

**Statisztikai modellezési módszerek az
agrártudományi kutatásokban**

című doktori mű
rövid összefoglalása (TÉZIS)



Ladányi Márta

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

Budapest, 2024.

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 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.

1.4 A dolgozat szerkezete a célkitűzésemet 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 körütekintő 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.

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 közzlése és vizualizációja

A modellező munka lépéseit a dolgozatban részletesen ismertetem. Az adatelőkészítés, adattisztítás és vizualizációkülönböző kihívásait és megoldási lehetőségeit tizenkét példával szemléltetem, melyek mindegyikét saját publikációból idézem.

3 LINEÁRIS MODELLEK

Az agrártudományi kutatásokban gyakran alkalmazott lineáris modellek rövid bemutatása során azok alkalmazásának lehetőségeire, adategényeire, feltételeire helyezem a hangsúlyt, és hat saját publikációból vett öt példával világítom meg a modellek sokszínűségét.

Az általános lineáris modelleken belül szó esik egy- és többtényezős ANOVA, egyszerű és többszörös lineáris regresszió, kovarianciaanalízis, többváltozós ANOVA (MANOVA) modellekről, kevert modellekről, hierarchikus modellekről, ismételt mérések modellekről, marginális modellekről, szakaszos lineáris modellekről, válaszfelület-modellekről. Kitekintünk a nemlineáris modellek irányába, valamint az általánosított lineáris modellek felé is.

4 A GÉPI TANULÁS

4.1 A gépi tanulás fogalma

A gépi tanulási módszerekkel kicsit bővebben foglalkozunk. Kitérünk a modellépítés lépéseire, a validáció folyamatára, a

paraméterek becslésének módjára, a becslések értékelésére, valamint a becslések, illetve modellek összehasonlítására. Hangsúlyozzuk a gépi tanulási és a hagyományos statisztikai módszerek szoros kapcsolatát. Öt példával szemléltjük ezeket, melyek tizenegy saját publikációból származnak.

A gépi tanulási módszerek agrártudományi alkalmazása napjainkban robbanásszerű fejlődésen megy keresztül. Ahogy a múlt században már igen mélyen kidolgozott statisztikai módszerek voltak a kezünkben, csak éppen az adatok előállítása, mérése, elemzése volt nehézkes, úgy ma már a kifinomult műszerek, a távérzékelés, az internet lehetősége, valamint a csaknem mindenki számára elérhető számítógép-kapacitás az esetenként elképesztő mennyiségben létrejövő adathoz való hozzájutást, és azok feldolgozását is lehetővé teszik. Ehhez járulnak hozzá a már régóta meglévő statisztikai módszereken alapuló, azokból kiinduló, de azok lehetőségeit jócskán meghaladó gépi tanulási módszerek. Ezek alkalmazása olyan páratlan lehetőség, amit az agrártudományi kutatásokban késedelem nélkül meg kell ragadnunk.

A gépi tanulás (machine learning) olyan gépi algoritmus, amely automatikus fejlődésre képes az adatok egyre mélyebb megismerése által. 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 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.

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.

A *felügyelt gépi tanulás* adathalmaza a megfigyelt elemekről nemcsak ismérvváltozókat tartalmaz, hanem egy, az ismérvektől függő válaszváltozót is. Ez a válaszváltozó lehet bináris,

nominális, ordinális, vagy folytonos mérési szintű, mely lényegében 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 becselő függvény, pontosabban egy becselő függvény paraméterhalmaza (*hyperparameters*). A becselő 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. 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.

A tesztelő adathalmazon vizsgálhatjuk a modell jóságát, amikor is a modell az előbbieken 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 lényegét 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*). 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.

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.

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.

4.3 A gépi tanulási módszerek feladatai

A gépi tanulás az agrártudományi kutatásokban az alábbi leggyakoribb alapeladatokra használható:

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

A *regresszió* ún. becselő 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 ériék el az előrejelzésben.

Az *osztályozás* a bemeneti jellemzőket egy diszkrét (faktor típusú) kimeneti változóra képezi. Bináris osztályozás esetében a kimeneti változó csak 0 vagy 1 lehet. Többszintű osztályozáskor a kimeneti változó több osztályból állhat. Az osztályozás a hasonlóság, illetve különbözőség valamely definíciója alapján a

hasznló 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.

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.

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.

4.4 A gépi tanulási módszerek lépései

A gépi tanulási projekt, mint minden modellező munka, az adatok gyűjtésével és azok előkészítésével kezdődik. Ennek utolsó lépése a gépi tanulásra speciálisan jellemző lépés. Az adathalmazt tanuló, illetve tesztelő adathalmazra osztjuk fel.

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. Szintén fontos, hogy a tanuló adathalmaz (is) reprezentatív legyen. Amennyiben a válaszváltozó diszkrét, azaz osztályozási feladatunk van, akkor 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*). 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 teszhalmazon 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.

4.5 A legfontosabb gépi tanulási módszerek

A gépi tanulási algoritmusok kutatásainkban alkalmazott legfontosabbjait, azok típusai, illetve feladatai szerint az 1. Táblázatban foglaltam össze. Itt csak azokat említem meg, melyeket kutatásom során használtam, és dolgozatom ‘Gépi tanulási módszerek’ fejezetében részletesen bemutatam.

1. Táblázat A legfontosabb gépi tanulási módszerek

Típus	Feladatok				
	Osztályozás	Becslés	Csoportosítás	Többcélú	Adatredukció
Felügyelt	K-legközelebbi szomszéd (K-NN) (lineáris, Naiv Bayes (NB) nemlineáris, Klasszifikációs döntési fák (DT) (LR, NLR, Logisztikus regresszió (LogR) Regressziós Diszkriminancia-analízis (lineáris, nemlineáris, DA, LDA, QDA)			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 varianciarányad (R^2), vagy az átlagos abszolút hiba (*Mean Absolute Error*, MAE) alapján állapítjuk meg.

A klasszifikációs modellek jóságát a legegyszerűbben a helyes, illetve helytelen pozitív, illetve negatív detektálás fogalmainak segítségével mérhetjük (2. Táblázat).

2. Táblázat A modellek jóságának mérőszámai

Elnevezés	Definíció	Angol elnevezés
Specifitás	$Sp = \frac{TN}{FP + TN}$	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	$P_{prev} = \frac{N}{TP + FN}$	Prevalence
Detekciós szint	$DR = \frac{N}{TP + FP}$	Detection Rate
Detekciós prevalencia	$DPrev = \frac{N}{TP + FN}$	Detection Prevalence
Pontosság	$A = \frac{TP + TN}{TP + FP + TN + FN}$	Accuracy
Kiegyensúlyozott pontosság	$BA = \frac{Se + Sp}{2}$	Balanced Accuracy
F1-szám	$F1 = \frac{2 \cdot PPV \cdot TPR}{PPV + TPR}$	F1 score
Cohen-féle kappa	$\frac{P(MM) - P(VM)}{1 - P(VM)}$	Cohen's Kappa

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

A *Cohen-féle kappa* statisztika 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 érünk el, ha minden elemet ebbe a csoportba sorolnánk, ami persze nem célunk

Ha a helyes pozitív arányt (érzékenységet) ($0 \leq \text{TPR} \leq 1$) ábrázoljuk a hibás pozitív arány ($0 \leq \text{FPR} \leq 1$, 1-specificitás) függvényében, akkor a modellek egy másik fontos értékelését, az ún. ROC-görbét (*receiver operating characteristic curve*) kapjuk. A ROC-görbe kiegyensúlyozatlan mintavételből származó adathalmaz esetén különösen informatív adaléka a modell értékelésének.

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. Minél magasabb az AUC értéke, annál jobb a modell elválasztó ereje.

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 tanulási 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 tanulási 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 tanulási 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 EREDMÉNYEK

Az Eredmények 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 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 aktivitássűrűsége (RAD) alapuló, dinamikus változó faktorú MANOVA alkalmazása, illetve (2) a speciális ordinális skálázás az aktivitássűrűségkülönbségre (ADD) vonatkozott (Flórián és mtsai, 2019, D1).
3. 16 éves nagyparcellás szántóművelés kísérlet adatait (kultúrnövény, talajművelés, növénytakaró), valamint csapadékindikátorokat felhasználva a teljes adathalmazra és a kukoricakultúra részadathalmazra vonatkozóan a lefolyás- és talajvesztesség-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 talajvesztesség és a 64%-ot a lefolyás esetében. A modell segítségével azonosítani lehetett a lefolyást és a talajvesztiséget

- befolyásoló legfontosabb tényezőket. Ezekkel az eredményekkel igazoltam, hogy a véletlen erdő modell alkalmas a lefolyás és a talajvesztesség 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) terméshozamadatain, 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 hatásai 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, Vének és mtsai, 2020, Q1).
7. Megmutattam, hogy a véletlen erdő, illetve a támogatóvektor-gépek alkalmas módszerek a borok 1H 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 (Nyitrai Sándor 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.

A kiemelt eredményekben idézett saját társszerzős publikációk

- 1) Flórián N., **Ladányi M.**, Ittész A., Kröel-Dulay G., Ónodi G., Mucsi M., Szili-Kovács T., Gergőcs 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.**, Véték 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) Véték 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) Nyitrai 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

A példákban idézett saját társszerős 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://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., Némédi 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) Mirmazloun 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.**, Loredó 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](https://doi.org/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) Nyitrai 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-Nótn 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) Véték 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) Véték 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**