FA-ALAPÚ MÓDSZEREK ALKALMAZÁSA A VÁLLALATI JÖVEDELMEZŐSÉG VIZSGÁLATÁBAN: VÁLTOZÓSZELEKCIÓ ÉS MODELLSZEGMENTÁCIÓ[1]

HAJDU OTTÓ – FÁRÓ JENŐ
ELTE Gazdálkodástudományi Intézet

Tanulmányunk célja, hogy bemutasson két – a vállalati jövedelmezőség vizsgálatában ez idáig nem alkalmazott –, egyaránt fa-alapú modellezési technikát, majd ezeket ötvözve feltérképezze a vállalati jövedelmezőség kapcsolatrendszerét, megvizsgálja, hogy egyes változók jövedelmezőségre kifejtett hatása stabilan alakul vagy instabilitást mutat bizonyos változók mentén. A két algoritmust, nevezetesen a véletlen erdőket és a modell-alapú rekurzív partícionálást egy európai országok mezőgazdasági vállalataiból álló 24 ezres mintára alkalmazzuk. A véletlen erdők változófontossági mutatóinak segítségével azonosítjuk az EBITDA-marginnal mért jövedelmezőség szempontjából legfontosabb változókat, majd ezeket egy lineáris modellbe illesztjük, és vizsgáljuk az egyes változók marginális hatását, valamint azt, hogy ezek a hatások összefüggnek-e olyan vállalati jellemzőkkel, mint a múltbeli jövedelmezőség, az eszköz- és forrásszerkezet, a likviditás, az eladósodottság, a vállalat országa és az idő múlása. Modellünk alapvető konklúziója, hogy egyetlen globális (*pooled*) modell illesztése helyett érdemes a jelenséget 14 lokális modell segítségével leírni, ugyanis a paraméterek erős változékonyságot mutatnak a múltbeli jövedelmezőség, az eszköz- és forrásszerkezet, a vállalatméret, illetve a vállalat országa szerint. Szegmentált modellünk a vállalatra szabott értékelésen túl lehetővé tette azon vállalatok profilírozását, melyek komoly jövedelmezőségi visszaesést szenvedtek el a gazdasági világválság kirobbanását követően. Mindemellett a tanulmányban felhívjuk a figyelmet további gépi tanulási algoritmusok alkalmazásának lehetőségére.

Kulcsszavak: véletlen erdők, modell-alapú rekurzív partícionálás, változószelekció, modellszegmentáció

[1]Hajdu Ottó, egyetemi tanár, ELTE Gazdálkodástudományi Intézet, NJE Gazdálkodástudományi Kar. Fáró Jenő, mesterszakos hallgató, ELTE Gazdálkodástudományi Intézet, CIB Bank Zrt. A tanulmány az Emberi Erőforrások Minisztériuma ÚNKP-18-2 kódszámú Új Nemzeti Kiválóság Programjának támogatásával készült. E helyen mondunk továbbá köszönetet az ismeretlen lektornak, aki észrevételeivel hozzájárult a tanulmány jobbá tételéhez. E-mail: `hajdu@gti.elte.hu, farojeno96@gmail.com`. Beérkezett 2019. augusztus 30.

1 Bevezetés

A vállalati jövedelmezőséget és annak meghatározóit számos szerző tanulmányozta az utóbbi évtizedekben[2]. Egyes kutatók a profitok varianciáját igyekeztek különféle eljárásokkal komponensekre bontani és hozzárendelni különböző (pl. ország-, iparági, vállalati) hatásokhoz (Schmalensee, 1985; Rumelt, 1991; Schiefer & Hartmann, 2009; Hirsch & Schiefer, 2016; Zouaghi, Hirsch & Garcia, 2016; Chaddad & Mondelli, 2013). Mások a jövedelmezőséget meghatározó változók kapcsolatrendszerét vizsgálták panel adatokra illesztett regressziók segítségével – a jövedelmezőségre kifejtett marginális hatások szignifikanciáját és szignifikáns esetben irányát feltárandó (Dencic-Mihajlov, 2014; Pervan & Mlikota, 2013; Lazăr, 2016; Pratheepan, 2014; Nunes, Serrasqueiro & Sequeira 2009; Pantea, Gligor & Anis, 2014; Yazdanfar, 2013). Utóbbiak tanulmányaikban nem ritkán több modellt is becsültek, ugyanakkor azzal a feltételezéssel éltek, hogy egy – a teljes mintára illesztett – globális (*pooled*) modell hatékonyan képes leírni a jövedelmezőség kapcsolatrendszerét, és jól illeszkedik az adatokra. A szerzők vizsgálódása nem terjedt ki arra, hogy mintáik bizonyos jellemzők által definiált szegmenseiben esetleg eltérnek-e az általuk vélelmezett összefüggést leíró modell paraméterei a globálisan illesztett modell, illetve más szegmensek paraméterértékeitől. Kivétel ez alól Yazdanfar (2013), aki az általa vizsgált 4 iparágra 4 különböző modellt becsült. Az egyes kutatások jellemzőit az *1. táblázat* tartalmazza.

A *pooled* modell adekvát mivoltának hipotézise természetesen lehet helyes, különösen akkor, ha hasonló profilú vállalatok alkotják a tanuló mintát. Jelen tanulmány ugyanakkor egy meglehetősen diverz (az EU-15 országaiban, illetve Magyarországon bejegyzett mezőgazdaság, erdészet, halászat tevékenységi körébe eső vállalatokat tartalmazó) mintán vizsgálja a jövedelmezőséget meghatározó változók kapcsolatrendszerét, a kapcsolatrendszert leíró egyenletek paramétereinek stabilitását, a jövedelmezőség szempontjából legfontosabb változókat, illetőleg a gazdasági válság hatására bekövetkezett jövedelmezőségbeli elváltozások mintázatait.

Mindezek érdekében tanulmányunk a jövedelmezőségi kutatások terén új megközelítésnek számító fa-alapú algoritmusokhoz folyamodik. Bemutatja a CART[3]-algoritmust, az előrejelzését számos regressziós fára alapozó véletlen erdőket – felhívva a figyelmet az alkalmazás során fellépő anomáliákra, melyekre megoldást javasol a létező algoritmusok közül. Ezt követően rátér a fák adatvezéreltségét és a statisztikai-ökonometriai modellezés elmélet-alapúságát ötvöző modell-alapú rekurzív partícionálásra (Kopf, Augustin & Strobl, 2013), bemutatja annak működési elvét, valamint az elemzéseket támogató R statisztikai rendszerben való implementációját.

[2] A hatékonyság, jövedelmezőség, üzleti környezet és működési feltételek sokváltozós elemzéséről lásd: Hajdu (1987).

[3] Classification and Regression Trees

Módszer	Eredmény-változó	Vizsgált vállalatok	Megfi-gyelések száma [db]	Időszak	R^2 [%]
Fix hatású panel regresszió szektor dummykkal (Dencic-Mihajlov, 2014)	ROTA, OPM	A belgrádi tőzsdén jegyzett közép- és nagyvállalatok, kivéve a pénzügyi vállalatokat	432	2008-2011	46,5-49,5
Dinamikus panel regresszió (Pervan & Mlikota, 2013)	EBITDA / Árbevétel	Horvát élelmiszeripar és italgyártás nagy és középvállalkozásai	1059	1999-2009	NA
Fix hatású panel regresszió év és iparág-év hatásokkal (Lazăr, 2016)	ROA	A bukaresti tőzsdén jegyzett nem pénzügyi vállalatok	608	2000-2011	61,7-65,7
OLS-, fix- és véletlen hatású statikus panel modellek (Pratheepan, 2014)	ROA	Tőzsdén jegyzett SríLanka-i gyártó vállalatok	550	2003-2012	13,9-23,2
Statikus és dinamikus panel modellek éves dummykkal (Nunes, Serrasqueiro & Sequeira, 2009)	Működési eredmény / Összes eszköz	Portugál szolgáltató szektorbeli vállalatok	375	1999-2003	16,1-24,4
Fix és véletlen hatású panel regresszió (Pantea, Gligor & Anis, 2014)	ROA & ROE	A bukaresti tőzsdén jegyzett ipari cégek	770	1999-2012	NA
SUR-regresszió (Yazdanfar, 2013)	ROA	4 iparág: egészségügy, kiskereskedelem, szállítás, fémgyártás	12530	2006-2007	42,3-49,1

Megjegyzés: ROTA: Return on Total Assets, OPM: Operating Profit Margin, EBITDA: Earnings before Interest, Taxes, Depreciation and Amortization, ROA: Return on Assets, ROE: Return on Equity, SUR-regresszió: Seemingly Unrelated Regression.

1. táblázat. Kutatási eredmények és módszertanok összefoglalója. *Forrás:* saját szerkesztés.

1.1 Klasszifikációs és regressziós fák

A manapság növekvő népszerűségnek örvendő fa-modellek között megkülönböztetünk klasszifikációs és regressziós fákat a magyarázott változó lehetséges kimeneteinek száma (véges vagy végtelen sok lehetséges kimenetel) alapján. Jelen tanulmányban a magyarázni kívánt változó folytonosságát tekintve[4] a regressziós fák kerülnek alkalmazásra. Ezek egy olyan fa gráffal szemléltethetők, melynek minden egyes csúcsa[5] egy-egy magyarázó változót és annak értéke alapján hozott döntést reprezentál: a szóban forgó magyarázó változó értéke nem nagyobb vagy nagyobb, mint egy meghatározott érték. A döntéseket reprezentáló *node*-oknál a változók és azok vágási (*cut-off*) értékeinek meghatározása a változók értéktartományának partícionálását teszi szükségessé, amely aztán az első felosztás után rekurzívan folytatódik a felosztott mintán valamilyen megállási kritérium eléréséig. Innen ered a felosztást végző algoritmus elnevezése: rekurzív partícionáló algoritmus, melynek több változata is ismert. Közéjük tartozik a Breiman, Friedman, Olshen és Stone

[4]Erről bővebben az Adatok és modellezési stratégia c. fejezetben lesz szó.

[5]Erre az angol nyelvű szakirodalomban *node*-ként hivatkoznak, melyhez mi is tartjuk magunkat a tanulmány hátralévő részében.

nevéhez fűződő CART-algoritmus[6] (Razi & Athappilly, 2005), amely a „ha-akkor" döntési szabályok megkonstruálásakor a következőképpen jár el: a fa gyökerétől lefele haladva minden egyes *node*-nál minden egyes magyarázó változó esetében megkeresi a változó értéktartományának azon pontját (egyetlenegyet, mivel a CART bináris felosztásokat hajt végre), amely mentén való mintafelosztás minimalizál egy ún. zavarossági mérőszámot (*impurity measure*), amely regressziós változatban nem más, mint a létrejövő csoportokban az eredményváltozó egyedi értékeinek a csoportátlagtól vett négyzetes eltéréseinek összege, majd kiválasztja a legjobban szeparáló változót, amely minimalizálja a fenti mérőszámot az összes változó tekintetében, és amely mentén történő felosztás a legnagyobb csoporton belüli homogenitást és csoportok közötti heterogenitást eredményezi az eredményváltozó szempontjából (Loh, 2011; Razi & Athappilly, 2005; Prasad, Iverson & Liaw, 2006). A megközelítést egy felülről-lefelé haladó (*top-down*) mohó (*greedy*) algoritmusnak nevezzük, mivel a fa gyökerétől halad lefelé és minden lépésnél az aktuálisan legjobb felosztást választja. A létrejövő fa a leveleiben[7] rendre a döntési szabályok alapján oda kerülő megfigyelések eredményváltozó szerinti átlagával készít előrejelzést (Tibshirani et al., 2013).

1.2 Véletlen erdők

Az *ensemble*-módszerek, így a közéjük tartozó véletlen erdők a fentebb bemutatott individuális fák azon kedvezőtlen tulajdonságát hivatottak kezelni, mely szerint azok igen érzékenyen reagálnak a tanuló mintában végbement változásokra (Garge, Bobashev & Egglestone, 2013), melynek köszönhetően megváltoznak az általuk készített előrejelzések is. Ugyanakkor ezen fák átlagosan már pontos előrejelzést adnak, alátámasztva azon ötlet létjogosultságát, hogy az előrejelzést érdemes fák egész halmazára alapozni (Strobl, Malley & Tutz, 2009).

A nagyszámú fát tartalmazó véletlen erdőknek számos különböző változata létezik, melyek elkülöníthetők aszerint, hogy

- milyen módon jönnek létre az erdőt alkotó egyedi fák,
- milyen eljárással (visszatevéssel vagy visszatevés nélkül) generálunk új mintákat, melyeken az egyes fák tanulnak, illetve
- miként kerülnek aggregálásra az egyes fák adta előrejelzések (Boulesteix et al., 2012).

Az eredeti – Breiman által javasolt és nagy gyakorisággal alkalmazott – véletlen erdők algoritmus[8] a fenti dimenziók mentén az alábbiak szerint karakterizálható:

[6] A további algoritmusok (C4.5, QUEST, GUIDE, M5, CRUISE) iránt érdeklődő Olvasó figyelmébe Loh (2011, 2014) műveit ajánljuk.

[7] Olyan *node*-ok, melyekből már nem indul elágazás.

[8] Ez került implementálásra az R statisztikai rendszer *randomForest* nevű package-ében azonos névvel (R Core Team, 2019).

- a korábban már bemutatott CART-algoritmus felelős az egyedi fák létrehozásáért annyi módosítással, hogy minden egyes felosztásnál változók egy véletlenszerűen kiválasztott részhalmazából választja ki a leginkább szeparáló változót,
- minden egyes fa egy visszatevéses mintavétellel előállított bootstrap mintán jön létre (Boulesteix et al., 2012),
- az egyes fák előrejelzéseinek súlyozatlan átlaga képezi az erdő előrejelzését (Segal, 2004).

Fontos elméleti hozzájárulás a véletlen erdők alkalmazásával kapcsolatban, hogy Breiman (2001) eredményei alapján egy erdő átlagos négyzetes általánosítási hibájának[9] felső korlátja egyenesen arányos az erdőt alkotó fák előrejelzési hibái közötti súlyozott korrelációval, illetve az erdőt alkotó fák általánosítási hibájával. Ebből kifolyólag a véletlen erdők pontosságának két legfontosabb feltétele a fák közötti alacsony korreláció és az egyes fák alacsony hibája. Ezeket optimalizálandó az egyes fák metszés nélkül, maximális mélységig kerülnek növesztésre a pontosság érdekében, míg a korrelációt a bootstrap mintán való tanítás és a felosztásoknál alkalmazott véletlen változóhalmazok hivatottak csökkenteni (Segal, 2004). Ezzel ugyan lokálisan szuboptimális fák jönnek létre (nem feltétlenül az a változó kerül kiválasztásra, melynek értéktartománya mentén való felosztás lokálisan optimalizálná a felosztás kritériumát), viszont a globális teljesítmény javítható ezáltal (Strobl, Malley & Tutz, 2009). Ezen kívül az ilyen változók bekerülése az erdőbe értékes interakciós hatásokat tárhat fel, melyek egyébként kimaradnának a modellből (Strobl et al., 2008).

Breiman (2001) másik említésre méltó eredménye, hogy az általánosítási hiba a nagy számok erős törvénye alapján konvergál, azaz megfelelően nagy számú fa esetén elkerülhető a túlilleszkedés.

A véletlen erdőknek további számos kívánatos tulajdonságát emelik ki a vonatkozó szakirodalomban. Ezek szerint az eljárás

- nagyon jól kezeli a változók közötti kapcsolat nemlinearitását (Strobl et al., 2007; Grömping, 2009)
- alkalmas az ún. „kicsi n – nagy p" probléma[10] kezelésére (Boulesteix et al., 2012; Ishwaran, 2007; Grömping, 2009)
- erősen korrelált változók esetén is jól alkalmazható (Strobl et al., 2008; Boulesteix et al., 2012)
- nagyon pontos előrejelzést tesz lehetővé (Strobl et al., 2008; Segal, 2004; Grömping, 2009)

[9] Az előrejelzett értékek tényleges értékektől vett négyzetes eltérésének feltételes várható értéke.

[10] A probléma abban áll, hogy a magyarázó változók p száma (akár nagyságrendekkel is) meghaladja a megfigyelések n számát, amely problémát a regresszióanalízis nem képes kezelni.

- robusztus az outlierekre nézve (Breiman, 2001)
- az egyes fák létrehozásához fel nem használt megfigyelések (OOB – *out of the bag*) belső validációs adathalmazként működnek, melyekkel becsülhető az előrejelzés hibája (Breiman, 2001; Segal, 2004; Grömping, 2009)
- változófontossági mérőszámokat szolgáltat (Segal, 2004; Grömping, 2009; Strobl et al., 2007; Strobl et al., 2008; Calle & Urrea, 2010; Boulesteix et al., 2012; Nicodemus, 2011), amelyek értékei alapján előálló rangsorok alkalmasak arra, hogy segítségükkel azonosítsuk a legfontosabb változókat és csökkentsük a probléma dimenzionalitását (Ishwaran, 2007).

A változófontossági mérőszámoknak számos szerző szentelt kitüntetett figyelmet a vonatkozó irodalomban (Grömping, 2009; Nicodemus et al., 2010; Louppe et al., 2013; Genuer, Poggi & Tuleau-Malot, 2010; Ishwaran, 2007; Nicodemus, 2011; Calle & Urrea, 2010; Strobl et al., 2007; Strobl et al., 2008), eredményeik azonban kizárólag a mérőszámok számítási módjának ismeretében értelmezhetőek. Egyikük az egyedi fák növesztésekor felosztási kritériumként funkcionáló zavarossági mérőszám értékében az egyes változók értékei mentén való felosztás hatására bekövetkezett javulást (a zavarosság csökkenését) veszi alapul és átlagolja ki ezen javulásokat (csökkenéseket)[11] (Louppe et al., 2013). A másik mutató leginkább permutációs fontossági mutató néven ismert. Az emögötti ráció, hogy egy változó értékeit véletlenszerűen permutálva az OOB-megfigyelések esetén az eredményváltozóval vett asszociáció (ha egyáltalán asszociált volt) megtörik (Strobl et al., 2007), így az eredeti állapothoz képest jelentős visszaesés következik be az előrejelzési pontosságban[12], amely regressziós esetben a tényleges értéktől vett átlagos négyzetes eltéréssel mérhető. Így a permutáció előtti és utáni előrejelzési pontosságok közötti különbségek kerülnek kiátlagolásra az erdőt alkotó fák számával (Grömping, 2009; Genuer, Poggi & Tuleau-Malot, 2010). A mutató – melynek nagy értéke (nagy romlás a permutáció után) eredetileg fontos változóra utal – gyakran használt rangsorolási célokra, előnyös vonásaként azt emelik ki, hogy nemcsak a változók individuális hatását, hanem a többi változóval vett interakciós hatását is figyelembe veszi (Strobl et al., 2007).

Ezeket a fontossági mutatókat az R statisztikai rendszerben könnyűszerrel ki tudjuk kalkulálni, ugyanakkor értelmezésükkor figyelemmel kell lennünk arra, hogy szimulációs tanulmányok tanúsága szerint számos hatás torzíthatja, illetve destabilizálhatja értéküket (Strobl et al., 2007; Strobl et al., 2008; Calle & Urrea, 2010; Nicodemus et al., 2010; Nicodemus, 2011). Az előbbi mutató például magán viseli az egyedi fák torzított változókiválasztásának

[11] Ezt a mutatót többféleképpen is hívják a vonatkozó irodalomban: átlagos zavarosságcsökkenés (*Mean decrease impurity*), átlagos Gini-csökkenés (*Mean decrease Gini*) (Louppe et al., 2013), illetve tisztaság-növekedésként is. Ez utóbbi jellemzi az R terminológiáját, ahol *IncNodePurity* néven jelenik meg (R Core Team, 2019).

[12] Erre reflektál, hogy a mutató átlagos pontosságcsökkenés (*Mean decrease accuracy*) néven is ismert (Calle & Urrea, 2010; Nicodemus, 2011). Az R implementációban *IncMSE* nevet viseli a mutató (R Core Team, 2019).

(a nagyszámú kategóriával rendelkező, illetve folytonos – összefoglalóan a sok különböző felosztást lehetővé tevő – változók mesterségesen preferáltak), illetve a visszatevéses bootstrap mintavétel torzító hatását. Ezt elkerülendő érdemes a változószelekciót hipotézisvizsgálati alapokra helyezni a torzítatlanság érdekében (ezáltal vezérelve fejlesztették ki Hothorn, Hornik & Zeileis (2006) a feltételes következtetési fákat[13], amely módszertan ún. permutációs tesztek p-értékén alapuló változószelekcióval tetszőleges mérési skálájú változók esetén képes elkerülni a torzítást, és amelynek R-beli alkalmazására a *party* package *ctree* nevű függvénye ad lehetőséget (R Core Team, 2019)), illetőleg visszatevés nélküli mintavétellel képzett részmintákon tanítani az erdőt alkotó fákat (Strobl et al., 2007). A fenti hatások a permutációs mutatót Strobl et al. (2007) szimulációja alapján csupán a variancia növelésén keresztül érintették, de nem torzították el. Eltorzítja viszont őket a magyarázó változók közötti erős korreláció (Nicodemus et al., 2010; Strobl et al., 2008), amely torzítás a releváns változókkal erősen korreláló, de az eredményváltozó szempontjából kevéssé releváns változók irányába mutat. Erre megoldásként Strobl et al. (2008) egy feltételes permutációs eljárást javasoltak, amely a magyarázó változók közötti korrelációs struktúra figyelembevételével – a permutációt csak feltételesen a többi változó rögzített $Z = z$ szintjén[14] elvégezve – vizsgálja az eredményváltozóra kifejtett hatást, ezáltal jobban tükrözi a tényleges hatást és megbízhatóbb eredményeket ad. Ez az eljárás az R statisztikai rendszerben a *varimp* függvény *conditional* logikai típusú argumentumának „igaz"-ra állításával érhető el (R Core Team, 2019).

1.3 Modell-alapú rekurzív partícionálás

A modell-alapú rekurzív partícionálás alapgondolata, hogy számos helyzetben nehezen tartható az a feltételezés, mely szerint egy globális modell illeszkedik az adatokra, ugyanakkor elképzelhető, hogy az adathalmaz egyes szegmenseire már jól illeszthető a modell (Zeileis, Hothorn & Hornik, 2008). Nem véletlen, hogy a strukturális változások vizsgálata – nevezetesen, hogy mikor és hogy változik meg az adatgeneráló mechanizmus – számos tudományterületen áll az érdeklődés középpontjában (Zeileis & Hornik, 2007). A medicinában például vizsgálják, hogy egy kezelés hatása miként függ össze a kezelt páciens egyedi jellemzőivel (pl. nem, kor stb.). Ezen jellemzők által a hasonló kezelési hatást mutató pácienscsoportok azonosítása fontos feladat, ugyanis a személyre szabott kezelés irányába mutat (Seibold, Zeileis & Hothorn, 2016). Úgy gondoljuk, hogy az orvostudományi példa analógiájára egy meglehetősen diverz vállalati mintában (6 év, 9 ország, 24 ezer megfigyelés[15]) is érdemes azzal a feltételezéssel élni, hogy a kezelés hatásához hasonlóan az egyes változók jövedelmezőségi hatása is változhat bizonyos változók mentén[16], ebből ki-

[13] Conditional inference trees

[14] A permutáció így egy ún. *conditioning grid* egyes szegmenseiben – a többi változó értékét rögzítve – megy végbe, amely *conditioning grid* minden egyes fánál az adott fa által meghatározott partíciót jelenti.

[15] Bővebben lásd az Adatok és modellezési stratégia c. fejezetben.

[16] Ezeket partícionáló változóknak hívjuk.

folyólag érdemes megkísérelni vállalati jellemzők által lehatárolt vállalati szegmensek azonosítását, melyekben a jövedelmezőségi hatások már stabilak. A modell-alapú rekurzív partícionálás épp ezt a stabilitást vizsgálja az alábbi lépésekben:

1. Globális modell illesztése az adatokra, melynek paramétereit az alkalmazott becslőfüggvény optimalizálásával lehet kiszámítani.

2. Annak értékelése, hogy a magyarázó változók (X-ek) paraméterei stabilak-e a partícionálóként megjelölt változók ($Z_1, Z_2, \ldots, Z_l$) tekintetében. Paraméter instabilitás detektálása esetén a legnagyobb instabilitást előidéző változó kiválasztása felosztásra, egyébként megállás.

3. A (2)-ben felosztásra kiválasztott változó értéktartománya mentén azon osztópont megkeresése, amely lokálisan optimalizálja a becslés során alkalmazott célfüggvény értékét.

4. A (3)-ban kiválasztott osztópont mentén az adathalmaz 2 részre (partícióra) osztása, majd a folyamat megismétlése (1)-től kezdve (Kopf, Augustin & Strobl, 2013; Zeileis, Hothorn & Hornik, 2008; Mozhgan et al., 2017; Zeileis & Hothorn, 2019).

Az eljárás előnyös vonásai közé tartozik, hogy ötvözi a fákat az elmélet által vezérelt modellezéssel, s ezáltal a tisztán adatvezérelt megközelítést finomítja a modell szegmentációjával, a szegmensek létrehozása során pedig a felírt összefüggés különböző változatait azonosítja (Kopf, Augustin & Strobl, 2013). Ezen kívül struktúrájából kifolyólag sokkal könnyebben értelmezhető, mint a magasabb rendű interakciók jelentése egy lineáris modellben (Seibold, Zeileis & Hothorn, 2016), számos parametrikus modell helyettesíthető R implementációjának (*mob*[17]) megfelelő argumentumába (*fit*), továbbá lehetővé teszi nemlineáris kapcsolatok modellezését, és a magyarázó változók közötti interakciók automatikus detektálását (Zeileis, Hothorn & Hornik, 2008). Az interakciók automatikus detektálása igen fontos vonása az algoritmusnak, ugyanis nem minden esetben van a priori hipotézisünk arra vonatkozóan, hogy melyek azok a szegmensei az adatállománynak, melyeken belül a paraméterek már stabilitást mutatnak (Seibold, Zeileis & Hothorn, 2016). Ezért mindezt meg kell tanulnia a rekurzív eljárásnak a mintából. Ennek során minden egyes iteráció második lépésében azt vizsgálja, hogy a becslőfüggvény[18] egyes paraméterek szerinti parciális deriváltjainak összetevői[19] véletlenszerűen ingadoznak 0 körül vagy szisztematikusan eltérnek tőle a partícionáló változók értékei mentén. Az eltérések nagyságrendje az ún. empirikus fluktuációs folyamat segítségével mérhető le, melyből tesztstatisztikák származtathatók mind folytonos, mind pedig kategóriakimenetű partícionáló változók esetén, melyek aszimptotikus eloszlásaiból már kalkulálhatók az instabilitási

[17] A függvény a *partykit* nevű package-ben található (R Core Team, 2019).

[18] Leggyakrabban OLS vagy log-likelihood.

[19] A derivált egy – a mintamérettel megegyező számosságú tagból álló – összeg, amelynek tagjai így párba állíthatók bármely más változó értékeivel.

tesztek p-értékei, melyek közül amennyiben a minimális az előre megadott szignifikanciaszint alá esik, akkor megtörténik a felosztás. További technikai részletekért az érdeklődő Olvasó Zeileis, Hothorn & Hornik (2008)-hoz, illusztratív példákért pedig Kopf, Augustin & Strobl (2013)-hoz fordulhat.

A tanulmány soron következő fejezeteiben bemutatásra kerül az adatállomány, a vizsgált pénzügyi probléma és az a keretrendszer, melyben a korábban bemutatott eljárások egymást támogatva, kiegészítve vesznek részt a vállalati jövedelmezőség modellezésében. Ezt követően a modellek eredményeinek értékelésével és az eredmények összegzésével, valamint további kutatási irányok megfogalmazásával zárul a tanulmány, melynek hozzájárulása a területhez kettős: azon kívül, hogy megismertette az Olvasóval a véletlen erdők és a modell-alapú rekurzív partícionálás eljárásait, célja még segítségükkel vállalatcsoport-specifikusan elvégezni az egyes változók jövedelmezőségi hatásának elemzését.

2 Adatok és modellezési stratégia

Az elemzéshez használt adatok az Amadeus adatbázisból származnak, ahonnét 30 ezernél is több vállalat ezer euróban mért adataihoz jutottunk hozzá. A belőlük való információszerzéshez a Microsoft Excel-t és az R statisztikai rendszert alkalmaztuk. Az adattisztítás keretein belül leszűrtük azon vállalatokat, melyeknél hiánytalanul rendelkezésre álltak az elemzéshez szükséges változók értékei a 2008-tól 2013-ig terjedő időszak minden egyes évében, kizártuk a mintából azon megfigyeléseket, ahol a jövedelmezőségi indikátorként kiválasztott EBITDA-margin[20] értéke abszolút értékben meghaladta az 1-et (100%-ot), majd ezen vállalatokból vettünk rétegzett mintát, úgy, hogy minden egyes évben véletlenszerűen kiválasztott 4000 vállalat került a mintába[21]. Ezen adatállomány szolgált az algoritmusok számára tanuló mintaként, ahol az EBITDA-margint az alábbi változókkal kívántuk modellezni, melyeknek egyúttal egy csoportosítását is megadjuk.

Az általunk alkalmazott egyszerűsített[22] eredménykimutatásból kalkulálható mutatók:

- X1: Árbevételarányos hozzáadott érték = Hozzáadott érték / Árbevétel, ahol Hozzáadott érték = Árbevétel – (ELÁBÉ + Anyagköltség)

[20] Az eredményváltozó ilyeténképpen történő kiválasztásának elsősorban az elemzés országhatárokon átívelő jellege miatt van jelentősége, ugyanis a mutató az EBITDA és az árbevétel hányadosaként áll elő, melyek közül előbbi a tényleges üzleti folyamatokat minősítő mutatóként lehetővé teszi, hogy az értékcsökkenési leírás elszámolása, a kamatkörnyezet vagy a helyi társasági adózás hatásai ne befolyásolják az eredményeket. Így a jövedelmezőségi indikátor kiválasztásakor a Pervan & Mlikota (2013) szerzőpároshoz hasonlóan járunk el.

[21] A vállalat-év megfigyelések országok szerinti megoszlását lásd a Melléklet 6. táblázatában.

[22] Az egyszerűsített jelző oka az Amadeus adatbázisból kinyerhető információk korlátossága, melyből fakadóan nem tudjuk teljes mértékben reprodukálni a vállalatok eredménykimutatását. Az általunk használt sémát lásd a Melléklet 3. táblázatában.

- X2: SZJR / Árbevétel – azt mutatja meg, hogy a vállalat a realizált árbevétel hány százalékának megfelelő összeget számolt el a munkavállalókhoz kapcsolódó ráfordításként
- X3: Növekedés = (Árbevétel$_t$ - Árbevétel$_{t-1}$) / Árbevétel$_{t-1}$ – az árbevétel dinamikáját jelzi

A beruházási tevékenység mutatói:

- X4: Nettó beruházás = Tárgyi eszközök és Immateriális javak záróértéke – nyitóérték + értékcsökkenési leírás. A mutató a tartós eszközök megújításáról informál, segítségével azt kívánjuk tesztelni, hogy van-e jövedelmezőségi hatása a beruházási tevékenységnek.
- X5: Nettó beruházás_1 – A Nettó beruházás késleltetett értéke. A beruházási tevékenység időben késleltetett hatásának tesztelését hivatott elősegíteni.

Késleltetett változók:

- X6: Likviditási ráta_1 = Forgóeszközök_1 / Rövid lejáratú kötelezettségek_1
- X7: Befektetett eszközök aránya_1 = Befektetett eszközök_1 / Mérlegfőösszeg_1
- X8: Idegen tőke aránya_1 = Idegen tőke_1 / Mérlegfőösszeg_1
- X9: Nettó adósság_1 = Teljes adósság – Pénzeszközök és pénzegyenértékesek, ahol Teljes adósság = Hosszú lejáratú kötelezettségek + Rövid lejáratú hitelek és kölcsönök
- X10: (Nettó adósság / EBITDA)_1 – azt mutatja meg, hogy a vállalat működéséből hány év alatt tudja kitermelni a nettó adósságát.
- X11: Cash-flow_1 – a cash-flow előző évi értéke
- X12: EBITDA_margin_1 – az eredményváltozó késleltetett értéke. Segítségével a profitok rövid távú perzisztenciáját (Hirsch & Gschwandtner, 2013) tanulmányozzuk.

Kategóriakimenetű változók:

- X13: Év = 2008; 2009; 2010; 2011; 2012; 2013[23]
- X14: Ország = HUN; POR; ESP; SWE; FIN; FRA; ITA; GER; BEL[24]

Vállalatméretet proxyzó változó:

[23] A változó kimeneteit *dummy* változókba is átkódoltuk, melyek kódolását lásd a Melléklet 7. táblázatában.

[24] A mintavételezés eredményeként bizonyos EU-15-ös országok kimaradtak az elemzésekből.

- X15: Árbevétel.

A fenti változókat és a korábban bemutatott algoritmusokat az alábbiak szerint fogjuk alkalmazni egy kétlépcsős modellezési eljárás keretein belül:

1. A véletlen erdőket az első 12 db változóval fogjuk lefuttatni feltételes következtetési fákkal mint erdőképzőkkel, majd 50 db futtatás után szemrevételezzük a fontossági mutatók eloszlását, hogy meggyőződhessünk azok stabilitásáról.

2. Az (1)-ben legfontosabbnak bizonyuló változókat további feldolgozásnak vetjük alá a modell-alapú rekurzív partícionálás keretein belül. Itt meghagyjuk azt a lehetőséget, hogy néhány kisebb fontosságú, de közgazdaságilag fontos mondanivalót tartogató változó is bekerüljön a modellezés második fázisába.

3. A (2)-ben előálló változók jelentik a modell-alapú rekurzív partícionálás magyarázó változóit, melyeket lineárisan szerepeltetünk az egyenletekben, míg a partícionáló változók körének kialakítása arra a hipotézisünkre reflektál, mely szerint a jövedelmezőséget leíró változók kapcsolatrendszere összefüggésben állhat a vállalatok eszköz- és forrásszerkezetével, likviditási rátával jellemezhető forgótőke-menedzsmentjével, eladósodottságával, múltbeli jövedelmezőségével és méretével. Ezen kívül a jövedelmezőségi hatások évenkénti és országonkénti megváltozását is elképzelhetőnek tartjuk. Ugyanakkor mindezek hagyományos módon való teszteléséhez számos interakciót kellene definiálnunk és megbecsülnünk, ami szükségtelen komplexitáshoz vezetne. Így a rekurzív eljárásra bízzuk, hogy tanulja meg az adatokból, melyek azok a csoportok, amelyeken belül hasonló a jövedelmezőségi változók kapcsolatrendszere. Tehát a regressziós fáktól eltérően nem az eredményváltozóban lévő mintázatok feltárása a cél, sokkal inkább a változók közötti asszociációban lévőké (Kopf, Augustin & Strobl, 2013), azaz modelleket kívánunk szegmentálni, nem pedig homogén jövedelmezőségi csoportokat.

Várakozásaink szerint a múltbeli jövedelmezőség fogja legnagyobb mértékben meghatározni a változók közötti kapcsolatrendszert. Mellette előzetesen nagy jelentőséget tulajdonítunk az Ország nevű változónak is országspecifikus jövedelmezőségi hatásokat vélelmezve. Arra számítunk továbbá, hogy a fentiek mellett a vállalatméret és az eladósodottság szintje is diszkrimináló erővel fog rendelkezni – különösen abban a tekintetben, hogy mekkora volt 2009-es év jövedelmezőségi visszaesése.

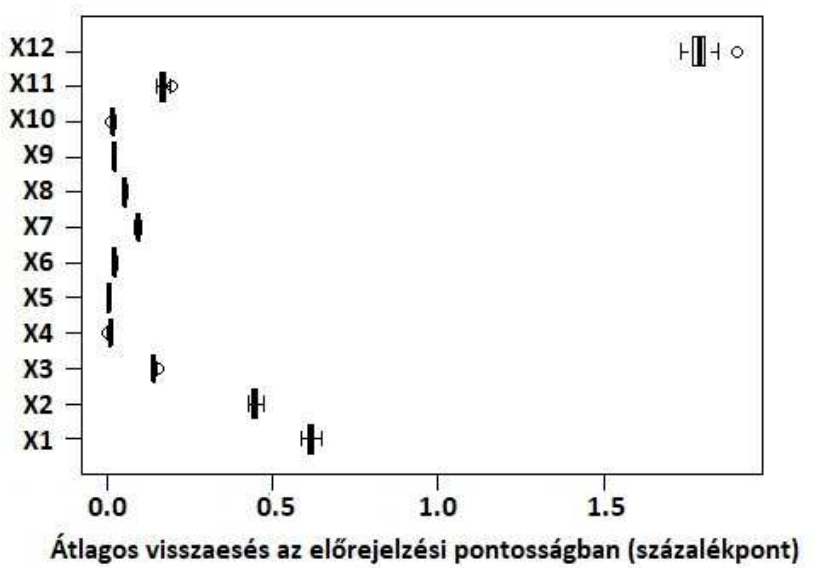
3 Eredmények és következtetések

A változófontossági mérőszámok előállításakor[25] az egyedi fákban a változószelekciót hipotézisvizsgálati alapokra helyeztük, tehát a *party* package-ben

[25]Az eredmények előállítása során futtatott R-kódok legfontosabb parancsait a Mellékletben követheti nyomon az Olvasó.

található *cforest* függvényt alkalmaztuk *ntree* = 100 db[26] egyedi fával, illetve *mtry* = 5 változóból[27] választotta ki az eljárás minden egyes felosztásnál a legkisebb p-értékkel rendelkezőt, és végre is hajtotta a felosztást, ha a p-érték kisebb volt, mint 1 – (*mincriterion* =) 0,95 = 0,05. Minden egyes fa megnövesztéséhez egy – az eredeti mintából visszatevés nélkül előálló – részminta szolgált alapul, míg a permutációs fontossági mutató kiszámításakor a kevésbé számításigényes, feltétel nélküli permutációs sémát alkalmaztuk, tekintve a magyarázó változók közötti alacsony korrelációt[28].

A fontosság megállapítása érdekében feltártuk a fontossági mutatók eloszlását 50 db futtatást követően, melyek igen alacsony változékonyságot mutattak, tehát a soron következő változók nem pusztán a véletlen folytán szerepelnek a legfontosabbak között.



1. ábra. A permutációs fontossági mutató eloszlása. *Forrás:* saját szerkesztés.

Mindezek alapján a legfontosabb változónak az EBITDA-margin késleltetett értéke (X12) bizonyult (a permutációját követően átlagosan több mint

[26]Itt, illetve más paramétereknél is ajánlatos addig próbálkozni a megfelelő beállítások keresésével, míg az érdeklődés középpontjában álló mennyiségek stabilizálódnak (Boulesteix et al., 2012, p. 496).

[27]Alapértelmezett érték. Megjegyzendő, hogy nagyszámú informatív változó esetén érdemes alacsonyra állítani a paraméter értékét, hogy a mérsékelt hatású változóknak nagyobb esélyt adjunk a fákban való megjelenésre, ezáltal jobban hasznosítva a rendelkezésre álló információt (Boulesteix et al., 2012, p. 496).

[28]A korrelációs mátrix (lásd a Melléklet 5. táblázatában) tanúsága szerint csupán minimális számú esetben fordul elő, hogy a változók között akár csak közepes erősségű korreláció lenne (csupán 4 esetben haladja meg abszolút értékben a 0,3-et), ezért nem tartjuk szükségesnek a számításigényes feltételes permutációs séma alkalmazását.

1,5 százalékponttal esett vissza az előrejelzés pontossága), melyet az Árbevételarányos hozzáadott érték (X1), az SZJR / Árbevétel (X2), az előző évi Cash-flow (X11) és a Növekedés (X3) követtek[29].

A fenti 5 változó mellett a beruházási tevékenység mutatóit tartottuk érdemesnek további feldolgozásra, illetve kiegészítettük őket éves *dummykkal*, melyek közül leginkább a 2009-es évet kódoló változó hatásának feltárásában voltunk érdekeltek, ugyanis e változó kapta a szerepet, hogy mérje le a gazdasági válság előidézte jövedelmezőségi visszaesést. Minden egyes változót lineárisan szerepeltettünk a modellben, ebből kifolyólag a modell-alapú rekurzív partícionálást annak egy speciális eseteként a *partykit* package-ben található *lmtree* függvény segítségével végeztük el (R Core Team, 2019).

A modell-alapú rekurzív partícionálás során alkalmazott magyarázó és partícionáló változók a Melléklet 4. táblázatában követhetők nyomon. A változók ilyeténképpen előálló szereposztásában a magyarázó változók rendre vagy már önmagukban *flow*-mutatók, vagy *flow*-mutatókból konstruált mutatók. Eközben a partícionálásra kijelölt változók inkább *stock*-mutatók, azon időszak elejéről[30], melyben a fenti *flow*-mutatók értékei realizálódtak.

Az *lmtree* függvény argumentumainak specifikálásakor a magyarázó- és partícionáló változói szerepek kiosztásán kívül az előmetszésről is gondoskodnunk kellett. Ennek érdekében az instabilitást mérő teszteknél alkalmazott szignifikanciaszintet 1%-ban ($alpha = 0{,}01$) határoztuk meg, a *node*-ba eső minimális megfigyelésszámot 1000-ben rögzítettük ($minsplit = 1000$), illetve Bonferroni-korrigált p-értékeket számítottunk ($bonferroni$ = TRUE).

Az algoritmus a legjobb felosztás kimerítő keresése miatt meglehetősen hosszan futott, melynek eredményeként egy 27 *node*-ból álló fa jött létre 14 levéllel, tehát 14 modellel írjuk le a jövedelmezőség alakulását egy darab globális modell illesztése helyett. A modellben csupán a döntési szabályokat szemügyre véve[31] adódnak az alábbi megállapítások:

- A kapcsolatrendszer erősen különbözik a múltbeli jövedelmezőség alapján
- Szintén csoportosító erővel rendelkezik a vállalati eszközszerkezet, melynek alapján különböző modell érvényes azon vállalatokra, melyeknél a Befektetett eszközök arányának nyitóértéke 40,3%-ot nem meghaladó, illetve meghaladó.

[29] Ezzel párhuzamosan lefuttattuk a CART-algoritmus létrehozta fákat tartalmazó *randomForest* függvényben implementált változatot is $ntree = 100$ db fával, melyek mindegyike egy visszatevés nélküli mintavétellel előállított részmintán jött létre, és minden egyes felosztásnál $mtry = 5$ db véletlenszerűen kiválasztott változóból került ki a legjobban szeparáló. Az itteni eredmények annyiban összhangban állnak a fentiekkel, hogy az ottani 5 legfontosabbnak mutatkozó változó mindkét típusú fontossági mutatónál az átlagos érték tekintetében a legfontosabb 6 változó között helyezkedett el, csupán a Nettó adósság / EBITDA mutató késleltetett értéke került be közéjük a permutációs fontossági mutatónál (*IncMSE*). Az előálló értékeket lásd a Melléklet 8. táblázatában.

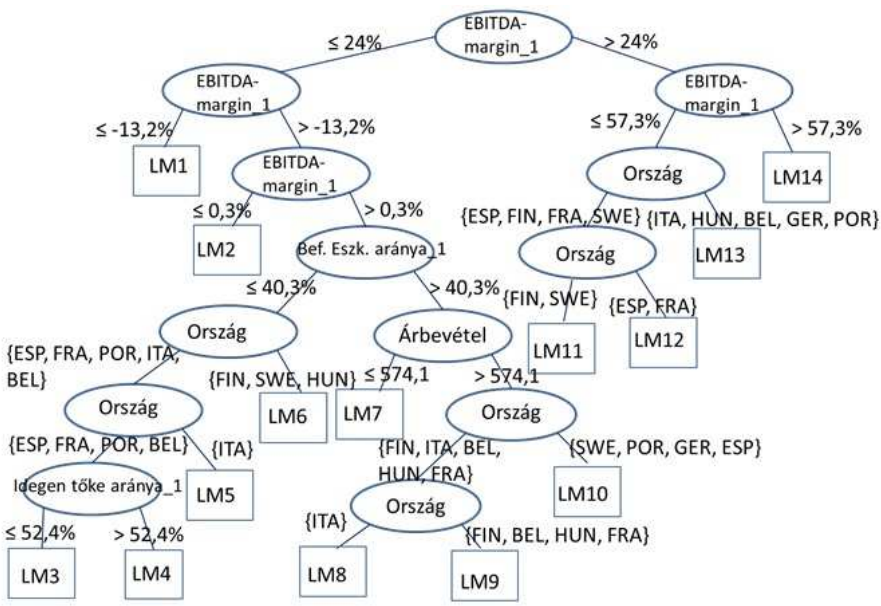
[30] Számviteli terminológiával élve nyitóértékekről van szó.

[31] A döntési szabályokat reprezentáló fa (lásd a 2. ábrán) minden egyes levelében egy-egy lineáris modell (LM) található.

- A vállalati forrásszerkezet is megjelenik a modellben mint a magyarázó változókkal interaktáló változó.

- Bizonyos vállalatoknál a vállalatméret is irányadó az együtthatók értékeinek homogenizálásában.

- Ezeken kívül az országok között is szóródnak az együtthatók. Az országok szerinti felosztás néhány esetben földrajzilag egymáshoz közel álló országokat (pl. a skandináv országokat) sorol egy csoportba, míg két esetben is elkülöníti az olasz vállalatokat, ezáltal náluk a többi országtól merőben különböző paraméterértékekre hívja fel a figyelmet.



2. ábra. A modell-alapú rekurzív partícionálás fája. *Forrás:* saját szerkesztés.

Most pedig vegyük szemügyre az egyes modellekben a változók együtthatóit[32], hogy leszűrhessük a bennük rejlő közgazdasági mondanivalót! Ennek során vegyük figyelembe, hogy az adott együttható által mért marginális hatás mely vállalati szegmensre érvényes!

[32] Az együtthatókat összefoglaló táblázatot lásd a 2. táblázatban.

	LM1	LM2	LM3	LM4	LM5	LM6	LM7	LM8	LM9	LM10	LM11	LM12	LM13	LM14
Konstans	-0.016	-0.004	0.019	0.006	-0.034***	-0.017***	0.042***	-0.006	0.016**	0.005	-0.084***	-0.036	0.072***	0.408***
Növekedés	0	0.088***	0.176***	-0.003	0.07***	0.024***	0.066***	0.035***	-0.006	0.041***	0.024***	0.036***	-0.006	-0.001*
Hozzáadott érték	0.115***	0.127***	0.084***	0.055***	0.134***	0.091***	0.175***	0.166***	0.1***	0.047***	0.403***	0.195***	0.046***	0.113***
SZJR / Árbevétel	-0.31***	-0.178***	-0.136***	-0.077***	-0.096***	-0.098***	-0.229***	-0.149***	-0.164***	-0.069***	-0.467***	-0.228***	-0.12***	-0.068***
Nettó beruházás	0.000006	-0.000006	0	0	0.000022	0.000005	0.000001	0	0.000001	0.000001	0.000006*	-0.000003	0.000002	0.000013
Nettó beruházás_1	0.000005	0.000006*	-0.000007	0.000001	-0.000009	0.000018*	0.000026***	0	-0.000001	0.000001	0.000007	0.000001	-0.000001	-0.000001
EBITDA-margin_1	0	0.087	0.624***	0.64***	0.641***	0.71***	0.442***	0.613***	0.721***	0.858***	0.574***	0.733***	0.552***	0.001
Cash-flow_1	0.000035**	0.000006	0.000001	0.000002	0	-0.000005	0.00015***	0	0	-0.000001	0.000012	0	0	-0.000012
T2009	0.004	0.023	-0.024**	0.001	0.013	0.006	-0.029***	0.006	-0.01	0.011	-0.011	0.006	-0.002	-0.103***
T2010	0.122***	0.056***	-0.006	0.006	0.009	0.013**	0.015	0.008	0.001	0.01	-0.019	0.011	0.047***	-0.004
T2011	0.006	0.024	-0.003	0.004	0.016	0.003	-0.012	-0.008	0.004	0.008	0.012	0.007	0.036**	-0.002
T2012	-0.041	0.018	-0.019	0.002	0.019*	0.006	-0.014	0.008	-0.004	-0.005	0.018	-0.013	0.036**	0.002
T2013	-0.004	0.02	-0.022*	-0.004	0.02*	0.003	-0.032***	0.003	-0.015*	0.005	0.007	-0.002	0.045***	0.046
R^2	0.1274	0.1613	0.4017	0.2567	0.2243	0.2346	0.1888	0.3008	0.2847	0.3744	0.4185	0.2639	0.07465	0.06438
F-statisztika	15.38	27.01	56.9	61.28	35.91	61.83	59.89	77.43	44.35	54.55	77.24	51.12	14.44	5.66
p-érték	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	1.769e-09
Megfigyelések száma	1277	1698	1030	2142	1503	2434	3100	2173	1350	1107	1301	1724	2161	1000

2. táblázat. A szegmentált modell együtthatói és szignifikanciái (*: $p < 0{,}1$; **: $p < 0{,}05$; ***: $p < 0{,}01$). *Forrás:* saját számítás.

A modellek arról tanúskodnak, hogy a legtöbb csoportban érvényesül a profitperzisztencia jelensége, azaz szignifikáns pozitív hatást fejt ki a jövedelmezőség késleltetett értéke a tárgyévi mutatóra. Ez a két legkisebb és a legnagyobb jövedelmezőségű szegmenst leszámítva mindenütt érvényesül, azaz a kiugróan magas vagy alacsony jövedelmezőségű vállalatok profitjai visszatérnek a kompetitív szintre, míg a többi vállalatcsoport esetében a profitabilitás fennmarad.

Ami az éves hatásokat illeti: a válság utáni visszaesést igazolja a 2009-es évet kódoló változó (T2009) szignifikáns negatív együtthatója 3 szegmensben. A szegmenseket alkotó vállalatok vonásait tanulmányozva megállapítható, hogy komoly visszaesést szenvedtek el

- az előző évben (2008-ban) 0,3 és 24% közötti jövedelmezőségű befektetett eszközökkel 40,3%-nál nagyobb arányban rendelkező és alacsony árbevételű (< 574,1 ezer euro) vállalatok (-2,9%)

- a fenti múltbeli jövedelmezőségi kategória azon vállalatai, melyek 40,3%-nál kisebb arányban rendelkeztek befektetett eszközökkel s Spanyolországban, Portugáliában, Franciaországban vagy Belgiumban vannak bejegyezve, forrásszerkezetük pedig 52,4%-nál kisebb arányban tartalmaz külső forrásokat (-2,4%); továbbá, érdekes módon

- a legmagasabb – 57,24%-ot meghaladó – múltbeli jövedelmezőségű vállalatok (-10,3%).

Ezek szerint a válság előidézte visszaesés a múltbeli jövedelmezőség, az eszköz- és forrásszerkezet és a vállalatméret függvénye, továbbá országok között is eltéréseket mutat, ugyanakkor várakozásainkkal szemben nem függ össze a vállalat múltbeli eladósodottságával.

Az árbevétel növekedésének hatása nem mindenütt szignifikáns, ami költségszerkezetbeli különbségekre hívja fel a figyelmet: a legtöbb csoportban ez elég nagy arányban tartalmaz fix és alacsony arányban változó költségeket ahhoz, hogy a fedezeti pont elérése után növekedéssel szignifikánsan javítani lehessen a jövedelmezőségen, míg másutt ez a hatás nem szignifikáns.

A nettó beruházás esetében a legtöbb csoportban inszignifikáns hatása van mindkét változatnak. Néhány szegmensben létezik pozitív beruházási hatás: a 24%-nál kisebb múltbeli jövedelmezőségű vállalatoknál ez rendre késleltetve jelentkezik, míg a magas jövedelmezőségű skandináv vállalatoknál késleltetés nélkül.

Nincs különbség abban, hogy az árbevételarányos hozzáadott érték mindenütt szignifikáns pozitív, míg a SZJR / Árbevétel szignifikáns negatív hatást fejt ki a jövedelmezőségre. Közülük az elemzés alapjául szolgáló egyszerűsített eredménykimutatást tanulmányozva egyik sem meglepő eredmény. Erős különbség mutatkozik azonban e két változó marginális hatásában az egyes szegmensek között, amelyből következik, hogy más-más stratégia teheti jövedelmezőbbé a vállalatokat: néhol az anyagjellegű ráfordítások optimalizálása vezet célra, míg másutt a személyi jellegű ráfordítások visszafogása.

Előbbi az árbevételarányos hozzáadott érték definíciójából kiindulva az anyagköltség és az ELÁBÉ visszafogásával lehetséges, melyre lehetőség van azáltal, ha nagyobb tételben rendelnek a vállalatok a szállítóiktól, vagy pedig kedvezmény realizálása érdekében hamarabb fizetnek nekik, és inkább egy hitelkeretet rulíroztatnak az eszközeik mögött.

A pénzeszközállományban előző évben bekövetkezett változások a legtöbb szegmensben nem bírtak jövedelmezőséget növelő hatással. Kivételt képeznek ez alól a legkisebb múltbeli jövedelmezőségű vállalatok, továbbá azon vállalatok, melyek múltbeli jövedelmezősége 0,3 és 24% közé esett, eszközszerkezetüket 40,3%-nál nagyobb mértékben a befektetett eszközök határozták meg, illetve alacsony árbevételszinten helyezkedtek el.

4 Összefoglalás

Tanulmányunkban a vállalati jövedelmezőséget meghatározó változók kapcsolatrendszerét vizsgáltuk mezőgazdasági vállalatok európai országhatárokon átívelő 24 ezres mintáján. A vizsgálatnak azzal a hipotézissel vágtunk neki, hogy ez a kapcsolatrendszer változékony, hisz a minta több országból, több évből tartalmazott számtalan – pénzügyi-számviteli mutatók mentén egymástól erősen – különböző vállalkozást. Számos változóról elképzelhetőnek tartottuk, hogy interaktál a jövedelmezőséget meghatározó változókkal és ezáltal megváltoztatja azok marginális hatását. Az összes ilyen interakció beparaméterezése azonban szükségtelenül bonyolulttá tette volna a modellt, márpedig kívánatos tulajdonsága egy modellnek a parszimónia, így egy gépi tanulási algoritmusra, a modell-alapú rekurzív partícionálásra bíztuk, hogy tanulja meg a rendelkezésére bocsátott adathalmazból, hogy mely változók vannak befolyással a többi változó marginális hatására. Ismételten a takarékosság jegyében egy olyan modellezési eljárást alkalmaztunk, amely tartalmazott egy elsőkörös változószelekciós fázist a probléma dimenzionalitásának csökkentése érdekében. Ezt a szerepet a véletlen erdők változófontossági mérőszámai töltötték be, melyek 5 magyarázó változó (Növekedés, Árbevételarányos hozzáadott érték, SZJR / ÉNÁ, illetve a Cash-flow és a jövedelmezőség késleltetett értéke) kiemelt jelentőségére világítottak rá (ezek aztán a későbbi tesztek során is legnagyobb részben relevánsnak bizonyultak), mely változókat éves *dummykkal*, illetve a beruházási tevékenységre reflektáló mutatókkal kiegészítve magyarázó változóként alkalmaztuk és teszteltük, hogy ezek marginális hatása összefüggésben van-e a vállalatok eszköz- és forrásszerkezetével, eladósodottságával, méretével, likviditásával, múltbeli jövedelmezőségével, avagy országonként és/vagy évenként különböznek a marginális hatások.

Összességében 14 darab lokális modellel tudtuk leírni a jövedelmezőség kapcsolatrendszerét egy globális (*pooled*) modell illesztése helyett. Eredményeink alapján az alkalmazott magyarázó változók marginális hatása erősen különbözik a múltbeli jövedelmezőség, az eszköz- és forrásszerkezet, az árbevétellel közelített vállalatméret szerint, valamint összefüggésben áll a vállalat országával, ugyanakkor nem változik az időmenzió mentén. Az egyes lokális

modellekhez elvezető és ezáltal egy-egy vállalati szegmenst azonosító döntési szabályokat nyomon követve profilírozni tudtuk azon vállalkozásokat, melyek a gazdasági világválság európai elterjedését követően a legnagyobb visszaesést szenvedték el. Megállapítottuk, hogy a profitok perzisztenciája majdnem mindenütt érvényesülő jelenség a mezőgazdaságban, ugyanis mindössze a legnagyobb és a legalacsonyabb múltbeli jövedelmezőségű vállalatok szegmensében nem volt szignifikáns az eredményváltozó késleltetett értékének hatása. Fényt derítettünk arra, hogy az árbevételarányos hozzáadott érték, valamint az SZJR / ÉNÁ változók mindenütt szignifikáns jövedelmezőségi hatást fejtenek ki. A legtöbb szegmensben az árbevétel növekedése is szignifikáns pozitív hatást fejt ki, amely arra utal, hogy a vizsgált vállalatok költségszerkezete néhány szegmenst leszámítva megfelelően magas arányban tartalmaz fix költségeket, így a növekedéssel jelentősen javítható a jövedelmezőség. Ugyanez már nem mondható el a beruházási aktivitásról, amely csak ritkán kapott szignifikáns együtthatót. Ha mégis szignifikáns hatást fejtett ki, azt jellemzően késleltetve tette.

Eredményeink természetesen magukon viselik azt, hogy a mezőgazdasági tevékenység szezonalitásából fakadóan az adott időpillanatra vonatkozó számviteli mutatók nem festenek teljes képet a vállalat működéséről, hisz bizonyos vállalatok készlet-, pénz- és vevőállománya erősen hullámozhat az év folyamán, így a befektetett eszközök év végi adatokból kalkulált aránya pontatlan képet fest az eszközszerkezetről, a likviditási ráta pedig a tényleges likviditásról. Ugyanez érvényes a forrásoldalra, ahol a finanszírozási (különösen forgóeszköz) igény erősen szezonális, ami azt eredményezi, hogy az idegen tőke év végi aránya alul- vagy felülbecsli a külső források tényleges arányát.

Ami a tanulmány módszertani kontribúcióját illeti, 2 igen kedvező tulajdonságokkal rendelkező modellezési eljárást mutattunk be. Közülük a véletlen erdőknél a változófontossági mérőszámokat emeltük ki. Ezzel kapcsolatban a releváns szakirodalom összefoglalásával rámutattunk arra, hogy alkalmazásukkor figyelembe kell venni a változók mérési skáláját, nominális és ordinális esetben a kategóriák számosságát, a mintavételezés módját, illetve a magyarázó változók közötti korrelációs struktúrát, és – szükség esetén – ezeknek megfelelően módosítani az eljárást. A modell-alapú rekurzív partícionálás legfőbb hozadékának a modell szegmentációjának adatvezérelt módon történő megvalósításának képessége bizonyult.

Lehetséges továbblépési irány a modellek előrejelző képességének validációs adathalmazokon történő tesztelése, mely jelen tanulmány keretein kívül esett, illetőleg további fa-alapú algoritmusok alkalmazása, ugyanis ahogy az egyedi fák stabilizálására létrejöttek a számos fával operáló *ensemble*-módszerek, úgy a modell-fák esetében is van már mód arra, hogy modell-fák egész erdejére alapozzuk az előrejelzést. Ennek implementációja a *mobForest* nevű R *package*-ben található (Garge, Bobashev & Egglestone, 2013). Ezen eljárás alkalmazására perspektivikus lehetőségként tekintünk és jövőbeli vizsgálódásaink során evidenciában fogjuk tartani.

Egyéb hasznosítható fa-alapú módszerként említést érdemel a QUINT-

algoritmus[33], melyet gyakran használnak az ún. kvalitatív kezelés–részcsoport interakciók[34] feltárásában, amely út már a személyre szabott orvoslás irányába vezet, ugyanis az eljárás célja, hogy egy RCT[35]-adathalmazban feltárja páciensek azon részcsoportjait különböző kezelés előtti karakterisztikák alapján, melyeknél az egyik típusú (A vagy alternatív) kezelés kedvezőbb eredményt hozott, mint a másik (B vagy hagyományos) (lásd Doove et al., 2016a; R-implementáció: Dusseldorp, Doove & Van Mechelen, 2016b). A jövedelmezőségi kontextusban a kezelés megfelelőjeként pl. a támogatás meglétét/hiányát alkalmazva profilírozhatóvá válnának azon vállalatok, melyeknél a támogatás a legnagyobb hozadékkal jár a jövedelmezőség szempontjából.

Irodalom

1. Boulesteix, A. L., Janitza, S., Kruppa, J., & König, I. R. (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2(6), 493–507. https://doi.org/10.1002/widm.1072
2. Breiman, L. (2001) Random forests. *Machine Learning,* 45(1), 5–32. https://doi.org/10.1023/a:1010933404324
3. Calle, M. L., & Urrea, V. (2010) Letter to the editor: stability of random forest importance measures. *Briefings in Bioinformatics,* 12(1), 86–89. https://doi.org/10.1093/bib/bbq011
4. Chaddad, F. R., & Mondelli, M. P. (2013) Sources of firm performance differences in the US food economy. *Journal of Agricultural Economics,* 64(2), 382–404. https://doi.org/10.1111/j.1477-9552.2012.00369.x
5. Dencic-Mihajlov, K. (2014) Profitability during the financial crisis: evidence from the regulated capital market in Serbia. *South-Eastern Europe Journal of Economics,* 12(1), 7–33 http://www.asecu.gr/Seeje/issue22/issue22-mihajlov.pdf
6. Doove, L. L., Van Deun, K., Dusseldorp, E., & Van Mechelen, I. (2016a) QUINT: A tool to detect qualitative treatment–subgroup interactions in randomized controlled trials. *Psychotherapy Research,* 26(5), 612–622. https://doi.org/10.1080/10503307.2015.1062934
7. Dusseldorp, E., Doove, L., & Van Mechelen, I. (2016b) Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them. *Behavior Research Methods,* 48(2), 650–663. https://doi.org/10.3758/s13428-015-0594-z
8. Garge, N. R., Bobashev, G., & Eggleston, B. (2013) Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics,* 14(1), 125. https://doi.org/10.1186/1471-2105-14-125
9. Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010) Variable selection using random forests. *Pattern Recognition Letters,* 31(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

[33]QUalitative INteraction Trees

[34]Treatment–subgroup interaction

[35]Randomized clinical trials, ahol véletlenszerűen rendelnek a kliensekhez legalább két kezelés közül egyet.

10. Grömping, U. (2009) Variable importance assessment in regression: linear regression versus random forest. *The American Statistician,* 63(4), 308–319. https://doi.org/10.1198/tast.2009.08199

11. Hajdu, O. (1987) *Sokváltozós statisztikai módszerek gyakorlati alkalmazása.* PRODINFORM, in Időszerű gazdaságirányítási kérdések, Budapest.

12. Hirsch, S., & Gschwandtner, A. (2013) Profit persistence in the food industry: evidence from five European countries. *European Review of Agricultural Economics,* 40(5), 741–759. https://doi.org/10.1093/erae/jbt007

13. Hirsch, S., & Schiefer, J. (2016) What causes firm profitability variation in the EU food industry? A redux of classical approaches of variance decomposition. *Agribusiness,* 32(1), 79–92. https://doi.org/10.1002/agr.21430

14. Hothorn, T., Hornik, K., & Zeileis, A. (2006) Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics,* 15(3), 651–674. https://doi.org/10.1198/106186006X133933

15. Ishwaran, H. (2007) Variable importance in binary regression trees and forests. *Electronic Journal of Statistics,* 1, 519–537. https://doi.org/10.1214/07-ejs039

16. Kopf, J., Augustin, T., & Strobl, C. (2013) The potential of model-based recursive partitioning in the social sciences: Revisiting Ockham's razor. In *Contemporary issues in exploratory data mining in the behavioral sciences,* Routledge, 97–117. https://epub.ub.uni-muenchen.de/11933/1/mob_techreport.pdf

17. Lazăr, S. (2016) Determinants of Firm Performance: Evidence from Romanian Listed Companies. *Review of Economic and Business Studies,* 9(1), 53–69. https://doi.org/10.1515/rebs-2016-0025

18. Loh, W. Y. (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1(1), 14–23. https://doi.org/10.1002/widm.8

19. Loh, W. Y. (2014) Fifty years of classification and regression trees. *International Statistical Review,* 82(3), 329–348. https://doi.org/10.1111/insr.12016

20. Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013) Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems* (431–439). http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-tre

21. Mozhgan, S. A. F. E., Faradmal, J., Poorolajal, J., & Mahjub, H. (2017) Model-based Recursive Partitioning for Survival of Iranian Female Breast Cancer Patients: Comparing with Parametric Survival Models. *Iranian Journal of Public Health,* 46(1), 35–43 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5401933/#__ffn_sectitle

22. Nicodemus, K. K. (2011) Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics,* 12(4), 369–373. https://doi.org/10.1093/bib/bbr016

23. Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010) The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics,* 11(1), 110. https://doi.org/10.1186/1471-2105-11-110

24. Nunes, P. J. M., Serrasqueiro, Z. M., & Sequeira, T. N. (2009) Profitability in Portuguese service industries: a panel data approach. *The Service Industries Journal,* 29(5), 693–707. https://doi.org/10.1080/02642060902720188

25. Pantea, M., Gligor, D., & Anis, C. (2014) Economic determinants of Romanian firms' financial performance. *Procedia – Social and Behavioral Sciences,* 124, 272–281. https://doi.org/10.1016/j.sbspro.2014.02.486

26. Pervan, M., & Mlikota, M. (2013) What determines the profitability of companies: case of Croatian food and beverage industry. *Economic Research – Ekonomska Istraživanja,* 26(1), 277–286. https://doi.org/10.1080/1331677x.2013.11517602

27. Prasad, A. M., Iverson, L. R., & Liaw, A. (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems,* 9(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

28. Pratheepan, T. (2014) A Panel Data Analysis of Profitability Determinants: Empirical Results from Sri Lankan Manufacturing Companies. *International Journal of Economics, Commerce and Management,* 2(12), 1–9. https://ssrn.com/abstract=2538927

29. R Core Team (2019) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.r-project.org

30. Razi, M. A., & Athappilly, K. (2005) A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications,* 29(1), 65–74. https://doi.org/10.1016/j.eswa.2005.01.006

31. Rumelt, R. P. (1991) How much does industry matter? *Strategic Management Journal,* 12(3), 167–185. https://doi.org/10.1002/smj.4250120302

32. Schiefer, J., & Hartmann, M. (2009, July) Industry, firm, year, and country effects on profitability: Evidence from a large sample of EU food processing firms. In AAEA & ACCI Joint Annual Meeting, Milwaukee, Wisconsin.

33. Schmalensee, R. (1985) Do markets differ much?, *American Economic Review,* 75(3), 341–351. https://www.jstor.org/stable/1814804?seq=1#metadata_info_tab_contents

34. Segal, M. R. (2004) Machine learning benchmarks and random forest regression. UCSF: Center for Bioinformatics and Molecular Biostatistics. https://escholarship.org/uc/item/35x3v9t4

35. Seibold, H., Zeileis, A., & Hothorn, T. (2016) Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics,* 12(1), 45–63. https://doi.org/10.1515/ijb-2015-0032

36. Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008) Conditional variable importance for random forests. *BMC Bioinformatics,* 9(1), 307. https://doi.org/10.1186/1471-2105-9-307

37. Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics,* 8(1), 25. https://doi.org/10.1186/1471-2105-8-25

38. Strobl, C., Malley, J., & Tutz, G. (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods,* 14(4), 323–348. https://doi.org/10.1037/a0016973

39. Tibshirani, R., James, G., Witten, D., & Hastie, T. (2013) *An introduction to statistical learning* (Vol. 112). New York: Springer.

40. Yazdanfar, D. (2013) Profitability determinants among micro firms: evidence from Swedish data, *International Journal of Managerial Finance,* 9(2), 151–160. https://doi.org/10.1108/17439131311307565

41. Zeileis, A., & Hornik, K. (2007) Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica,* 61(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

42. Zeileis, A., & Hothorn, T.: Parties, Models, Mobsters: A New Implementation of Model-Based Recursive Partitioning in R. Letöltve: 2019. április 14. https://www.rdrr.io/rforge/partykit/f/inst/doc/mob.pdf

43. Zeileis, A., Hothorn, T., & Hornik, K. (2008) Model-based recursive partitioning. *Journal of Computational and Graphical Statistics,* 17(2), 492–514. https://doi.org/10.1198/106186008x319331

44. Zouaghi, F., Hirsch, S., & Garcia, M. S. (2016) What makes firms profitable? A multilevel approach to the Spanish agri-food sector. In *Proceedings of the 153th EAAE Seminar: New dimensions of market power and bargaining in the agri-food sector: Organisations, policies and models,* Gaeta, Italy.

Mellékletek

R-kódok

```
#A megfelelő package betöltése az első modellezési fázishoz
library("party")

#A fontossági mutatókkal feltöltendő tömb létrehozása
varimp_global_4000_cforest <- array(dim = c(50,12))

#A véletlen erdők beágyazása egy 50 iterációs lépésből álló ciklusba, melynek
minden egyes lépésében kinyerjük az aktuális modell fontossági mutatóit és
elraktározzuk a fenti tömb i-edik sorában
for (i in 1:50) {

      set.seed(i)

      model_cforest_4000 <- cforest(EBITDA_margin ~ Növekedés +
      Hozzáadott.érték + Net.Investment + Net.Investment_1 + SZJR.Sales
      + Likviditási.ráta_1 + Befektetett.eszközök.aránya_1 +
      Idegen.tőke.aránya_1 + Net.debt_1 + Net.debt.EBITDA_1 + Cash.flow_1
      + EBITDA_margin_1, data = subsample_of_adatok, controls =
      cforest_control(mincriterion = 0.95, ntree = 100, mtry = 5, trace
      = TRUE, replace = FALSE))

      set.seed(i)

      varimp_global_4000_cforest[i, ] <- varimp(model_cforest_4000,
      conditional = FALSE)

}

#A megfelelő package betöltése a második modellezési fázishoz
library('partykit')

#Modell-alapú rekurzív partícionálás az lmtree függvény segítségével
mob_model_4000_évekkel <- lmtree(EBITDA_margin ~ Növekedés + Hozzáadott.érték
+ SZJR.Sales + Net.Investment + Net.Investment_1 + EBITDA_margin_1 +
Cash.flow_1 + T2009 + T2010 + T2011 + T2012 + T2013 | EBITDA_margin_1 + Year.1
+ Befektetett.eszközök.aránya_1 + Net.debt.EBITDA_1 + Likviditási.ráta_1 +
Idegen.tőke.aránya_1 + Net.Investment_1 + Country + Sales.  , data =
subsample_of_adatok, alpha = 0.01, bonferroni = TRUE, minsplit = 1000, verbose
= TRUE)

#Együtthatók modellenkénti kiírása
coef(mob_model_4000_évekkel)
```

I.	Értékesítés nettó árbevétele
II.	Anyagköltség + ELÁBÉ
A = I - II	Hozzáadott érték
III.	Egyéb tételek
IV.	Személyi jellegű ráfordítások
B = A + III.- IV.	EBITDA
V.	Értékcsökkenési leírás
C = B – V.	Üzemi (üzleti) eredmény
VI.	Pénzügyi eredmény
VII.	Egyéb rendkívüli tételek
VIII.	Adófizetési kötelezettség
D = C + VI. + VII. - VIII	Adózott eredmény

3. táblázat. Az elemzések alapjául szolgáló eredménykimutatás.
Forrás: saját összeállítás.

Változó neve	Véletlen erdők	Modell-alapú rekurzív partícionálás
EBITDA-margin	Y	Y
Növekedés: az árbevétel előző évhez képesti %-os növekedése	X	X
SZJR/Árbevétel	X	X
Árbevételarányos hozzáadott érték = = Hozzáadott érték / Árbevétel, ahol Hozzáadott érték = Árbevétel − − (ELÁBÉ + Anyagköltség)	X	X
Nettó beruházás	X	X
Nettó beruházás_1	X	X
EBITDA_margin_1	X	X
Cash flow_1	X	X
Likviditási ráta_1	X	Z
Idegen tőke aránya_1	X	Z
Befektetett eszközök aránya_1	X	Z
Nettó adósság_1	X	Z
Nettó adósság / EBITDA_1	X	Z
Árbevétel		Z
Év		Z
Ország		Z
T2009		X
T2010		X
T2011		X
T2012		X
T2013		X

4. táblázat. Változók és a modellezés különböző fázisaiban betöltött szerepük.

	Növekedés	Hozzáadott érték	Nettó beruházás	SZJR / Árbevétel	Likviditási ráta_1	Nettó beruházás_1	Befektetett eszközök aránya_1	Idegen tőke aránya_1	Nettó adósság_1	EBITDA-margin_1	Cash-flow_1	(Nettó adósság / EBITDA)_1
Növekedés	1	-0.002	0	-0.007	0.004	0.001	0.004	0.006	0	-0.123***	-0.002	-0.001
Hozzáadott érték	-0.002	1	0.001	0.223***	0.015**	-0.003	0.162***	-0.094***	0.002	0.014**	0.007	-0.009
Nettó beruházás	0	0.001	1	-0.018***	-0.003	-0.297***	0.032***	0.005	0.743***	0.002	0.359***	0.001
SZJR / Árbevétel	-0.007	0.223***	-0.018***	1	0.035***	-0.002	0.124***	-0.052***	-0.018***	0.007	-0.028***	-0.014**
Likviditási ráta_1	0.004	0.015**	-0.003	0.035***	1	-0.002	-0.03***	-0.074***	-0.003	-0.003	-0.003	0.002
Nettó beruházás_1	0.001	-0.003	-0.297***	-0.002	-0.002	1	0.024***	-0.006	-0.389***	0.002	0.282***	0
Befektetett eszközök aránya_1	0.004	0.162***	0.032***	0.124***	-0.03***	0.024***	1	-0.022***	0.029***	0.02***	0.018***	-0.006
Idegen tőke aránya_1	0.006	-0.094***	0.005	-0.052***	-0.074***	-0.006	-0.022***	1	0.01	-0.029***	-0.018***	0.003
Nettó adósság_1	0	0.002	0.743***	-0.018***	-0.003	-0.389***	0.029***	0.01	1	0	0.637***	0.001
EBITDA-margin_1	-0.123***	0.014**	0.002	0.007	-0.003	0.002	0.02***	-0.029***	0	1	0.007	0.002
Cash-flow_1	-0.002	0.007	0.359***	-0.028***	-0.003	0.282***	0.018***	-0.018***	0.637***	0.007	1	0
(Nettó adósság / EBITDA)_1	-0.001	-0.009	0.001	-0.014**	0.002	0	-0.006	0.003	0.001	0.002	0	1

5. táblázat. Korrelációs mátrix (*: $p < 0{,}1$; **: $p < 0{,}05$; ***: $p < 0{,}01$).
Forrás: saját számítás.

Ország	Megfigyelések száma, db
HUN	534
POR	5726
ESP	3751
FIN	1378
SWE	3199
FRA	3715
ITA	5584
BEL	84
GER	29

6. táblázat. A minta országok szerinti megoszlása.

	2008	2009	2010	2011	2012	2013
T2008	1	0	0	0	0	0
T2009	0	1	0	0	0	0
T2010	0	0	1	0	0	0
T2011	0	0	0	1	0	0
T2012	0	0	0	0	1	0
T2013	0	0	0	0	0	1

7. táblázat. Az éves dummy változók kódolása.

		Növekedés	Hozzáadott érték	Nettó beruházás	Nettó beruházás_1	SZJR / Árbevétel	Likviditási ráta_1	Befektetett eszközök aránya_1	Idegen tőke aránya_1	Nettó adósság_1	(Nettó adósság/ EBITDA)_1	Cash-flow_1	EBITDA-margin_1
IncMSE	Átlag	0.002756	0.008222	0.000947	0.00059	0.006997	0.001354	0.001906	0.001155	0.001934	0.003913	0.004799	0.032032
	Szórás	0.000124	0.000207	6.45E-05	7.44E-05	0.000222	0.000113	0.000117	7.96E-05	0.000103	0.000209	0.000216	0.000392
	Rang	6	2	12	11	3	9	8	10	7	5	4	1
IncNode Purity	Átlag	59.55497	60.60933	25.22647	23.18471	63.82748	28.29021	29.60743	26.58713	21.34682	31.22162	54.03535	230.9733
	Szórás	0.605279	1.612364	0.318626	0.362442	0.998576	0.366859	0.376934	0.368795	0.338141	0.42134	2.020447	3.625565
	Rang	4	3	10	11	2	8	7	9	12	6	5	1

8. táblázat. A véletlen erdők (randomForest) fontossági mutatói.
Forrás: saját számítás.

TREE-BASED METHODS IN THE INVESTIGATION OF CORPORATE PROFITABILITY: VARIABLE SELECTION AND MODEL SEGMENTATION

The aim of the study is to introduce 2 tree-based modeling tool in the field of corporate profitability – where they have not been used yet – and apply them together in order to reveal the relationship between the variables determining the profitability, to investigate whether the marginal effect of certain variables is stable over the range of other variables or it shows significant instabilities in addition to revealing the patterns in the profitability drop as a consequence of the economic crisis. The 2 algorithms – random forests whose prediction is based on several random trees as well as the model-based recursive partitioning incorporating the data-driven nature of the trees and the theory-based nature of the statistical-econometrical modeling – turned out to be applicable for the purposes of the study based on their favourable characteristics highlighted in the relevant literature.

During the application of these algorithms within a two-stage modeling procedure we make use of a sample consisting of 24 thousand companies operating either in one of the EU-15 countries or Hungary engaged in agriculture, forestry and fishery as their main business activity. We utilize their financial data from the 2008-2013 period while conducting the analyses. The profitability of these companies is measured by EBITDA-margin the application of which is especially important if we take into consideration the cross-border nature of the study since the indicator being the ratio of EBITDA and Net sales the former of which as a measure of the profitability of the core business activity enables to avoid distortions originating from the differences in the depreciation and amortization methodology, interest rates and the local corporate taxation. To model this profitability indicator we apply several explanatory variables calculated from the components of our simplified profit and loss statement, those of the investment activity, year and country, net sales as a proxy of company size complemented by some lagged financial indicators. From the set of explanatory variables we identify the most important ones with the help of the variable importance measures of the random forests – one of the most favourable characteristics of them is the provision of these measures which can be extremely useful for variable selection purposes – and complement them with variables having low importance measures but bearing useful information from the economic point of view. We conduct this variable selection step in order to reduce the dimensionality of the problem as a result of which we obtain a smaller set of variables which we utilize as the explanatory variables of our linear regression models and investigate the marginal effect of certain variables as well as whether they are dependent of some other characteristics of the companies measured by the so-called partitioning variables. We establish the set of partitioning variables in line with our hypotheses according to what the marginal effect of the variables might depend on the asset- and capital structure of the companies, their working capital management which can be measured by the liquidity ratio, indebtedness, lagged profitability and size. Furthermore we deem conceivable that the profitability effects even vary year by year and across countries. Though in order to test it in the traditional way we would have needed to define a high number of interaction terms and estimate their effects leading to unnecessary complexity which is of course to be avoided. Thus we make use of the recursive algorithm to learn from the data it has been provided what are the segments of our sample where the relationship between the variables is homogeneous. This way we practically segment models with the help of certain company characteristics.

We use the R-implementation of the algorithms for variable selection and model segmentation purposes the results of which let us conclude that instead of fitting one global (*pooled*) model the phenomenon can be better described by fitting 14 local models since the parameters show significant instability over lagged profitability, asset- and capital structure, company size as well as the country of operation.

Studying the parameters of the local models we made the conclusion that the phenomenon of profit persistence prevails in almost every segment. However there is no persistence as for the companies with the highest and lowest past profitability – here the lagged profitability does not have significant effect. We shed light on the fact that the variables Value added / Net sales as well as Personnel cost / Net sales have significant effect in each of the segmented models. The direction of these effects is not surprising at all. However their effect size has strategic importance since it shows which type of costs shall be reduced to make a company more profitable. In most of the segments the growth has also positive effect on the profitability highlighting that the cost-structure of the companies under investigation contains fix costs in a sufficiently high proportion enabling growth to improve the profitability. Interestingly the same does not hold for the investment activity whose parameter is significant only in a few segments and – if it is – its effect prevails basically with delay. Our segmented model – beyond the firm-specific assessment – made it possible to characterize the companies suffering a significant profitability drop in 2009 – after the outburst of the global economic crisis. This characterization lets us conclude that the above mentioned profitability drop depends on past profitability, asset- and capital-structure as well as size although – contrary to our expectations – it has nothing to do with indebtedness. As for random forests we highlighted the variable importance measures in the connection of which we directed the attention towards the fact that before their application the measurement level of the variables, in the case of nominal and ordinal data the number of categories, the sampling method as well as the correlation structure between the variables have to be taken into account and – if necessary – the method has to be altered accordingly. The most important result of the model-based recursive partitioning was the ability to conduct the model segmentation in a data-driven way without having a priori hypotheses about the segments across which the marginal effects vary.

As a limitation of the study we have to mention that we made use of financial data of agricultural companies whose operation can be highly seasonal accompanied with strong fluctuations in cash and equivalents, inventories, receivables and working capital financing of the companies that might not properly be reflected in the balance sheet of the companies which depicts the company's assets and liabilities at one single instant. Consequently the data we have do not necessarily reflects the reality and the value of the liquidity and indebtedness indicators calculated from the balance sheet might be distorted accordingly. As a possible step forward we mentioned the application of validation data sets in order to test the predictive power of our models created by the use of the learning data set. Beyond these we direct the attention towards the applicability of further machine learning algorithms such as QUINT and the forest of model trees implemented in the R package called *mobForest*.

Key words: random forests, model-based recursive partitioning, variable selection, model segmentation