Közgazdasági és
Gazdaságinformatikai
Doktori Iskola

TÉZISGYŰJTEMÉNY

Takács Olga

Bérfüggvények becslése Magyarországon gépi tanuló algoritmusokkal

című Ph.D. értekezéséhez

Témavezető:

Vincze János, DSc
egyetemi tanár

Budapest, 2022

Közgazdaságtan Intézet

TÉZISGYŰJTEMÉNY

Takács Olga

Bérfüggvények becslése Magyarországon gépi tanuló algoritmusokkal

című Ph.D. értekezéséhez

Témavezető:

Vincze János, DSc
egyetemi tanár

© Takács Olga

Tartalomjegyzék

1. Kutatási előzmények és a téma indoklása

A munkagazdaságtan egyik leginkább kutatott témája az, hogy milyen tényezők határozzák meg a béreket. A téma kutatása során olyan kérdések merülhetnek fel, minthogy milyen mértékben számít az iskolai végzettség. Az életkor előrehaladtával mennyivel emelkedik a munkavállalók bére? Hol fizetnek jobban: a hazai vagy a külföldi tulajdonú cégeknél? Melyik ágazatban fizetnek a legtöbbet és a legkevesebbet?

A bérezési kérdésekkel foglalkozó kutatások kiindulópontja Mincer (1958) és Becker (1964) munkásságához köthető. Az általuk meghatározott humán tőke modell a tanulás, az életkor és a tapasztalat bérekre gyakorolt hatását vette figyelembe. Későbbi kutatásokban újabb személyes, vállalati, illetve munkaerőpiaci jellemzőket vontak be a vizsgálatokba. (Wachtel–Betsey (1972), Blanchflower et al. (1992)) A további változók körének bevonása mellett más kutatások a modellek illeszkedésének javítására törekedtek. (Heckman (1979), Chamberlain (1991))

Ezekkel a kutatásokkal párhuzamosan, Becker (1971) munkásságából kiindulva, a közgazdászokat egyre inkább foglalkoztatta a munkaerőpiacon megjelenő diszkrimináció témája is. A munkapiaci megkülönböztetés mérésével kapcsolatos kutatások sarokköve Blinder (1973) és Oaxaca (1973) munkássága. A szerzőkről elnevezett Blinder-Oaxaca felbontás azóta alapvető eszközzé vált a csoportok közti különbségek vizsgálatánál (Jann (2008), Fortin et al. (2011)), és kiemelten használják a nemi bérkülönbségek elemzésénél. (Weichselbaumer–Winter-Ebmer (2005)) A dekompozíció alkalmas arra, hogy az átlagos különbségen belül elkülönítése a munkaerőpiaci jellemzőkkel magyarázott és nem magyarázott részt. A felbontáshoz pedig minden esetben szükséges a vizsgált csoportokhoz tartozó bérfüggvények becslése.

A bérfüggvények meghatározásánál és a Blinder-Oaxaca dekompo-

zíciónál is történtek módszertani változtatások, melyek hozzájárultak ahhoz, hogy a bérek, illetve a bérkülönbségek és a munkapiaci jellemzők közti strukturális összefüggéseket jobban lehessen vizsgálni. Azonban az eddigi kutatások nagy részében, mind a bérezési kérdések tárgyalásánál, mind a nemi bérkülönbségek kutatásánál a bérfüggvények meghatározásához paraméteres – általában regresszión alapuló – becsléseket használtak. Ezek a becslések pedig minden esetben megkövetelnek előzetes feltételezéseket az adatgeneráló folyamatról, melyek általában elméleti modellekből származnak, azonban nem minden esetben megfelelően alátámasztottak.

A bérfüggvények meghatározása kapcsán továbbra is lehetnek – eddig tényként kezelt, azonban megkérdőjelezhető – vagy fel nem tárt elméleti összefüggések. Ezeknek az összefüggéseknek a beépítése vagy kihagyása a bérfüggvények becslésénél félrevihetik a gondolkodást. A disszertációmban a bérfüggvények meghatározása kapcsán újítás a gépi tanuló algoritmusok használata. A tanuló algoritmusok a hagyományos statisztikai módszerekkel összehasonlítva más feltételezésekből indulnak ki. A gépi tanuló eljárások igyekeznek a lehető legszabadabb függvényformákat alkalmazni, ami által „adatvezérelt" módon a lehető legtöbb információ kinyerhető az adatokból. Ezek az eljárások képesek rátanulni az adatban rejlő összefüggésekre, emiatt nem kell előzetes feltételezéseket tenni az adatgeneráló folyamatról.

Az értekezésben a gépi tanuló algoritmusok használata kettős célt szolgál. Először a magyarországi bérfüggvényt becslem tanuló algoritmussal, mellyel feltárhatóak a változók közötti kapcsolatok. Az algoritmus által feltárt összefüggések egy része lehet olyan, melyek már korábban is ismertek voltak, illetve lehetnek olyanok, melyek további kérdéseket vethetnek fel. A tanuló algoritmusok alkalmazásának további célja a lehető legpontosabb előrejelzések készítése, melyeket a nemi bérkülönbségek vizsgálatánál használok ki. Az átlagos nemi bér-

különbségeket leggyakrabban Blinder-Oaxaca dekompozícióval elemzik, melyhez mindenképpen szükséges egy férfi és egy női bérfüggvény becslése. A disszertációmban gépi tanuló algoritmussal becslem meg ezeket a függvényeket.

A disszertáció felépítése a következő. A 2. fejezet tartalmazza a kapcsolódó szakirodalom Magyarországra vonatkozó eredményeinek összefoglalását és a saját kutatásokhoz használt adatbázis leírását. A gépi tanulási algoritmusokat és néhány interpretációs eljárást, amelyeket a saját kutatásaimban használok a 3. fejezetben mutatom be.

A 4. fejezet Takács–Vincze (2018) átdolgozott változata, melyben a magyarországi bérfüggvényt döntési fákkal becsültük meg. A 2. fejezetben bemutatott kutatásokban a bérfüggvények becslése hagyományos ökonometriai módszerekkel történt, amiből adódóan a függvényforma meglehetősen kötött volt, így csak azok az összefüggések szerepeltek bennük, melyeket előre definiáltak a szerzők. A döntési fa használata lehetővé teszi, hogy az adatgeneráló folyamatra tett előzetes feltevések nélkül lehessen béreket becsülni és a hagyományostól eltérő módon tudjuk vizsgálni a változók közti kapcsolatokat. A kutatás tekinthető előzetes adatelemzésnek is, melyben a cél az eddig rejtett kapcsolatok feltárása tanuló algoritmusok használatával.

Az 5. és a 6. fejezetek a nemi bérkülönbségek becslésével foglalkozó irodalomhoz kapcsolódnak. Az 5. fejezet Takács (2021) szerkesztett változata. Ez a fejezet tartalmazza az átlagos nemi bérkülönbségek vizsgálatához használt Blinder-Oaxaca dekompozíció bemutatását és a felbontás véletlen erdőre átírt változatát. Ebben a fejezetben összehasonlítok egy OLS-sel és egy általánosított véletlen erdővel készített bérfüggvény becslést. Ezen felül összevetem a felhasználásukkal készített Blinder-Oaxaca dekompozícióból származó eredményeket. A célom annak bemutatása, hogy a módszertan megválasztása milyen mérték-

ben befolyásolja a dekompozícióból kapott eredményeket.

A 6. fejezet Takács–Vincze (2020) átdolgozott változata, melyben az általánosított véletlen erdővel készült Blinder-Oaxaca felbontás eredményeinek időbeli alakulását vizsgálom. Bemutatom, hogy a Blinder-Oaxaca felbontásból kapott magyarázott és nem magyarázott rész nagysága hogyan alakult a vizsgált időszakban. A fejezet végén elkülönítem azokat a csoportokat, amelyek a leginkább, illetve a legkevésbé járultak hozzá az átlagos nemi bérkülönbségek kialakulásához és fennmaradásához. A 7. fejezet összefoglalja az értekezés eredményeit.

2. A felhasznált módszerek

Adatbázis

A disszertációban a Nemzeti Munkaügyi Hivatal által gyűjtött Bértarifa-felvételből származó (továbbiakban Bértarifa) adatokat használom. Az adatbázis a jövedelmi adatok (alapilletmény, bónuszok, pótlékok) mellett szolgáltat információt vállalatokról (létszám, tulajdon, ágazat, telephely) és a munkavállalókról (iskolai végzettség, életkor) egyaránt. Az eredményváltozó mindegyik fejezetben az átlagos havi keresetek logaritmusa. Az átlagos havi kereset magába foglalja az alapilletményt és az egyéb, havi szintre vetített juttatásokat is.

Az elemzésekben csak a teljes munkaidőben a versenyszférában foglalkoztatottakat vizsgálom. Továbbá kizárom a 20 főnél kisebb vállalatokat, ahogyan Lovász (2013) is tette hivatkozva Elek és szerzőtársai (2009) eredményére miszerint a kisebb cégeknél gyakoribb a fekete- és szürkefoglalkoztatás. Emiatt az ezektől a vállalatoktól származó béradatok nem megbízhatóak.

A 4. és az 5. fejezetekben kizárólag a 2016-os adatokat használom. A 6. fejezetben pedig a nemi bérkülönbségek időbeli alakulásának vizs-

gálatánál a 2008 és 2016 közti adatokra támaszkodom. A 4., az 5. és a 6. fejezetben bemutatásra kerülő kutatásoknál a felhasznált magyarázóváltozók köre eltér: az 1. táblázat összefoglalóan mutatja az egyes fejezetekben használt magyarázóváltozókat.

1. táblázat. Felhasznált magyarázóváltozók köre

Változó	Mértékegység	Fejezet		
		4.	5.	6.
Nem	0: nő, 1: férfi	X		
Iskolai végzettség	9 kategória	X	X	X
Életkor	Évek	X	X	X
Szolgálati idő	Hónapok	X	X	X
Foglalkozás	Egyjegyű FEOR	X		
Foglalkozás	Kétjegyű FEOR		X	X
Külföldi részarány	4 kategória	X	X	X
Állami (önkormányzati) részarány	4 kategória	X	X	X
Vállalati létszám	Foglalkoztatottak száma	X	X	X
Régió	NUTS 2 kategóriák	X	X	X
Településtípus	1: Bp., 2: város, 3: egyéb	X	X	X
Ágazat	Egyjegyű TEÁOR	X	X	X
Vállalati kollektív szerződés	0: nincs, 1: van	X	X	X
Ágazati kollektív szerződés	0: nincs, 1: van	X	X	
Több munkáltatóra kiterjedő szerződés	0: nincs, 1: van	X	X	

Forrás: Bértarifa

Az 1. táblázatban a szolgálati idő az adott munkáltatónál eltöltött hónapokat takarja. Az ágazatok közül a kevés számú megfigyelés miatt a közigazgatást, védelmet és kötelező társadalombiztosítást tartalmazó O ágazatot kizárom. A 6. fejezetben a kollektív szerződéseket összevontan, egy változóval vizsgálom: ha adott személy rendelkezett bármilyen típusú kollektív szerződéssel, akkor a változó értéke 1, különben 0.

Gépi tanuló algoritmusok

A gépi tanuló algoritmusok közül döntési fákra épülő eljárásokat használok. A fáknál a kiindulópontot jelentő adathalmazt a fa tetején látható gyökér jelenít meg. A fa minden lépésnél a sokaságot két

diszjunkt halmazra osztja szét, melyekhez tartozóan kiátlagolja az eredményváltozót. Ezek a csoportokhoz rendelt átlagok jelentik a csoport tagjaira vonatkozó előrejelzést. A kiinduló adathalmaz, a gyökér csomóponttá válik és meghatározódik az első vágási pont.

A következő vágás meghatározásához az eljárás a két levelet függetlenül kezeli és az egyik levelet további két csoportra osztja az egyik változó – a vágóváltozó – szerint. Ezzel az érintett levél csomóponttá válik és két újabb levél keletkezik. Majd az eljárás újabb vágásokkal, újabb leveleket hoz létre, ezzel növelve a döntési fa méretét. A fa növekedése megáll, ha elér egy előre meghatározott kilépési kritériumot.

A statisztikai teszteket használó algoritmusoknál, melyeket én is használok az értekezésben, a vágóváltozó kiválasztása az adott magyarázóváltozó és az eredményváltozó közti kapcsolat szorosságát mérő statisztikai teszt alapján történik. (Jiao et al. (2020)) A statisztikai tesztek használata segít elkerülni a túlillesztés problémáját, valamint a CART-on alapuló eljárások azon hátrányát, hogy ezek az algoritmusok előnyben részesítik a több lehetséges vágási ponttal rendelkező változókat.

A döntési fák hátránya, hogy magas az előrejelzési hibájuk és nem robusztusok az adathalmaz változására. Azonban a robusztusság és az előrejelzések pontossága több fa eredményének aggregálásával javítható. Három eljárást alkalmaznak erre a célra: a bagging és a boosting eljárásokat, valamint a véletlen erdőket. (James et al. (2013)) A véletlen erdő bootstrap mintákon készíti el a döntési fákat. Emellett a vágási pontokban az algoritmus a változók egy részét tekinti potenciális vágóváltozónak és nem az összeset. (Hastie et al. (2009))

A disszertációm 4. fejezetében döntési fát építek, hogy a bérek és a felhasznált magyarázóváltozók közti kapcsolatokat vizsgáljam. Az elérhető algoritmusok közül a feltételes következtetési fákat (Conditio-

nal Inference Tree, CTree) használom. A véletlen erdő eljárást pedig azoknál a becsléseknél alkalmazom, ahol az előjelzések pontosságának van lényeges szerepe. Emiatt az 5. és a 6. fejezetekben a Blinder-Oaxaca dekompozícióhoz tartozó bérfüggvényeket az Athey és szerzőtársai (2019) munkásságán alapuló regressziós erdőkkel (Regression Forest) becslem meg. Emellett a 4. fejezetben a CTree eredményeinek robusztusságát a következtetési fákat aggregáló feltételes véletlen erdővel (Conditional Random Forest, CForest) ellenőrzöm.

Feltételes következtetési fa – CTree

A CTree a vágási pontok meghatározásánál első lépésként az összes magyarázóváltozó és az eredményváltozó együttes függetlenségére vonatkozó nullhipotézist teszteli. Amennyiben ezt a globális nullhipotézist elveti az algoritmus, akkor minden változóhoz hozzárendeli az eredmény- és az adott inputváltozó szorosságára vonatkozó tesztstatisztikához tartozó p-értéket. Amennyiben az algoritmus talál olyan magyarázóváltozókat, ahol az adott változó függetlenségére vonatkozó nullhipotézist el tudja vetni, akkor azok közül kiválasztja a legkisebb p-értékkel rendelkezőt. A legkisebb p-érték mutatja, hogy ennek a változónak a legszorosabb a kapcsolata az eredményváltozóval.

A következő lépésben az algoritmus a kiválasztott változónál megvizsgálja az összes lehetséges vágási pontot, mely alapján két részre osztható a sokaság. A lehetséges vágási pontok közül pedig szintén statisztikai teszttel határozza meg, hogy melyik vágási pontot érdemes választani. A CTree addig folytatja a vágásokat, amíg egy előre meghatározott kilépési kritérium nem teljesül.

Feltételes véletlen erdő – CForest

A CTree algoritmussal készített fákat összegző erdő a CForest. Ez az erdő az adatok visszatevés nélküli almintáján épít CTree eljárással fákat. Emellett a CForest esetében az előrejelzések azoknak a megfigyeléseknek az átlagaiként adódnak, amelyekkel az adott megfigyelés minden fán egy levélre esett. (Levshina (2020))

Feltételes változófontossági érték

A feltételes változófontossági értékekek egy mennyiségi skálán fejezik ki, hogy az adott modellben az inputváltozók milyen mértékben hatnak az eredményváltozóra. (Inglis et al. (2022)) Ennek ismeretében könnyebbel el lehet dönteni, hogy az eredményváltozó szempontjából melyek a lényeges és a kevésbé lényeges változók anélkül, hogy az egész fát részletesen meg kellene vizsgálni. A disszertációmban a Strobl és szerzőtársai (2008) által kidolgozott feltételes permutációra épülő fontosságokat használom.

Általánosított véletlen erdő – regressziós erdő

Az általánosított véletlen erdőt Athey és szerzőtársai (2019) alkották meg. Ez az algoritmus felhasználja a véletlen erdő néhány tulajdonságát úgymint a fa építésénél az egymást követő bináris vágásokat, a mintavételt és a lehetséges vágóváltozók számának korlátozását a vágási pontokban. Azonban az algoritmus a fák építésénél még egy lényeges tulajdonságot, a „becsületesség" (honesty) feltételét is figyelembe veszi: az algoritmus a tanulómintát két részre osztja és az egyik felén végzi a vágásokat, a másik mintarészből pedig a levelekhez tartozó becsléseket számszerűsíti. A regressziós erdőknél a minta visszatevés nélküli mintavételt jelent.

A regressziós erdőben az előrejelzések készítésénél azokat a megfigyeléseket átlagolja ki az algoritmus, amelyek az adott megfigyeléssel egy levélre esnek. Így egy megfigyelés, amely a váltózótérben közel áll a vizsgált megfigyeléshez, akár többször is beleszámolódhat az átlagba.

Klasszifikációs analízis – CLAN

A klasszifikációs analízis (Classification Analysis, CLAN) módszertanát Chernozhukov és szerzőtársai (2018) dolgozták ki. A szerzők célja egy olyan eljárás kifejlesztése volt, amellyel a gépi tanuló algoritmusokkal számolt feltételes átlagos kezelési hatások fő tulajdonságai vizsgálhatóak. A feltételes átlagos kezelési hatás két – randomizált kísérletek esetében a kezelt és a kezeletlen – csoport közti átlagos különbséget mutatja.

A CLAN arra a kérdésre keresi a választ, hogy milyen tulajdonságokkal rendelkeznek a leginkább és a legkevésbé érintett csoportok. A „leginkább" érintett csoport a legnagyobb, a „legkevésbé" érintett pedig a legkisebb kezelési hatással rendelkezőket foglalja magába. Ezeknek a csoportoknak a karakterisztikáit leíró vektorok mutatják, hogy milyen általános tulajdonságokkal rendelkeznek a két csoport tagjai. A karakterisztikáit leíró vektorok elemei a közvetlenül megfigyelt változókból számolódnak. Továbbá ezen vektorok különbsége is számszerűsíthető, ami az átlagos különbséget mutatja a legkevésbé és a leginkább érintett csoport között.

Véletlen erdő a nemi bérkülönbségek elemzésénél

A disszertációm 5. és 6. fejezetében a bérek feltételes várhatóérték függvényét közelítem OLS regresszióval és regressziós erdővel is. OLS regresszió esetén a Blinder-Oaxaca dekompozíció a következőként írható

fel:

$$\overline{Y_F} - \overline{Y_N} = \left(\overline{X_F} - \overline{X_N}\right)\beta_F - \overline{X_N}\left(\beta_F - \beta_N\right) + \text{torzítás}, \qquad (1)$$

ahol a $\overline{Y_F} - \overline{Y_N}$ a férfiak (F) és nők (N) átlagbéreinek különbségét – a nyer bérkülönbséget – ragadja meg. $\overline{X_F}$ és $\overline{X_N}$ a megfigyelt magyarázóváltozók átlagos értékei a férfiak és a nők esetében. A β_F és a β_N a két nemre külön-külön becsült OLS regressziók együtthatói. Az (1) jobb oldalán az első kifejezés az összetételhatást, a második a bérstruktúrahatást ragadja meg. Az összetételhatás a megfigyelt jellemzőkkel magyarázható részt, a bérstruktúrahatás a nem magyarázható részt mutatja az átlago különbségen belül. Az OLS esetében a tanulómintán a torzítás definíció szerint nulla, míg a tesztmintán ettől eltérhet.

Az (1)-ben szereplő összefüggés megegyezik Blinder (1973) és Oaxaca (1973) eredeti felírásával. A szerzők úgy gondolták, hogy a bérstruktúrahatás tökéletesen méri a diszkriminációt. Későbbi tanulmányok azonban kimutatták, hogy ez a rész egyéb statisztikai hibákat is tartalmaz: például a kihagyott változók hatását vagy a munkatapasztalat hibás mérését. (Reilly (2001)) Emiatt a bérstruktúrahatás fontos információt hordoz a nemi különbségek tekintetében, azonban érdemes óvatosnak lenni az értelmezésénél.

A heterogén bérstruktúrahatás az i-edik megfigyeléshez tartozóan, az adott személy jellemzői alapján a következő formában írható fel:

$$HWS(i) = X_i^{'}(\beta_F - \beta_N), \qquad (2)$$

ahol X_i az adott személy megfigyelt jellemzőinek vektora, míg a β_F és β_N a férfi és a női almintán becsült OLS regresszióból származó együtthatók. A nőknél kiszámolt heterogén bérstruktúrahatások átlaga megegyezik az (1) jobb oldalának második tagjával, az átlagos bérstruktúrahatással.

A regressziós erdő esetében nem becsülhetők a β együtthatók, hanem csak közvetlenül az előrejelzések. Így a következő összefüggés mutatja a Blinder-Oaxaca dekompozíció véletlen erdőre való kiterjesztését:

$$\overline{Y_F} - \overline{Y_N} = \left(\overline{P_F} - \overline{P_{N(F)}}\right) - \left(\overline{P_{N(F)}} - \overline{P_N}\right) + \text{torzítás}, \tag{3}$$

ahol a $\overline{P_F}$ a véletlen erdővel a férfi almintán, míg $\overline{P_N}$ az erdővel a női almintán készült modellekből az adott nemhez tartozó almintán számolt előrejelzések átlaga. A $\overline{P_{N(F)}}$ pedig a férfi modellel a női almintán számolt egyedi előrejelzések átlaga.

A heterogén bérstruktúrahatás felírása az i -edik személyre a véletlen erdővel becsült bérfüggvények esetében a következő:

$$HWS(i) = P_F(i) - P_N(i), \tag{4}$$

ahol a $P_F(i)$ a férfi, míg a $P_N(i)$ a női bérfüggvénnyel becsült bér az i-edik megfigyelésre.

A heterogén bérstruktúrahatások változónkénti elemzésével mérem az egyes magyarázóváltozók bérstruktúrára gyakorolt hatását. Az OLS és az erdő esetében is a heterogén bérstruktúrahatás kizárólag a nőknél van értelmezve. Emiatt a heterogén bérstruktúrahatás elemzését csak a nők esetében végzem el. Amennyiben ez a heterogén bérstruktúrahatás pozitív, akkor az adott nő alul-; ha pedig negatív, akkor felülárazott egy azonos munkaerőpiaci jellemzőkkel rendelkező férfihoz képest.

A heterogén bérstruktúrahatás elemzésénél első lépésben a két módszer szerint számolt bérstruktúrahatások átlagait és eloszlásait hasonlítom össze. Második lépésben a magyarázóváltozók kategóriáihoz a heterogén bérstruktúrahatásokból számolt átlagokat hasonlítom össze leíróstatisztikai módszerekkel. Az összehasonlítás megmutatja, hogy az egyes változók milyen mértékben hatnak a bérstruktúrahatás nagysá-

gára.

A 6. fejezetben a legszélsőségesebb csoportok vizsgálatánál a heterogén bérstruktúra alapján a legfelső és a legalsó tíz százalékhoz tartozó személyeket vizsgálom klasszifikációs analízissel, CLAN-nal. A CLAN megmutatja az ezekhez a csoportokhoz tartozó megfigyelések általános jellemzőit és ezek elemzése rámutat arra, hogy a magyar munkaerőpiacon milyen tulajdonságokkal rendelkeznek azok a nők, akik a legkevésbé vagy a leginkább vannak másként árazva, mint az azonos adottságokkal rendelkező férfiak.

3. Az értekezés tudományos eredményei

Bérfüggvény-becslés feltételes következtetési fákkal

- A fák és a változófontossági táblázat is egyértelműen kimutatta, hogy a foglalkozás a legfontosabb magyarázóváltozó, ami egybevág a szakirodalommal. Ahogyan a foglalkozások egyre rutinszerűbbé válnak és a FEOR kód egyre nagyobb értéket vesz fel, úgy a keresetek csökkennek.

- A második legfontosabb változó az iskolai végzettség, melynél a fákban látszott rendezettség: magasabb iskolai végzettség, nagyobb bért jelent. Azonban a végzettségi szintek keresetnövelő hatása nem egyenletes: a felsőfokú végzettség esetén megugrik a bérszint.

- Az optimális fa alapján az ágazat és a külföldi tulajdon egyformán fontos, viszont a közepes fában még a külföldi tulajdon hatása volt jelentősebb. A szurrogátumok használatával az ágazat jelentősége megemelkedett és a szurrogátumok vizsgálata rámutatott, hogy ez a változó mindegyik másik változónak jó helyettesítője lehet. Ezek az eredmények igazolják Kőrösi (2006) megállapításait mi-

szerint az ágazat hatása jelentősebb, mint a külföldi tulajdoné és hogy a kettő erősen összefügg.

- A külföldi tulajdon erősen hat a bérekre. A bármilyen külföldi tulajdonnal rendelkező vállalatok magasabb béreket fizetnek, mint a nulla százalék külföldi részesedéssel rendelkezők. Ez alátámasztja Kőrösi (2006) azonos megállapítását. A külföldi tulajdon sorrendje „kijön" a fából: a két változó közt monoton a kapcsolat.

- A változófontosságok alapján az életkor közel sem olyan fontos, mint előzetesen várható volt. A fák elemzésénél látszott a bérek és az életkor közti nem-monoton kapcsolat, azonban ez nem feltétlenül négyzetes mint ahogyan Gábor (2008) kimutatta. Az adott munkáltatónál eltöltött szolgálati idő lényegesebbnek bizonyult mint az életkor. Ez az eredmény arra enged következtetni, hogy a bérek szempontjából az adott munkahelyen megszerzett specifikus tudás lehet a meghatározó, szemben az általános munkatapasztalattal, melyet a fáknál az életkor ragadott meg.

- A régióknál a Közép-magyarországi, a Közép-dunántúli területek azok, ahol a munkavállalók jobban keresnek, ami egybevág Fazekas (2005) és Bartus (2003) eredményeivel. Fazekas (2005) kiemeli, hogy a külföldi vállalatok rendszerváltás utáni elhelyezkedését befolyásolta a nyugati határhoz való közelség, míg a magyar vállalatok elhelyezkedésére ez a tényező nem hatott. A szurrogátumok vizsgálatánál azonban az látszott, hogy a régió mind a külföldi, mind az állami részaránynak jó szurrogátuma volt. Tehát a saját eredmények részben igazolják Fazekas (2005) állítását és arra mutatnak, hogy mindkét vállalati kör elhelyezkedésére hatnak földrajzi tényezők, azonban ez a tényező nem feltétlenül a nyugati határhoz való közelség.

- A főváros általában a jobban fizető területekhez tartozik. Azon-

ban a nagy fában Budapest mellé sok esetben bekerült az egyéb településtípus is. Emiatt Köllő (2003) eredménye, miszerint Budapest jelentős bérelőnnyel rendelkezik a többi településtípussal összevetve, egyértelműen nem igazolható.

- A nem szerinti megkülönböztetés látható a CTree-ben, mivel egy kivétellel a férfiak mindig a jobban fizető csoportot jelentik. Továbbá igazolva Csillag (2006) eredményeit a férfiak és a nők foglalkozásbeli megosztottsága egyre kevésbé van jelen a magyar munkaerőpiacon, azonban látszik a nemek ágazati megosztottága.

- Rigó (2012) bemutatta, hogy Magyarországon a kollektív szerződések létrejötte nem független az ágazattól és a földrajzi régiótól. A saját eredmények alapján a kollektív szerződések jó helyettesítője lehet az ágazat és a régió is. Továbbá Rigó (2012) megállapította, hogy a kollektív szerződések bérekre gyakorolt hatása alacsony, amit a CTree alapján számolt változófontossági táblázat alátámasztott. Rigó (2012) mindkét megállapítását a CTree tudta igazolni.

Nemi bérkülönbségek Magyarországon: OLS regresszióval és regressziós erdővel kapott eredmények összehasonlítása

- Az OLS-sel és a regressziós erdővel kapott becslések összehasonlítása rámutat arra, hogy az erdővel készített bérelőrejelzések pontosabbak. A véletlen erdők használatát igazolja, hogy a Blinder-Oaxaca felbontásban megjelenő torzítás elfogadhatóan kicsi. Tehát a módszertani váltás nem befolyásolja jelentősen a torzítottságot és a bérelőrejelzések is pontosabbak.

- A felbontás alapján az összetételhatás kisebb a véletlen erdő esetében, ami miatt bérstruktúrahatás nagyobb. Azonban a két

eljárással számolt átlagos bérstruktúrahatások így is hasonlóan alakulnak.

- A heterogén bérstruktúrahatás vizsgálata alapján bár az átlagok hasonlóak, a két eloszlás statisztikailag szignifikánsan eltér egymástól és a becslések közti korreláció is csak közepes mértékű. Az eredmények arra mutatnak, hogy a két módszer alapján van különbség aközött, hogy egy nő milyen mértékben van félreárazva egy azonos adottságú férfihoz képest. A heterogén bérstruktúrahatás vizsgálata alapján ha egy kategóriához több megfigyelés tartozik, akkor a két módszer szerinti átlagos bérstruktúrahatások közelebb esnek egymáshoz. A leíró statisztikai elemzés alapján nincs egy változó, amely felelős lenne a két módszerrel számolt átlagos bérstruktúrahatások eltéréséért. Inkább a kevesebb megfigyelést tartalmazó csoportok közti átlagos bérstruktúrahatások közötti eltérésre vezethető vissza az eloszlások különbözősége és a becslési eredmények közti közepes korreláció.

- A megfigyeléseim igazolják Weichselbaumer–Winter-Ebmer (2005) eredményeit miszerint a dekompozíciónál a különböző módszerek hasonló eredményekre vezetnek, azonban a változókat másként értékelik az egyes módszerek. Tehát ha az egyedi változók hatásait akarjuk vizsgálni, akkor körültekintően kell megválasztani a módszertant, mert az befolyásolja az egyes változóknál mért összetétel- és bérstruktúrahatás nagyságát.

Heterogén bérstruktúrahatások időbeli alakulása és a legszélsőségesebb csoportok

- A regressziós erdő és a magasabb rendű tagokat is tartalmazó OLS teljesítőképességének összehasonlítása alapján az erdő jobb előrejelzési képességgel rendelkezik, mint az OLS becslés. A teszt-

mintán mindkét módszerrel számolt átlagos négyzetes hibák emelkednek, azonban a regressziós erdő hibái így is alacsonyabbak.

- A Blinder-Oaxaca felbontás összehasonlításánál a regressziós erdővel kapott összetételhatások általában alacsonyabbak, mint az OLS-becsléssel kapottak, azonban nagyságrendileg megegyeznek és azonos irányba mutatnak. Az összetételhatás kezdetben negatív volt és növekvő tendenciát mutatott. A nyers bérkülönbség időben emelkedett, a bérstruktúrahatás a vizsgált időszakban állandónak tekinthető, ami megegyezik Lovász (2009) eredményeivel. A szerző kimutatta, hogy Magyarországon a rendszerváltást követően a bérstruktúrahatás drasztikusan lecsökkent, majd 1992-től sokkal kisebb ütemben mérséklődött, szinte stagnált.

- A negatív összetételhatás vélhetően a rendszerváltást követő folyamatok hosszútávú eredménye. A rendszerváltást követően az oktatás megtérülése nőtt Magyarországon és mivel a nők átlagosan képzettebbek voltak, mint a férfiak, ami az összetételhatás növekedése irányába hatott. (Kertesi–Köllő (1996), Brainerd (2000)) Emellett a nők munkakínálata csökkent Magyarországon ebben az időszakban (Brainerd (2000)), mivel vélhetően az alacsonyabban képzett, kevésbé termelékeny nők kiléptek a munkaerőpiacról. Így összességében a nők átlagos munkaerőpiaci jellemzői javultak.

- A magyar munkaerőpiacon a férfiak jelentős része rendelkezik középfokú végzettséggel, de érettségivel nem, míg a nőknél az érettségivel rendelkezők vannak a legtöbben. A tendenciák az évek között hasonlók mindkét nemnél: nő a felsőfokú végzettségűek aránya, mialatt az érettségivel nem rendelkezők aránya csökken. Azonban ez az átmenet nagyon lassan megy végbe.

- 2008 és 2016 között a férfiak aránya a 100 százalékban külföl-

di tulajdonban lévő vállalatoknál növekedett, míg a nőké inkább csökkent. Ez az átmenet – mivel a külföldi vállalatok általában magasabb béreket fizetnek (Kőrösi (2006)) – az összetételhatás növekedése irányába hatott.

- A heterogén bérstruktúrahatás vizsgálata rámutatott, hogy az átlagos hatások és a végzettségi szintek kapcsolata nem monoton. Az átlagos bérstruktúrahatások az életkor és a szolgálati idő növekedésével kezdetben jobban emelkednek, majd enyhén csökkennek, de továbbra is pozitívak, ami megerősíti Gábor (2008) eredményeit.

- Az átlagos bérstruktúrahatások és a külföldi részarány összevetésénél monoton kapcsolat látszik: a külföldi tulajdon csökkenésével az átlagos hatás is mérséklődik. Köllő (2000) felveti a lehetőségét annak, hogy a rendszerváltás után azok a cégek kerültek külföldi kézre, ahol nagyobbak voltak a nemi bérkülönbségek, emiatt a külföldi cégeknél növekedett az átlagos bérkülönbség, míg a hazaiaknál csökkent. Vélhetően ennek a változásnak a hosszútávú hatása figyelhető meg a saját eredményekben is.

- A CLAN elemzés alapján azok a nők, akiknél a bérstruktúrahatás alacsony volt, nem túl nagy hazai vállalatoknál dolgoznak a szolgáltatószektorban a Közép-magyarországi régióban. Ezeket a nőket a bérfüggvények alapján hasonlóan vagy még jobban megfizethetik, mint a férfiakat. A legnagyobb bérstruktúrahatással rendelkező nők általában nagyobb, külföldi tulajdonú, feldolgozóipari vállalatoknál dolgoznak a dunántúli régiókban.

- A CLAN elemzés eredményei alátámasztják a monopszonikus munkaerőpiac feltételezését. Az elmélet szerint a nemek közti eltérő ingázási, illetve keresési költségek, valamint nem pénzügyi juttatások fontossága miatt a férfiak és a nők vállalati rugal-

massága különbözik, ami a munkáltató erőfölényéhez vezethet. (Hirsch (2009), Hirsch–Schnabel (2010), Webber (2013), Heinze–Wolf (2010)) A központi régióban a nők könnyebben juthatnak új állásokhoz, jobb munkaerőpiaci pozícióval rendelkeznek, a munkáltatók közti verseny feltehetően nagyobb, ami miatt a vállalat monopszonikus ereje kisebb lehet ezen a földrajzi területen. Emellett az legkisebb bérstruktúrahatással rendelkező csoport tagjai között nagyobb arányban vannak azok, akik legalább érettségivel rendelkeznek – érettségizettek és diplomások együtt –, ami tovább erősíti a munkaerőpiaci helyzetüket ezeknek a nőknek. Természetesen még ebben a szegmensben is előfordulhat, hogy a nők vállalati rugalmassága kisebb, mint a férfiaké, azonban így már alacsonyabb az esélye annak, hogy a nemi bérkülönbség a vállalat monopszonikus erejéből származzon. A nemek között továbbra is megjelenhetnek termelékenységi és/vagy diszkriminációból fakadó különbségek.

4. Főbb hivatkozások

Athey, S. – Tibshirani, J. – Wager, S. (2019): Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178, DOI: http://dx.doi.org/doi10.1214/18-AOS1709.

Bartus, T. (2003): Ingázás. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2003*, pp. 88–101. MTA Közgazdaságtudományi Kutatóközpont.

Becker, G. S. (1964): *Human Capital*. Columbia University Press for the National Bureau of Economic Research, New York, NY, first edition.

Becker, G. S. (1971): *The Economics of Discrimination*. The University of Chicago Press, Chicago.

Blanchflower, D. G. – Oswald, A. J. – Sanfey, P. (1992): Wages, profits and rent-sharing. Working Paper 4222, National Bureau of Economic Research, DOI: http://dx.doi.org/doi10.3386/w4222.

Blinder, A. S. (1973): Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8(4):436–455, DOI: http://dx.doi.org/doi10.2307/144855.

Brainerd, E. (2000): Women in transition: Changes in gender wage differentials in eastern europe and the former soviet union. *Industrial and Labor Relations Review*, 54(1):138–162, DOI: http://dx.doi.org/doi10.2307/2696036.

Chamberlain, G. (1991): Quantile regression, censoring, and the structure of wages. In Sims, C. A., szerk., *Advances in Econometrics*, pp. 171–210. Cambridge University Press, DOI: http://dx.doi.org/doi10.1017/ccol0521444594.005.

Chernozhukov, V. – Demirer, M. – Duflo, E. – Fernández-Val, I. (2018): Generic machine learning inference on heterogenous treatment effects in randomized experiments. NBER Working Paper 24678, National Bureau of Econimic Research, DOI: http://dx.doi.org/doi10.3386/w24678.

Csillag, M. (2006): Női munka és nemek szerinti kereseti különbségek a késő szocializmustól napjainkig. In Fazekas, K. – Kézdi, G., szerk., *Munkaerőpiaci tükör 2006*, pp. 100–105. MTA Közgazdaságtudományi Intézet.

Elek, P. – Scharle, Á. – Szabó, B. – Szabó, P. A. (2009): A bérekhez kapcsolódó adóeltitkolás magyarországon. Közpénzügyi füzetek 23, Pénzügyminisztérium.

Fazekas, K. (2005): A hazai és a külföldi tulajdonú vállalkozások területi koncentrációjának hatása a foglalkoztatás és munkanélküliség területi különbségeire. In Faluvégi, A. – Fazekas, K. – Nemes-Nagy, J. – Németh, N., szerk., *A hely és a fej: Munakpiac és regionalitás Magyarországon, Budapest*, pp. 47–74. MTA Közgazdaságtudományi Kutatóközpont.

Fortin, N. – Lemieux, T. – Firpo, S. (2011): Decomposition Methods in Economics. In Ashenfelter, O. – Card, D., szerk., *Handbook of Labor Economics*, volume 4 of *Handbook of Labor Economics*, chapter 1, pp. 1–102. Elsevier, DOI: http://dx.doi.org/doi10.1016/s0169-7218(11)00407-2.

Gábor, R. I. (2008): A hiányzó láncszem? Életpálya-keresetek és keresetingadozás. *Közgazdasági Szemle*, 55(december):1057–1074.

Hastie, T. – Tibshirani, R. – Friedman, J. (2009): *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition, DOI: http://dx.doi.org/doi10.1007/b94608.

Heckman, J. J. (1979): Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, DOI: http://dx.doi.org/doi10.2307/1912352.

Heinze, A. – Wolf, E. (2010): The intra-firm gender wage gap: a new view on wage differentials based on linked employer–employee data. *Journal of Population Economics*, 23(3):851–879, DOI: http://dx.doi.org/doi10.1007/s00148-008-0229-0.

Hirsch, B. (2009): The gender pay gap under duopsony: Joan robinson meets harold hotelling. *Scottish Journal of Political Economy*, 56(5):543–558, DOI: http://dx.doi.org/doi10.1111/j.1467-9485.2009.00497.x.

Hirsch, Borisand Schank, T. – Schnabel, C. (2010): Differences in labor supply to monopsonistic firms and the gender pay gap: An empirical analysis using linked employer-employee data from germany. *Journal of Labor Economics*, 28(2):291–330, DOI: http://dx.doi.org/doi10.1086/651208.

Inglis, A. – Parnell, A. – Hurley, C. B. (2022): Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, pp. 1–13, DOI: http://dx.doi.org/doi10.1080/10618600.2021.2007935.

James, G. – Witten, D. – Hastie, T. – Tibshirani, R. (2013): *An Introduction to Statistical Learning: with Applications in R*. Springer, DOI: http://dx.doi.org/doi10.1007/978-1-4614-7138-7.

Jann, B. (2008): The blinder–oaxaca decomposition for linear regression models. *Stata Journal: Promoting communications on statistics and Stata*, 8(4):453–479, DOI: http://dx.doi.org/doi10.1177/1536867x0800800401.

Jiao, S. R. – Song, J. – Liu, B. (2020): A review of decision tree classification algorithms for continuous variables. *Journal of Physics: Conference Series*, 1651(1):012–083, DOI: http://dx.doi.org/doi10.1088/1742-6596/1651/1/012083.

Kertesi, G. – Köllő, J. (1996): A bér alakulását meghatározó tényezők. In Halpern, L., szerk., *Bérköltség és versenyképesség*, pp. 74–76. MTA Közgazdaságtudományi Intézet, január.

Köllő, J. (2000): Tulajdoni szektorok. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2000*, pp. 100–106. MTA Közgazdaságtudományi Kutatóközpont.

Köllő, J. (2003): Regionális kereseti és bérköltség-különbségek. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2003*, pp. 65–78. MTA Közgazdaságtudományi Kutatóközpont.

Kőrösi, G. (2006): Vállalatok közti bérkülönbségek dinamikája. In Fazekas, K. – Kézdi, G., szerk., *Munkaerőpiaci tükör 2006*, pp. 48–59. MTA Közgazdaságtudományi Kutatóközpont.

Levshina, N. (2020): Conditional inference trees and random forests. In Paquot, M. – Gries, S. T., szerk., *A Practical Handbook of Corpus Linguistics*, pp. 611–643. Springer International Publishing, Cham, DOI: http://dx.doi.org/doi10.1007/978-3-030-46216-1_25.

Lovász, A. (2009): A verseny hatása a női-férfi bérkülönbségre magyarországon 1986 és 2003 között. In Fazekas, K. – Lovász, A. – Telegdy, Á., szerk., *Munkaerőpiaci tükör 2009*, pp. 149–158. MTA Közgazdaságtudományi Intézet.

Lovász, A. (2013): Jobbak a nők esélyei a közszférában? a nők és férfiak bérei közötti különbség és a foglalkozási szegregáció vizsgálata a köz- és magánszférában. *Közgazdasági Szemle*, 60:1057–1074.

Mincer, J. (1958): Investment in human capital and personal income distribution. *Journal of Political Economy*, 66(4):281–302, DOI: http://dx.doi.org/doi10.1086/258055.

Oaxaca, R. (1973): Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709, DOI: http://dx.doi.org/doi10.2307/2525981.

Reilly, B. (2001): A nemek közötti bérkülönbségek elemzésének statisztikai módszerei. *Statisztikai Szemle*, 79(1):5–17.

Rigó, M. (2012): A szakszervezeti bérrés becslése magyarországon. In Fazekas, K. – Benczúr, P. – Telegdy, Á., szerk., *Munkaerőpiaci tükör 2012*, pp. 200–214. MTA Közgazdaságtudományi Kutatóközpont.

Strobl, C. – Boulesteix, A.-L. – Kneib, T. – Augustin, T. – Zeileis, A. (2008): Conditional variable importance for random forests. *BMC bioinformatics*, 9:307, DOI: http://dx.doi.org/doi10.1186/1471-2105-9-307.

Takács, O. (2021): Nemek közötti bérkülönbségek magyarországon: a véletlenerdő- és az ols-becslésen alapuló blinder–oaxaca-dekompozíció eredményeinek összehasonlítása. *Statisztikai Szemle*, 99(1):5–45, DOI: http://dx.doi.org/doi10.20311/stat2021.1.hu0005.

Takács, O. – Vincze, J. (2018): Bérelőrejelzések – prediktorok és tanulságok. *Közgazdasági Szemle*, 55(6):592–618, DOI: http://dx.doi.org/doi10.18414/ksz.2018.6.592.

Takács, O. – Vincze, J. (2020): The gender-dependent structure of wages in hungary: results using machine learning techniques. KRTK-

KTI műhelytanulmány 44, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Wachtel, H. M. – Betsey, C. (1972): Employment at low wages. *The Review of Economics and Statistics*, 54(2):121–129, DOI: http://dx.doi.org/doi10.2307/1926272.

Webber, D. A. (2013): Firm-level monopsony and the gender pay gap. IZA Discussion Papers 7343, Institute of Labor Economics (IZA).

Weichselbaumer, D. – Winter-Ebmer, R. (2005): A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*, 19(3):479–511, DOI: http://dx.doi.org/doi10.1111/j.0950-0804.2005.00256.x.

5. A témakörrel kapcsolatos saját (ill. társszerzős) publikációk jegyzéke

Takács, O. – Vincze, J. (2018a): Bérelőrejelzések - prediktorok és tanulságok. *Közgazdasági Szemle*, 55(6):592–618, DOI: http://dx.doi.org/10.18414/ksz.2018.6.592.

Takács, O. – Vincze, J. (2018b): The within-job gender pay gap in Hungary. Közgazdaságtudományi Intézet műhelytanulmány 34, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Takács, O. – Vincze, J. (2019a): Blinder-oaxaca decomposition with recursive tree-based methods: a technical note. Közgazdaságtudományi Intézet műhelytanulmány 23, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Takács, O. – Vincze, J. (2019b): The gender pay gap in hungary: new results with a new methodology. Közgazdaságtudományi Intézet műhelytanulmány 24, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.

Takács, O. – Vincze, J. (2020): The gender-dependent structure of wages in Hungary: results using machine learning techniques. CERS-IE WORKING PAPERS 2044, Institute of Economics, Centre for Economic and Regional Studies.

Takács, O. (2021): Nemek közötti bér különbségek Magyarországon: a véletlenerdő- és az OLS-becslésen alapuló Blinder–Oaxaca dekompozíció eredményeinek összehasonlítása. *Statisztikai Szemle*, 99(1): 5–45. DOI: http://dx.doi.org/10.20311/stat2021.1.hu0005