictive Modeling'. Springer. ISBN: 1461468485,978-1-4614-6848-6,978-1-4614-6849-3 p. 615

Kutner M.H., Nachtsheim C. J., Neter J., Li W. 2005. 'Applied Linear Statistical Models'. McGraw Hill, Ötödik kiadás. ISBN 0-07-238688-6 p. 1415

Lactin D.J., Holliday N.J., Johnson D.L., Craigen R. 1995. 'Improved Rate Model of Temperature-Dependent Development by Arthropods, Environmental Entomology'. 24(1), pp. 68–75, https://doi.org/10.1093/ee/24.1.68

Landis J. R., Koch G.G. 1977. 'The Measurement of Observer Agreement for Categorical Data'. Biometrics 33(1)159. doi: 10.2307/2529310.

Lantz B. 2015. 'Machine Learning with R. Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R'. Második kiadás. Birmingham - Mumbai: Packt Publishing. ISBN 978-1-78439-390-8. p. 452

Lardeux F.J., Tejerina R.H., Quispe V., Chavez T.K. 2008. 'A physiological time analysis of the duration of the gonotrophic cycle of *Anopheles pseudopunctipennis* and its implications for malaria transmission in Bolivia'. Malar *J* **7**, 141 https://doi.org/10.1186/1475-2875-7-141

Li D. 1998. 'A linear model for description of the relationship between the lower threshold temperature and thermal constant in spiders (Araneae: Arachnida)'. Journal of Thermal Biology, 23, pp. 23-30. DOI: 10.1016/S0306-4565(97)00042-9

Li X., Deng S., Li, L., Jiang, Y. 2019. ' Outlier Detection Based on Robust Mahalanobis Distance and Its Application'. Open Journal of Statistics, 9, 15-26. doi: 10.4236/ojs.2019.91002.

Liaw A., Wiener M. 2002. 'Classification and Regression by randomForest'. R News 2(3), 18-22.

Liland K., Mevik B., Wehrens R. 2023. 'Partial Least Squares and Principal Component Regression'. R package version 2.8-3, https://CRAN.R-project.org/package=pls

Lindon J.C., Nicholson J.K., Holmes E., Everett J.R. 2000. 'Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids'. Concepts Magn. Reson. 12, pp. 289–320. https://doi.org/10.1002/1099-0534(2000)12:5<289::AID-CMR3>3.0.CO;2-WCitations: 370

Liu M., Nicholson J.K., Lindon J.C. 1996. 'High resolution diffusion and relaxation edited one- and two-dimensional ^{1}H NMR spectroscopy of biological fluids'. Anal. Chem. 68, pp. 3370–3376. https://doi.org/10.1021/ac960426p

Logan J. A., Wollkind D. J., Hoyt S. C., Tanigoshi L. K. 1976. ’An analytic model for description of temperature dependent rate phenomena in arthropods’. Environmental Entomology, 5, pp. 1133-1140. DOI: 10.1093/ee/5.6.1133

Louppe G., Wehenkel L., Sutera A., Geurts P. 2013. 'Understanding variable importances in forests of randomized trees'. Advances in Neural Information Processing Systems; Burges C.J.C., Bottou L., Welling M., Ghahramani Z., Weinberger K.Q., Eds.; Curran Associates, Inc. (Lake Tahoe, CA), pp. 431–439.

Lovatti B.P.O., Nascimento M.H., Rainha K.P., Oliveira E.C.S., Neto Á.C., Castro E.V.R., Filgueiras P.R. 2020. ‘Different strategies for the use of random forest in NMR spectra’. Journal of Chemometrics. 34:e3231. https://doi.org/10.1002/cem.3231

Maechler M., Rousseeuw P., Struyf A., Hubert M., Hornik, K. 2023. 'cluster: Cluster Analysis Basics and Extensions'. R package version 2.1.6.

Magdas D.A., Pirnau A., Feher I., Guyon F., Cozar B.I. 2019. 'Alternative approach of applying ^{1}H NMR in conjunction with chemometrics for wine classification'. Lebensm. Wiss. Technol. 109, pp. 422–428. https://doi.org/10.1016/j.lwt.2019.04.054

Mahalanobis P.C. 1936. 'On the Generalized Distance in Statistics'. National Institute of Science of India, 2, pp. 49–55. https://link.springer.com/article/10.1007/s13171-019-00164-5.

Mallarino A.P., Oyarzabal E.S., Hinz P.N. 1999. ‘Interpreting within-field relationships between crop yields and soil and plant variables using factor analysis.’ Precis. Agric., 1. pp.15–25.

Marascuilo L.A. 1966. ‘Large-sample multiple comparisons'. Psychological Bulletin, 65, pp. 280–290. https://doi.org/10.1037/h0023189.

Martelo-Vidal M.J., Vázquez M. 2016. 'Polyphenolic Profile of Red Wines for the Discrimination of Controlled Designation of Origin'. Food Anal. Methods 9, 332–341. https://doi.org/10.1007/s12161-015-0193-8

Mascellani A., Hoca G., Babisz M., Krska P., Kloucek P., Havlik J. 2021. '^{1}H NMR chemometric models for classification of Czech wine type and variety'. Food Chem. 339, 127852. https://doi.org/10.1016/j.foodchem.2020.127852

Masetti O., Sorbo A., Nisini L. 2021. 'NMR Tracing of Food Geographical Origin: The Impact of Seasonality, Cultivar and Production Year on Data Analysis'. Separations 8, 230. https://doi.org/10.3390/separations8120230.

Mathieu A., Dumont Y., Chiroleu F., Duyck P. F., Flores O., Lebreton G., Reynaud B., Quilici S. 2014. ’Predicting the altitudinal distribution of an introduced phytophagous insect against an invasive alien plant from laboratory-controlled experiments: case of *Cibdela janthina* (Hymenoptera: Argidae) and *Rubus alceifolius* (Rosaceae) in La Réunion’. BioControl, 59, pp. 461–471. DOI: 10.1007/s10526-014-9574-y

Mazzei P., Francesca N., Moschetti G., Piccolo A. 2010. 'NMR spectroscopy evaluation of direct relationship between soils and molecular composition of red wines from Aglianico grapes'. Anal. Chim. Acta 673, pp. 167–172. https://doi.org/10.1016/j.aca.2010.06.003

McClure C.K. 2017. 'Structural Chemistry Using NMR Spectroscopy, Organic Molecules'. In Encyclopedia of Spectroscopy and Spectrometry, Harmadik kiadás, Lindon J.C., Tranter, G.E., Koppenaal D.W., Eds. Academic Press: Oxford, UK, pp. 281–292. https://doi.org/10.1016/B978-0-12-803224-4.00294-6

Meila M. 2007. 'Comparing clusterings-an information based distance.' Journal of Multivariate Analysis, 98(5) pp. 873–895. doi:10.1016/j.jmva.2006.11.013.

Mellin W.D. 1957. ‘Work With New Electronic 'Brains' Opens Field For Army Math Experts’. The Hammond Times. November 10, 1957. p. 65.

Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F. 2023a. 'e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien'. R package version 1.7-14, <https://CRAN.R-project.org/package=e1071>.

Meyer D., Zeileis A., Hornik K., Friendly M. 2023b. 'vcd: Visualizing Categorical Data'. R package version 1.4-12, https://CRAN.R-project.org/package=vcd.

Millar R.B., Anderson M.J. 2004. 'Remedies for pseudoreplication', Fisheries Research, 70(2–3) pp. 397–407 https://doi.org/10.1016/j.fishres.2004.08.016.

Minoja A.P., Napoli C. 2014. 'NMR screening in the quality control of food and nutraceuticals'. Food. Res. Int. 2014, 63, pp. 126–131. https://doi.org/10.1016/j.foodres.2014.04.056

Mitchell T.M. 1996. 'Machine learning'. McGraw-Hill, ISBN: 0070428077 p. 432

Moallem Z., Karimi-Malati A., Sahragard A., Zibaee A. 2017. 'Modeling Temperature-Dependent Development of *Glyphodes pyloalis* (Lepidoptera: Pyralidae)'. Journal of Insect Science, 17(1) 37, https://doi.org/10.1093/jisesa/iex001

Molnar C. 2022. 'Interpretable Machine Learning is a comprehensive guide to making machine learning models interpretable' ISBN 979-8411463330. p. 328. Independently published, Creative Commons Licensed

Monakhova Y.B., Godelmann R., Kuballa T., Mushtakova S.P., Rutledge D.H. 2015. 'Independent components analysis to increase efficiency of discriminant analysis methods (FDA and LDA): Application to NMR fingerprinting of wine'. Talanta, pp. 141, pp. 60–65. https://doi.org/10.1016/j.talanta.2015.03.037.

Monkahova Y.B., Schäfer H., Humpfer E., Spraul M., Kuballa T., Lachenmeier, D.W. 2011. 'Application of automated eightfold suppression of water and ethanol signals in 1H NMR to provide sensitivity for analyzing alcoholic beverages'. Magn. Reson. Chem. 49, pp. 734–739. https://doi.org/10.1002/mrc.2823

Morais C.L.M., Santos M.C.D, Lima K.M.G., Martin F.L. 2019. 'Improving Data Splitting for Classification Applications in Spectrochemical Analyses Employing a Random-Mutation Kennard-Stone Algorithm Approach' Bioinformatics, 35(24) pp. 5257–5263.

Motulsky H. Christopoulos A. 2003. 'Fitting Models to Biological data using Linear and Nonlinear Regression. A practical guide to curve fitting'. Graph pad Software Inc., san Diego CA, url

Muggeo V.M.R. 2003. 'Estimating regression models with unknown break-points'. Statist. Med., 22 pp. 3055–3071. https://doi.org/10.1002/sim.1545

Nahhas R.W. 2004 (in press) 'Introduction to Regression Methods Using R With Examples from Public Health'. CRC Press. https://www.bookdown.org/rwnahhas/RMPH/

Naumova E.N., Must A., Laird N.M. 2001. 'Tutorial in Biostatistics: Evaluating the impact of 'critical periods' in longitudinal studies of growth using piecewise mixed effects models'. International Journal of Epidemiology, 30(6) pp. 1332–1341. https://doi.org/10.1093/ije/30.6.1332

Németh M. 1966. 'Borszőlőfajták határozókulcsa'. Mezőgazdasági Kiadó: Budapest, Hungary, p. 240.

Neyman J., és Pearson E.S. 1933. 'On the Problem of the Most Efficient Tests of Statistical Hypotheses'. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231(694–706), pp. 289–337. https://doi.org/10.1098/rsta.1933.0009

Nisbet R., Miner G., Yale K. 2018. 'Handbook of Statistical Analysis and Data Mining Applications', Második kiadás; Academic Press: Cambridge, MA, USA, 2018; ISBN 9780124166325.

O.I.V., 2009a. 'OIV descriptor list for grape varieties and Vitis species'. O.I.V. 18, rue d'Aguesseau – 75008 Paris. 2. kiadás.

O.I.V., 2009b. 'Description of world vine varieties'. O.I.V. 18, rue d'Aguesseau – 75008 Paris. 560.

O'Brien R.M. 2007. 'A Caution Regarding Rules of Thumb for Variance Inflation Factors'. Qual Quant 41, pp. 673–690. https://doi.org/10.1007/s11135–006-9018-6

O'Connor B.P. 2023. 'DFA.CANCOR: Linear Discriminant Function and Canonical Correlation Analysis'. R package version 0.2.8, <https://CRAN.R-project.org/package=DFA.CANCOR>.

Pachú J.K.S., Malaquias J.B., Godoy W.A.C., Ramalho F.S., Almeida B.R., Rossi F. 2018. 'Models to describe the thermal development rates of *Cycloneda sanguinea* L. (Coleoptera: Coccinelidae)'. Journal of Thermal Biology, 73 pp. 1–7, https://doi.org/10.1016/j.jtherbio.2018.01.006.

Pálfai I. 2002. 'Probability of drought occurrence in Hungary'. Quarterly J. Hungarian Meteorological Service, 106(3-4) pp. 265–275.

Papotti G., Bertelli D., Graziosi R., Silvestri M., Bertacchini L., Durante C., Plessi, M. 2013. 'Application of One and two-dimensional NMR spectroscopy for the characterization of Protected Designation of Origin Lambrusco wines of Modena'. J. Agric. Food Chem. 61, pp. 1741–1746. https://doi.org/10.1021/jf302728b.

Pearson E. S. D'Agostino R.B. és Bowman K.O. 1977. 'Tests for departure from normality, comparison of powers'. Biometrika, 64, pp. 231–246.

Pereira G.E., Gaudillere J.P., Van Leeuwen C., Hilbert G., Lavialle O., Maucourt M., Deborde C., Moing A., Rolin, D. 2005. '^{1}H NMR and chemometrics to characterize mature grape berries in four wine-growing areas in Bordeaux', France. J. Agric. Food Chem. 53, pp. 6382–6389. https://doi.org/10.1021/jf058058q.

Ping J.L., Green C.J., Bronson K.F., Zartman R.E., Dober-mann A. 2004. 'Identification of relationships between cot¬ton yield, quality, and soil properties.' Agron. J., 96(6) pp. 1588–1597.

Pinheiro J.C., Bates D.M. 2000. 'Mixed-Effects Models in S and S-PLUS'. Springer, New York. ISBN: 0-387-98957-9 p. 528. doi:10.1007/b98882 <https://doi.org/10.1007/b98882>

Pinheiro J.C., Bates D.M., R Core Team 2023. 'nlme: Linear and Nonlinear Mixed Effects Models'. R package version 3.1-164, <https://CRAN.R-project.org/package=nlme>.

Podani J. 1997. 'Bevezetés a többváltozós biológiai adatfeltárás rejtelmeibe'. Scientia Kiadó, ISBN 963 8326 06 9 p. 411

Podani J. 1999. 'Extending Gower's general coefficient of similarity to ordinal characters'. Taxon. 48 (2) pp. 331–340. doi:10.2307/1224438. JSTOR 1224438.

Powers D.M.W. 2011. 'Evaluation: From Precision, Recall and F-Measure to ROC; Informedness, Markedness and Correlation'. J. Mach. Learn. Technol. 2, pp. 37–63. https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf

R Core Team 2023. 'R: A Language and Environment for Statistical Computing'. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Rand W. M. 1971. 'Objective criteria for the evaluation of clustering methods'. Journal of the American Statistical Association. 66(336). American Statistical Association. pp. 846–850. doi:10.2307/2284239. JSTOR 2284239.

Rao R.C. 1948. 'The utilization of multiple measurements in problems of biological classification'. J. R. Stat. Soc. Ser. B 10, pp. 159–203.

Rebaudo F., Rabhi, V.-B. 2018. 'Modeling temperature-dependent development rate and phenology in insects: review of major developments, challenges, and future directions'. Entomol Exp Appl, 166. pp. 607–617. https://doi.org/10.1111/eea.12693

Régnier B., Legrand J., Rebaudo F. 2022. 'Modeling Temperature-Dependent Development Rate in Insects and Implications of Experimental Design'. Environmental Entomology, 2022, 51 (1), pp.132–144. DOI: 10.1093/ee/nvab115.hal-03982548

Revelle W. 2024. 'psych: Procedures for Psychological, Psychometric, and Personality Research'. Northwestern University, Evanston, Illinois. R package version 2.4.3, <https://CRAN.R-project.org/package=psych>.

Ridgeway G., Developers G. 2024. 'gbm: Generalized Boosted Regression Models'. R package version 2.2.2, <https://CRAN.R-project.org/package=gbm>.

Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J-S., Müller M. 2011. 'pROC: an open-source package for R and S+ to analyze and compare ROC curves'. BMC Bioinformatics, 12, p. 77. DOI:10.1186/1471-2105-12-77 url

Robinson D., Hayes A., Couch S. 2023. 'broom: Convert Statistical Objects into Tidy Tibbles'. R package version 1.0.5, <https://CRAN.R-project.org/package=broom>.

Rousseeuw P.J. 1987. 'Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis'. Computational and Applied Mathematics. 20 pp. 53–65. doi:10.1016/0377-0427(87)90125-7.

Roy M., Brodeur J., Cloutier C. 2002. 'Relationship Between Temperature and Developmental Rate of *Stethorus punctillum* (Coleoptera: Coccinellidae) and Its Prey *Tetranychus mcdanieli* (Acarina: Tetranychidae)'. Environmental Entomology, 31(1), pp. 177–187, https://doi.org/10.1603/0046-225X-31.1.177

Rumelhart D.E., Hinton G.E., Williams R.J. 1986. 'Learning Representations by Back-Propagating Errors'. Nature 323(6088) pp. 533–36. doi: 10.1038/323533a0.

Rutherford A. 2011. 'Anova and Ancova – A GLM Approach' Wiley. Második kiadás ISBN 978-0-470-38555-5 p. 344.

Santos H.T., Marchioro C.A. 2021. 'Selection of models to describe the temperature-dependent development of *Neoleucinodes elegantalis* (Lepidoptera: Crambidae) and its application to predict the species voltinism under future climate conditions'. Bulletin of Entomological Research. 111(4) pp. 476–484. doi:10.1017/S0007485321000195

Schwarz G. 1978. 'Estimating the dimension of a model', Annals of Statistics 6(2) pp. 461–464. DOI: 10.1214/aos/1176344136

Scikit. 2021. 'Https://Scikit-Learn.Org/'.

Sebastien L., Josse J., Husson F. 2008. 'FactoMineR: An R Package for Multivariate Analysis'. Journal of Statistical Software, 25(1) pp. 1–18. 10.18637/jss.v025.i01

Sharpe D. 2015. 'Chi-Square Test is Statistically Significant: Now What?', Practical Assessment, Research, and Evaluation 20(1): 8. doi: https://doi.org/10.7275/tbfa-x148

Shi P., Ge F., Sun Y., Chen C. 2011a. 'A simple model for describing the effect of temperature on insect developmental rate'. J. Asia Pac. Entomol. 14. pp. 15–20. https://doi.org/10.1016/j.aspen.2010.11.008

Shi P., Ikemoto T., Egami C., Sun Y., Ge F. 2011b. 'A modified program for estimating the parameters of the SSI model'. Environ. Entomol. 40. pp. 462–469. https://doi.org/10.1603/EN10265

Shukla M.K., Lal R., Ebinger M. 2004. 'Principal compo¬nent analysis for predicting corn biomass and grain yields.' Soil Sci., 169(3) pp. 215–224.

Simmler C., Napolitano J.G., McAlpine J.B., Chen S.N., Pauli G.F. 2014. 'Universal quantitative NMR analysis of complex natural samples'. Curr. Opin. Biotechnol. 25, pp. 51–59. https://doi.org/10.1016/j.copbio.2013.08.004

Sing T., Sander O., Beerenwinkel N., Lengauer T. 2022. 'ROCR: Visualizing classifier performance in R'. Bioinformatics 21, 7881. url

Somers, R. H. 1962. 'A new asymmetric measure of association for ordinal variables'. American Sociological Review. 27 (6). doi:10.2307/2090408. JSTOR 2090408.

Son H.S., Ki M.K., Van Den Berg F., Hwang G.S., Park W.M., Lee C.H., Hong Y.S. 2008. '^{s}H nuclear magnetic resonance-based metabolomic characterization of wines by grape varieties and production areas'. J. Agric. Food Chem. 56, pp. 8007–8016. https://doi.org/10.1021/jf801424u.

Spearman C. 1904. 'The Proof and Measurement of Association between Two Things'. The American Journal of Psychology. 15 (1) pp. 72–101. doi:10.2307/1412159. JSTOR 1412159.

Spurgeon D.W. 2019. 'Common Statistical Mistakes in Entomology: Pseudoreplication' American Entomologist, 65(1) pp. 16–18, https://doi.org/10.1093/ae/tmz003

Spyros A. 2016. 'Application of NMR in food analysis'. In Specialist Periodical Reports: Nuclear Magnetic Resonance, Ramesh V. Ed., London RSC: London, UK, 45, pp. 269–308. https://doi.org/10.1039/9781782624103-00269

Spyros A., Dais P. 2012. 'NMR Spectroscopy in Food Analysis', Cambridge RSC: Cambridge, UK, p. 329. ISBN 978-1-84973-175-1. https://doi.org/10.1039/9781849735339 PDF ISBN: 978-1-84973-533-9 EPUB ISBN: 978-1-78262-566-79.

Stecanella B. 2021. 'Support Vector Machines (SVM) Algorithm Explained' https://monkeylearn.com/blog/introduction-tosupport-vector-machines-svm

Stenberg B. 1998. 'Soil attributes as predictors of crop production under standardized conditions.' Biol. Fert. Soils, 27, pp. 104–112.

Strobl C., 2005. 'Variable Selection in Classification Trees Based on Imprecise Probabilities'. In ISIPTA Vol. 5, pp. 340–348.

Strobl C., Boulesteix A.L., Augustin T. 2006. 'Unbiased split selection for classification trees based on the Gini Index'. Computational Statistics & Data Anal. 52, pp. 483–501. https://doi.org/10.1016/j.csda.2006.12.030

Strobl C., Boulesteix A.L., Kneib T., Augustin, T., Zeileis, A. 2008. 'Conditional variable importance for random forests'. BMC Bioinformatics 9, 307. doi:10.1186/1471-2105-9-307

Tabachnick B.G., Fidell L.S. 2019. 'Using multivariate statistics' Hetedik kiadás. Pearson. ISBN 9780134790541 p. 832

Tang F., Ishwaran H. 2017. 'Random Forest missing data algorithms'. Statistical Anal. and Data Mining 10, pp. 363–377.

Tharwat A. 2021. 'Classification assessment methods'. Appl. Comput. Inform. 17, pp. 168–192. Emerald Publishing Limited. e-ISSN: 2210-8327 p-ISSN: 2634-1964.DOI 10.1016/j.aci.2018.08.003https://www.emerald.com/insight/2210-8327.htm

Therneau T, Atkinson B 2023. 'rpart: Recursive Partitioning and Regression Trees'. R package version 4.1.23, <https://CRAN.R-project.org/package=rpart>.

Thioulouse J., Dray S., Dufour A., Siberchicot A., Jombart T., Pavoine S. 2018. 'Multivariate Analysis of Ecological Data with ade4'. Springer. doi:10.1007/978-1-4939-8850-1 <https://doi.org/10.1007/978-1-4939-8850-1>.

Tittonell P., Sheperd K.D., Vanlauwe B., Giller K.E. 2007. ‘Unravelling the effects of soil and crop management on maize productivity in smallholder agricultural systems of western Kenya - An application of classification and regres¬sion tree analysis.’ Agric. Ecosys. Environ., 123, pp. 137–150.

Tobias R. 1996. ‘An Introduction to Partial Least Squares Regression’. Semantic Scholar. url

Toda S., Murai T. 2007. 'Phylogenetic analysis based on mitochondrial COI gene sequences in *Thrips tabaci* Lindeman (Thysanoptera: Thripidae) in relation to reproductive forms and geographic distribution'. Appl. Entomol. Zool., 42 pp. 309-316. https://doi.org/10.1303/aez.2007.309

Tuszynski J. 2021. 'caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc'. R package version 1.18.2, <https://CRAN.R-project.org/package=caTools>.

UPOV. International Union for the Protection of New Varieties of Plants (2005) 'Guidelines for the conduct of tests for distinctness, uniformity and stability'. Apple. Technical Guideline TG/14/9. módosített verzió (2023): http://www.upov.int/edocs/tgdocs/en/tg014.pdf

Venables W. N., Ripley B. D. 2002. 'Modern Applied Statistics with S'. Negyedik kiadás. Springer, New York. ISBN 0-387-95457-0

Viggiani L., Castiglione Morelli M.A. 2008. 'Characterization of wines by Nuclear Magnetic Resonance: A work study on wines from the Basilicata region in Italy'. J. Agric. Food Chem. 56, pp. 8273–8279. https://doi.org/doi.org/10.1021/jf801513u.

Ward J.H., Jr. 1963. 'Hierarchical Grouping to Optimize an Objective Function', Journal of the American Statistical Association, 58, pp. 236–244. https://doi.org/10.1080/01621459.1963.10500845

Warnes G.R., Gorjanc G., Magnusson A., Andronic L., Rogers J., MacQueen D., Korosec A. 2023. 'gdata: Various R Programming Tools for Data Manipulation'. R package version 3.0.0, <https://CRAN.R-project.org/package=gdata>.

Weihs C., Ligges U., Luebke K., Raabe N. 2005. 'klaR Analyzing German Business Cycles'. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). Data Analysis and Decision Support, 335-343, Springer-Verlag, Berlin.

Wenck S., Mix T., Fischer M., Hackl T., Seifert S. 2023. ‘Opening the Random Forest Black Box of 1H NMR Metabolomics Data by the Exploitation of Surrogate Variables’. Metabolites. 13(10):1075. https://doi.org/10.3390/metabo13101075

West S.G., Finch J.F., Curran P.J. 1995 'Structural equation models with nonnormal variables: problems and remedies'. In RH Hoyle (Ed.). Structural equation modeling: Concepts, issues and applications. Newbery Park, CA: Sage; pp. 56–75.

Wickham H. 2023. 'modelr: Modelling Functions that Work with the Pipe'. R package version 0.1.11, <https://CRAN.R-project.org/package=modelr>.

Wickham H., Averick M., Bryan J., Chang W., McGowan L.D., François R., Grolemund G., Hayes A., Henry L., Hester J., Kuhn M., Pedersen T.L., Miller E., Bache S.M., Müller K., Ooms J., Robinson D., Seidel D.P., Spinu V., Takahashi K., Vaughan D., Wilke C., Woo K., Yutani H. 2019. 'Welcome to the tidyverse.' Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686

Wickham H., François R., Henry L., Müller K., Vaughan D. 2023. 'dplyr: A Grammar of Data Manipulation'. R package version 1.1.4, <https://CRAN.R-project.org/package=dplyr>.

Wickham H. 2016. 'ggplot2: Elegant Graphics for Data Analysis'. Springer-Verlag New York.

Wickham H., Grolemund G. 2023. 'R for Data Science: Import, Tidy, Transform, Visualize, and Model Data'. O'Reilly Media, Inc. ISBN: 10-1492097403, 13-978-1492097402 p. 550 https://r4ds.hadley.nz/

Wright M.N., Ziegler A. 2017. 'ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R'. Journal of Statistical Software, 77(1), pp. 1–17. doi:10.18637/jss.v077.i01

Zhao L.L, Qiu X.J, Wang W.B, Li R.M, Wang D.S 2019. 'NMR Metabolomics and Random Forests Models to Identify Potential Plasma Biomarkers of Blood Stasis Syndrome With Coronary Heart Disease Patients'. Front. Physiol. 10:1109. doi: 10.3389/fphys.2019.01109

Zheng H., Chen L., Han X., Zhao X., Ma Y. 2009. 'Classification and regression tree (CART) or analysis of soybean yield variability among fields in Northeast China: The importance of phosphorus application rates under drought conditions.' Agric. Ecosys. Environ., 132, pp. 98–105.

11 KÖSZÖNETNYILVÁNÍTÁS

Szeretném kifejezni hálámat Harnos Zsoltnak, aki PhD témavezetőm, és egyben tanszékvezetőm is volt. Az ő irányítása alatt olyan utat járhattam be, amely lehetővé tette számomra a folyamatos szakmai fejlődést, miközben a munka és a családi kötelezettségek közötti egyensúly megteremtésében mindig támogatóan állt mellettem. Mély szomorúsággal tölt el, hogy már nem lehet közöttünk.

Külön köszönetemet szeretném kifejezni a Magyar Agrár- és Élettudományi Egyetem Matematika és Természettudományi Alapok Intézetének vezetőjének, Székely Lászlónak, valamint helyettesének, Veres Antalnak. A velük való együttműködés szakmai életem számára rendkívül gazdagító és örömteli tapasztalat volt. Az ő vezetésük és támogatásuk lehetőséget biztosított számomra, hogy rugalmasan és elmélyülten végezhessem munkámat, ami jelentős mértékben hozzájárult hatékonyságom növeléséhez.

Közvetlen kollégáim, az Alkalmazott Statisztika Tanszék munkatársai között dolgozni igazi kiváltság. Jelenlétükben békés, barátságos és segítőkész légkör vesz körül, amely még a legnehezebb napokat is széppé varázsolja.

Szeretném megköszönni 360-at meghaladó számú társszerzőmnek az élményekben gazdag, inspiráló közös munkát. És bár felsorolni őket egyenként e helyen nincs lehetőségem, közös felfedezéseink emlékezetesek számomra. Rengeteget tanultam tőlük, tudások sokszor lenyűgözött, ami folyamatosan motivált engem.

A Magyar Agrár- és Élettudományi Egyetem Kutatási Kiválóság Programjában való részvétel megtisztelő lehetőség volt számomra, ami ösztönzőleg hatott rám.

Férjem és három felnőtt fiam biztatása, valamint lelkesítő érdeklődése óriási erőforrást jelentett. Odaadó szeretetük, nem fogyatkozó türelmük és tapintatos figyelmük nélkül ez a munka nem születhetett volna meg.