

Takács Olga

Bérfüggvények becslése
Magyarországon
gépi tanuló algoritmusokkal

Közgazdaságtan Intézet

Témavezető:

Vincze János, DSc

egyetemi tanár

©Takács Olga

Budapesti Corvinus Egyetem
Közgazdasági és Gazdaságinformatikai Doktori Iskola

Bérfüggvények becslése Magyarországon
gépi tanuló algoritmusokkal
doktori értekezés

Takács Olga

Budapest, 2022

Tartalomjegyzék

Ábrák jegyzéke	8
Táblázatok jegyzéke	12
1. Bevezetés	16
2. Magyar munkaerőpiaci sajátosságok	21
2.1. Bérek alakulására ható tényezők Magyarországon	21
2.2. Nemi bérkülönbségek Magyarországon	25
2.3. Bértarifa adatbázis	29
3. Gépi tanuló algoritmusok	32
3.1. Feltételes következtetési fa – CTree	38
3.2. Feltételes véletlen erdő – CForest	40
3.3. Feltételes változófontossági érték	41
3.4. Általánosított véletlen erdő – regressziós erdő	44
3.5. Klasszifikációs analízis – CLAN	48

4. Bérfüggvény-becslés feltételes következtetési fákkal	49
4.1. Bevezetés	49
4.2. Adatbázis és módszertan	52
4.3. Eredmények	54
4.3.1. Legkisebb fa	56
4.3.2. Közepes fa	57
4.3.3. Optimális fa	60
4.3.4. Feltételes változófontossági mértékek	64
4.4. Robusztusságvizsgálat	69
4.5. Összefoglalás	71
5. Nemi bérkülönbségek Magyarországon: OLS regresszióval és regressziós erdővel kapott eredmények összehasonlítása	74
5.1. Bevezetés	74
5.2. Adatbázis	77
5.3. Módszertan	79
5.3.1. Blinder-Oaxaca felbontás elméleti oldalról	79
5.3.2. Blinder-Oaxaca felbontás gyakorlati oldalról	82
5.3.3. Regressziós erdő hiperparaméterei	84
5.4. Eredmények	85
5.4.1. Férfi és női bérfüggvények összehasonlítása	86
5.4.2. Blinder-Oaxaca felbontás	88
5.4.3. Heterogén bérstruktúrahatások	90

5.4.4. Heterogén bérstruktúrahatások változonként	93
5.5. Összefoglalás	102
6. Heterogén bérstruktúrahatások időbeli alakulása és a legszélsőségebb csoportok	104
6.1. Bevezetés	104
6.2. Adatbázis és módszertan	105
6.3. Eredmények	108
6.3.1. Blinder-Oaxaca felbontás	110
6.3.2. Heterogén bérstruktúrahatások	113
6.3.3. Extrém csoportok meghatározása klasszifikációs analízissel	117
6.4. Diskusszió	119
7. Összefoglalás	123
Irodalomjegyzék	127
Függelékek	143
F.1. Függelék: Bértarifa adatbázisban szereplő változók leírása . . .	143
F.2. Függelék: Foglalkozás besorolás és a hozzájuk kapcsolódó végzettségi szintek	148
F.3. Függelék: Adatbázist leíró statisztikák	149
F.4. Függelék: Közepes fa	152
F.5. Függelék: Szurrogátumok a változóknál	154
F.6. Függelék: A 100 véletlen erdő leíró statisztikái	160

F.7. Függelék: Heterogén bérstruktúrahatás elemzéséhez tartozó változónkénti megoszlások	161
F.8. Függelék: OLS-becslésben szereplő változók	166
F.9. Függelék: Átlagos heterogén bérstruktúrahatás és az egyes változók szerinti relatív gyakoriságai	168
Saját publikációk a témában	175

Ábrák jegyzéke

1.	Döntési fa szemléltetése	35
2.	Döntési fa eljárások csoportosítása	36
3.	Feltételes permutációs módszer illusztrációja	43
4.	Változótér-felosztás illusztrálása a feltételes változófontosságoknál	44
5.	Illusztráció a regressziós erdő fáinak aggregálására	45
6.	Determinációs együttható a CTree mélységének függvényében .	55
7.	Legkisebb fa	56
8.	Közepes fa	58
9.	Szolgálati idő szerinti vágások megoszlása az optimális fában . .	63
10.	Átlagos négyzetes eltérés a regressziós erdővel (GRF) és az OLS regresszióval (OLS) készült nemi bérfüggvények esetében a tanuló- és a tesztmintán	86
11.	Átlagok megoszlása a 100 erdő esetében a tanulómintán	87
12.	Átlagok megoszlása a 100 erdő esetében a tesztmintán	88
13.	Blinder-Oaxaca felbontás a regressziós erdő (GRF) és az OLS regresszió (OLS) alapján a tanuló- és a tesztmintán	89

14.	Heterogén bérstruktúrahatások az OLS regresszióval (OLS) és a regressziós erdővel (GRF) becsült bérfüggvények esetében a tanulómintán	90
15.	Heterogén bérstruktúrahatás eloszlása a két módszer szerint a tanulómintán	91
16.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az iskolai végzettségnél	93
17.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az életkornál	94
18.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai a szolgálati éveknél	95
19.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az egyjegyű FEOR-oknál	96
20.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai a kétjegyű FEOR-oknál	97
21.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az ágazatoknál	98
22.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az állami és külföldi tulajdonú vállalatoknál	99
23.	Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai a magyar régiókban a különböző településtípusoknál	100
24.	Átlagos bérstruktúrahatás és azok 95 százalékos konfidencia intervallumai a vállalatméret (létszám) függvényében	101
25.	Átlagos négyzetes hiba az OLS és a regressziós erdő (GRF) esetében	109
26.	Nyers bérkülönbségek és bérstruktúrahatások 2008 és 2016 között	110

27. Blinder-Oaxaca felbontás (torzítás nélkül) 2008 és 2016 között (nyers bérkülönbség százalékában)	112
28. Heterogén bérstruktúrahatás és iskolai végzettség	113
29. Heterogén bérstruktúrahatás és az életkor kategóriák	114
30. Heterogén bérstruktúrahatás és a szolgálati idő	115
31. Heterogén bérstruktúrahatás és a külföldi részarány	115
32. Egyéni bérstruktúrahatás és a női részarány	116
F.7.1Megfigyelések megoszlása iskolai végzettség szerint	161
F.7.2Megfigyelések megoszlása az életkor szerint	162
F.7.3Megfigyelések megoszlása a szolgálati idő (évek) szerint	162
F.7.4Megfigyelések megoszlása a foglalkozások – egyjegyű FEOR- kódok – szerint	163
F.7.5Megfigyelések megoszlása a foglalkozások – kétjegyű FEOR- kódok – szerint	163
F.7.6Megfigyelések megoszlása a nemzetgazdasági ágak – egyjegyű TEÁOR-kódok – szerint	164
F.7.7Megfigyelések megoszlása a tulajdontípusok szerint	164
F.7.8Megfigyelések megoszlása a régió és a településtípusok szerint	165
F.7.9Megfigyelések megoszlása a vállalatméret (létszám) szerint	165

Táblázatok jegyzéke

1.	Felhasznált magyarázóváltozók köre	30
2.	Havi keresetek logaritmusának leíró statisztikája	52
3.	Hiperparaméterek a különböző méretű fáknál	53
4.	Hiperparaméterek az erdőnél	54
5.	Különböző nagyságú fák leíróstatisztikája	55
6.	Ágazatok felosztása a közepes fában	59
7.	Legmagasabb iskolai végzettség felosztása az optimális fában	61
8.	Régiók felosztása a közepes fában	61
9.	Ágazatok felosztása az optimális fában	62
10.	Feltételes változófontossági mértékek (százalék) és vágások száma (darab) a különböző fáknál	65
11.	A szurrogátumok nélkül és az azokkal számolt feltételes változófontossági mértékek (százalék) az optimális fában	67
12.	Változófontossági mértékek a CTreenél és a CForestnél (százalék), illetve azok különbsége (százalékpont)	70
13.	Női és férfi bérek leíróstatisztikái a tanuló- és a tesztmintán	78
14.	Regressziós erdő hiperparaméterei	85

15.	Heterogén bérstruktúrahatalás leíró statisztikái a tanulómintán és a kilógó értékek (outlierek) nélküli mintán	92
16.	A bérelőrejelzések átlagai (a bérek logaritmusában mérve) a kilógó értékek nélküli mintán	92
17.	Adatok leíró statisztikái	106
18.	Regressziós erdő hiperparaméterei a férfiaknál	107
19.	Regressziós erdő hiperparaméterei a nőknél	107
20.	A nyers bérkülönbség Blinder-Oaxaca felbontása regressziós erdővel (GRF) és OLS regresszióval (OLS) a tanulómintán	111
21.	A legnagyobb és a legkisebb bérstruktúrahatalással rendelkező csoportok (M- és A-csoportok) megoszlásainak különbsége (százalékpont)	118
22.	Iskolai végzettség relatív gyakorisága nemek szerint (%)	120
23.	Külföldi részarány relatív gyakorisága nemek szerint (százalék)	121
F.1.1	Foglalkozások – FEOR – lista	145
F.1.2	Nemzetgazdasági ágak – TEÁOR – listája	147
F.2.1	Foglalkozásokhoz – FEOR-hoz kötött – elvárt iskolai végzettségi szintek	148
F.3.1	Nemek szerinti megoszlás	149
F.3.2	Iskolai végzettség szerinti megoszlás	149
F.3.3	Foglalkozás szerinti megoszlás	149
F.3.4	Régió szerinti megoszlás	150
F.3.5	Településtípus szerinti megoszlás	150
F.3.6	Ágazat szerinti megoszlás	150

F.3.7	Kollektív szerződések szerinti megoszlás	151
F.3.8	Tulajdoni részarány szerinti megoszlás	151
F.3.9	Folytonos változók leíró statisztikái	151
F.4.1	Közepes fa	152
F.5.1	A legfontosabb szurrogátumok az optimális fában	155
F.5.2	A szurrogátumok használatának százalékos aránya az összes szurrogátumhoz viszonyítva	158
F.8.1	Változók az OLS becslésekben 2008 és 2016 között	166
F.9.1	A heterogén bérstruktúrahatás (BS hatás) nagysága és az egyes változók relatív gyakoriságai (százalék) a tanulómintában	168
F.9.2	A legkisebb bérstruktúrahatással (BS hatás) rendelkező csoportban (A-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tanulómintán	169
F.9.3	A legnagyobb bérstruktúrahatással (BS hatás) rendelkező csoportban (M-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tanulómintán	170
F.9.4	A legnagyobb és a legkisebb bérstruktúrahatással rendelkező csoportok (M- és A-csoportok) különbsége (százalékpont) a tesztmintán	171
F.9.5	A heterogén bérstruktúrahatás (BS hatás) nagysága és az egyes változók relatív gyakoriságai (százalék) a tesztmintában	172
F.9.6	A legkisebb bérstruktúrahatással (BS hatás) rendelkező csoportban (A-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tesztmintán	173
F.9.7	A legnagyobb bérstruktúrahatással (BS hatás) rendelkező csoportban (M-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tesztmintán	174

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Vincze Jánosnak, aki a kezdetektől fogva támogatott a kutatói pályámon, illetve értékes idejével és szaktudásával mindig támogatta a közös és az egyéni kutatásokat is. Az együttműködésünk során rengeteget tanultam tőle. Köszönöm a műhelyvitára felkért bírálóknak, Elek Péternek és Szabó-Morvai Ágnesnek a szakmai hozzászólásaikat és javaslataikat az értekezéstervezethez. Tovább szeretnék köszönetet mondani a családomnak és a barátaimnak, akik mindig lelkesítettek és kitartásra biztattak.

1. fejezet

Bevezetés

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 vállalatoknál? A kollektív szerződések ténylegesen képesek hatni a bérekre? Melyik ágazatban, foglalkozásban fizetnek a legtöbbet és melyikben 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 munkatapasztalat 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 javításával igyekeztek hozzájárulni a bérezési kérdéseket érintő strukturális összefüggések megértéséhez. (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 esz-

közzé vált a csoportok közti különbségek vizsgálatánál¹ (Jann (2008), Fortin et al. (2011)), és kiemelten használják a nemi bérkülönbségek vizsgálatánál. (Weichselbaumer–Winter-Ebmer (2005)) A dekompozíció alkalmas arra, hogy az átlagos különbségen belül elkülönítse a munkaerőpiaci jellemzőkkel magyarázott és nem magyarázott részt. A felbontáshoz pedig minden esetben szükséges a vizsgált csoportokhoz tartozó egyéni bérfüggvények becslése.

Az egyéni bérfüggvények meghatározásánál és a Blinder-Oaxaca dekompozíciónál 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 regresszió 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. Remek példája ennek az a vita, melyben azt taglalták, hogy az iskolában eltöltött évek száma vagy az iskolai végzettség a meghatározó a kereset szempontjából. Az iskolaévek mellett érvelők azt állították, hogy a plusz egy iskolában töltött év keresletnövelő hatása egyenletes, míg a végzettség mellett érvelők a megszerzett fokozatot tekintették a szellemi képesség indikátorának. (Hungerford–Solon (1987), Jaeger–Page (1996), Rodríguez–Muro (2015)) A két megközelítés hívei így mindenképpen másként mérik és használják fel az adott változót a becsléseikben.

A bérfüggvények meghatározása kapcsán továbbra is lehetnek ehhez hasonló – 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 össze-

¹A Blinder-Oaxaca dekompozícióval bármilyen csoportok közötti átlag felbontható magyarázott és nem magyarázott részre. Ezzel a dekompozícióval például vizsgálták a boldogságot (Arrosa–Gandelman (2016)) és az elhízást (Taber et al. (2016)) a férfiak és a nők között.

hasonlítva más feltételezésekből indulnak ki². 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. A torzítás minimalizálása helyett a tanuló algoritmusok a variancia és a torzítás együttes minimalizálására, valamint a legjobb előrejelzések készítésére törekszenek.

A disszertációmban 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. Ezt a bérfüggvénybecslést és a változók közti kapcsolatok elemzését tartalmazza a 4. fejezet. A tanuló algoritmusok alkalmazásának további célja a lehető legpontosabb előrejelzések készítése, melyeket a nemi bérkülönbségek vizsgálatánál használok ki az 5. és a 6. fejezetben. Az átlagos nemi bérkülönbségeket leggyakrabban Blinder-Oaxaca dekompozícióval elemzik, melyhez mindenképpen szükséges egy férfi és egy női bérfüggvény becslése. Leggyakrabban ezeket a bérfüggvényeket OLS regresszióval becsülték. A disszertációmban azonban gépi tanuló algoritmussal becslem meg ezeket a függvényeket. A célom annak bemutatása, hogy a bérfüggvények becslési módszertanának megváltoztatása milyen mértékben befolyásolja a Blinder-Oaxaca dekompozícióból kapott eredményeket, majd azok további elemzése.

A disszertációm felépítése a következő. A 2. fejezet tartalmazza a kapcsolódó szakirodalom Magyarországra vonatkozó eredményeinek összefoglalását. A 2.1. alfejezet foglalja össze azokat a kutatási eredményeket, melyek a bérekre ható tényezőket vizsgálták. A 2.2. alfejezet a nemi bérkülönbségekkel kapcsolatos kutatásokat tárgyalja. A 2.1. és 2.2. alfejezetekben bemutatott kutatások jelentős része a Bértarifa adatbázis felhasználásával készült. Az értekezésemben én is ezt az adatbázist használom és az ismételések elkerülése érdekében egyszer, egyben mutatom be az adatokat a 2.3. alfejezetben.

²Az általam használt feltételes következtetési fák (conditional inference tree, CTree) és feltételes véletlen erdők (conditional random forest, CForest) esetében Levshina (2020) kiemeli, hogy ezeknek az eljárásoknak az alkalmazásánál nem kell megkövetelni a homoszkedaszticitásra, illetve a hibatagok normális eloszlására vonatkozó feltételezést.

A gépi tanulási algoritmusokat és néhány interpretációs eljárást, amelyeket a későbbi fejezetekben használok, a 3. fejezetben mutatom be. Ebben a fejezetben tárgyalom a feltételes következtetési fákat (3.1. alfejezet), a feltételes véletlen erdőket (3.2. alfejezet) és az ezekből számolt feltételes változófontossági értékek módszertanát (3.3. alfejezet), az általánosított véletlen erdőket (3.4. alfejezet), valamint a klasszifikációs analízis módszerét (3.5. alfejezet).

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. Gondolhatunk itt például a regressziókra, melyekben előre meg kell határozni a függvényformát és a változók közötti kapcsolatokat. 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 véletlen erdőre átirított 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 mértékben 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. Ez a fejezet egyrészt tekinthető úgy, mint az 5. fejezetben bemutatott eredmények robusztusságvizsgálata, mivel az 5. fejezetben egy egyszerűbb, míg a 6.-ban egy magasabb rendű tagokat és interakciókat is tartalmazó OLS becslés eredményét hasonlítom össze egy erdőből kapottal. Másrészt rámutat, hogy a Blinder-Oaxaca felbontásból kapott magyarázott

és nem magyarázott rész nagysága hogyan alakult a vizsgált időszakban. A fejezet legvégén elkülönítem azokat a társadalmi csoportokat, amelyek a leginkább, illetve a legkevésbé járultak hozzá az átlagos nemi bérkülönbségek kialakulásához és fennmaradásához. A 7. fejezet összefoglalja az értekezés eredményeit.

2. fejezet

Magyar munkaerőpiaci sajátosságok

A 2.1. alfejezetben azokat a kutatási eredményeket mutatom be, amelyek Magyarországra vonatkozóan vizsgálták a bérek alakulására ható tényezőket. A 2.2. alfejezetben a nemi bérkülönbségekkel foglalkozom. Ebben az alfejezetben bemutatom a nemi bérkülönbségek mérésével kapcsolatos problémákat és a Magyarországra vonatkozó kutatási eredményeket. A szakirodalom feldolgozás során eltekintek a hazai kutatások alapjául szolgáló irodalom feltárásától és a nemzetközi eredmények bemutatásától. A 2.3. alfejezet a magyar munkaerőpiaci kutatásokban – és a disszertációmban is használt – Bértarifa adatbázis leírását tartalmazza.

2.1. Bérek alakulására ható tényezők Magyarországon

A magyarországi béreket Gábor (2008) vizsgálta Mincer-típusú bérfüggvénnyel és igazolta, hogy Mincer (1958, 1974) korábbi megfigyelései Magyarországon is érvényesek: 1) A magasabb iskolai végzettség esetében a bérgörbék

magasabbról indulnak. 2) A gyakorlati idő és a bérek között Magyarország esetében is konkáv a kapcsolat. Továbbá megállapította, hogy az azonos iskolai végzettséggel rendelkező férfiak és nők esetében a férfiak bérfüggvényei magasabbról indulnak és „erősebben konkávak” (Gábor (2008), p. 1059).

Földrajzi tényezőket figyelembe véve a magyar régiók közötti bérkülönbségek fennállása leginkább keresleti oldali tényezőkkel magyarázható. Ugyanis a rendszerváltást követően a megszűnők helyett az új munkahelyek leginkább a városiasodott, magasabb iskolai végzettséggel és jobb infrastruktúrával rendelkező területekre koncentráálódtak. A rendszerváltást követően a nyugati határhoz való közelség erősen hatott a külföldi tulajdonú vállalatok földrajzi elhelyezkedésére, míg a magyarokét nem befolyásolta. Az ipari hagyományok és a magasabb iskolázottságú területek helyzete számított a külföldi és a magyar tulajdonosoknak is (Fazekas (2005)), így hozva létre jelentős területi különbségeket a munkanélküliség tekintetében. Továbbá jellemző, hogy a magas munkanélküliséggel rendelkező régiók általában alacsonyabb termelékenységgel párosulnak (Köllő (2000b), Szabó (2006)), ami nehezíti a bérek területi kiegyenlítését. Fazekas (2005) Immarino–Santangelo (2000) tanulmányát idézi, mely egy mondatban foglalja össze a lényegét: „A hazai és a külföldi tulajdonú vállalatok termelésének földrajzi polarizációja ha egyszer már bekövetkezett, igen stabil, önfenntartó és visszafordíthatatlan folyamatokat eredményez, előrevetítve a centrális és perifériális helyzetű régiók tartós fennmaradását az országhatárokon belül.” (Fazekas (2005), p. 59).

Köllő (2000b) kimutatta, hogy a rendszerváltást követően a bérkülönbségek nőttek, majd később csökkentek mind a Budapest-város, mind a város-falu viszonylatban. Város-falu viszonylatban pedig meg is szűntek a 2000-es évekre. A régiók tekintetében a Budapest nélkül számított központi régió, a Nyugat- és a Dél-Dunántúl bérelőnye továbbra is fennmaradt a referenciának tekintett Észak-alföldi régióhoz képest. A mérés során a regionális különbségek mérséklődtek, amennyiben a munkanélküliséget és a termelékenységet is figyelembe vették. (Köllő (2003), Szabó (2006)) Kőrösi (2006) szerint a 90-es évek közepére a regionális munkanélküliség hatása beépült a bérekbe, azóta pedig viszonylag stabilan alakult, ami a bérkülönbségek további fennmaradásához vezetett.

Kínálati oldalról a munkavállalók ingázása a munkahely és az otthon kö-

zött, illetve a költözés mérsékelheti a területi különbségeket. Azonban ahogyan Kertesi (2000), Bartus (2003) és Fazekas (2005) is megállapították, Magyarországon magasak az ingázási költségek. A költözést pedig az olyan intézményi tényezők mint a magánbérlekások hiánya és piac szabályozatlansága nehezítik. Így amellet, hogy a költözés költséges, még az azzal járó kockázatokat is munkavállaló viseli. (Hegedüs (2003))

Köllő (2003) elemezte a városok és falvak közti bérkülönbségeket, amelyekkel kapcsolatban arra a következtetésre jutott, hogy a közsférában kisebbek, mint a vállalati szektorban. Az eltérés leginkább azzal magyarázható, hogy a kisvárosi és falusi intézmények bérkiadásai jellemzően kisebbek, mivel ezen a területeken jellemzőbbek a kisebb iskolák, egyszerűbb feladatokat ellátó egészségügyi intézmények és alapszintű közigazgatási létesítmények. Azonban pontosan ez a szerkezeti eltérés – hogy Budapesten nagyobb és komplexebb intézményi struktúrák vannak jelen mind az oktatás, mind a közigazgatás, mind az egészségügy területén – jelentős bérelőnyt jelent Budapestnek a többi településtípushoz képest. Az azonos településtípusok közötti régiós különbségek elhanyagolhatóak vagy nem szignifikánsak.

Kőrösi (2006) kiemeli, hogy a magyar munkaerőpiaci kutatások lényeges eredménye, hogy a külföldi tulajdonú vállalatok magasabb béreket fizetnek az alkalmazottaiknak, mint a hazaiak. Earle–Telegdy (2012) kimutatták, hogy ha a rendszerváltást követő időszakban – megközelítően 1998-ig – az adott vállalat előtte állami kézben volt, akkor a felvásárlás után a külföldi tulajdonba került vállalatnál a bérek emelkedtek. Ha egy belföldi vállalat vette meg az addig államinak számító céget, akkor ott a keresetek mérséklődtek azokhoz a vállalatokhoz képest, amelyek az állam tulajdonában maradtak. Amennyiben a szerzők kontrolláltak a vállalati fix hatásokra, vagyis éltek azzal a feltétellezzel, hogy a felvásárlás szempontjából a cégek különböznek, akkor a tulajdon szerinti bérkülönbségek ugyan mérséklődtek, azonban továbbra is fennmaradtak. Továbbá a szerzők kimutatták, hogy ha külföldi tulajdonba került egy cég, akkor bár minden dolgozói csoport bére emelkedett, azonban a magasan képzett fiatalok jártak a legjobban. Ezt az eredményt megerősítik Kertesi–Köllő (2001) és Köllő (2000c) kutatásai is.

Köllő (2000c) a tulajdonviszonyokkal kapcsolatban kiemelte, hogy a külföldi vállalatok általában nagyobb és jobban szervezett intézményeket takarnak,

szemben a magyar kisvállalkozásokkal. Továbbá látható, hogy az alacsonyabb bérű ágazatokban magasabb béreket fizetnek a külföldi vállalatok, mint a hazaiak, és a magasabb bérű ágazatok felé haladva egyre kisebb a külföldi és belföldi vállalatok közti bérkülönbség. Kőrösi (2006) felhívta a figyelmet arra, hogy a technológia szintje, a kialakult bérezési stratégia és a tulajdoni szerkezet időben stabil lehet az ágazatokban. Így az ágazat hatása erősebb, mint a külföldi tulajdoné. A saját kutatásában, ha a mintát nem bontotta meg ágazatok szerint, akkor több esetben is szignifikáns volt a külföldi tulajdon változója, ugyanis a tulajdon vette át a kihagyott ágazati változó hatását.

Kertesi–Köllő (2003) az ágazati bérkülönbségeket vizsgálta és kiemelte az ágazati koncentráció, valamint a vállalati szintű szervezettség hatását a bérekre. Az előbbi a monopoljáradék mértékét hivatott mutatni, az utóbbi pedig a munkavállalók alkuerejét méri. A szerzők arra a megállapításra jutottak, hogy ha az adott ágazat közepesen vagy erősen koncentrált, akkor a vállalaton belüli szervezettség bérekre gyakorolt hatása nagyobb. A szervezettség hatása erős koncentráltság mellett jelentősebb, mint közepes koncentrációnál. A koncentráció és a szervezettség közötti kapcsolat fordított esetben is megállja a helyét: minél nagyobb a szervezettség, annál inkább hat a koncentráció a bérekre. Emellett a szerzők kimutatták, hogy ha a szervezettség magas az adott ágazatban, akkor azoknál a cégeknél is magasabbak a fizetések, ahol nincsen béralku.

Rigó (2012) a vállalati és a több munkáltatóra kiterjedő szerződések hatását vizsgálta a bérekre. Magyarországra jellemző, hogy a vállalati szintű szerződések vannak túlsúlyban, az ágazati szerződésekhez képest, illetve inkább nagyobb vállalatoknál és bizonyos ágazatokban kötnek kollektív szerződéseket. Így a vállalatok eloszlása a kollektív szerződéssel lefedett és nem lefedett ágazatokban nem véletlenszerű. A kollektív szerződéssel rendelkező cégek általában az átlagnál jobb vállalatok¹, melyek nagyobbak és előnyös földrajzi régióban helyezkednek el. Emiatt elképzelhető, hogy ezek a vállalatok alapvetően magasabb béreket fizetnek. A szerző kiemeli a központi régiót a maga 60 százalékos lefedettségével. Összességében a magyar rendszer decentralizáltsága miatt a kollektív szerződések bérekre gyakorolt hatása kicsi.

¹Rigó (2012) a „jobb vállalatok” kifejezést azokra a cégekre használja, amelyeknél az OLS specifikációjában a vállalati fixhatások pozitívak. Tehát ezek a vállalatok a nem megfigyelt jellemzőikben tekinthetők jobbaknak.

Az ebben az alfejezetben bemutatott kutatásokban a bérfüggvényeket paraméteres módszerekkel – leginkább regressziókkal – becsülték, ahol általában alacsony volt az interakciós tagok száma, illetve csak bizonyos, a kutatás szempontjából lényeges interakciók jelentek meg. A függvényformának ezen korlátozását oldom fel a disszertációmban azzal, hogy tanuló algoritmusokat használok a bérfüggvények becsléséhez. Így bármilyen jellegű kapcsolat megjelenhet a modellekben.

2.2. Nemi bérkülönbségek Magyarországon

Nemzetközi összevetésben azt láthatjuk, hogy a férfiak és a nők közti bérkülönbségek időben csökkentek, azonban továbbra is fennállnak. (OECD (2018), Weichselbaumer–Winter-Ebmer (2005)) A nemi bérkülönbségek tekintetében pedig lényeges annak megértése, hogy milyen tényezők állnak a bérezési különbségek hátterében. A leggyakoribb gondolat az lehet, hogy diszkrimináció okozza a férfiak és a nők bérezésbeli különbségeit. Közgazdasági értelemben azonban akkor beszélünk diszkriminációról, ha a megkülönböztetés nem termelékenységi különbségből származik. Tehát ha két azonos termelékenyséű csoport valamilyen megfigyelhető jellemző alapján elkülöníthető és különböző elbánásban részesülnek, akkor az tekinthető diszkriminációnak. A nemi bérkülönbségek irodalmában a diszkrimináció az azonos jellemzőkkel rendelkező férfiak és nők kereseti különbségeivel ragadható meg. A munkagazdaságtani kutatásokban igyekeznek minél szélesebb információs bázist felhasználni, azonban így is előfordulhat, hogy nem minden, a bérezés szempontjából lényeges változót sikerül megfigyelni vagy helyesen mérni. (Lovász–Telegdy (2009))

A változók helytelen mérésének leggyakoribb példája a munkatapasztalat mérése. Általában a kutatásokban csak a potenciális munkatapasztalatot tudják mérni, ami az aktuális életkor és a legmagasabb iskolai végzettséghez tartozó életkor különbségeként adódik. Ez a mutató azonban nem veszi figyelembe a munkapiactól távol töltött éveket, például a nők esetében az otthoni, gyermekneveléssel töltött időt. (OECD (2018)) Emiatt a potenciális munkatapasztalat

a termelékenység túlbecsléséhez vezet.²

A termelékenység mérésénél a kihagyott változók problémája általában abból fakad, hogy az adatbázis készítésénél nem tudják mérni a férfiak és a nők közt fennálló hozzáállásbeli, értékrendbeli és nem kognitív képességbeli különbségeket. Így a megfigyelt jellemzők alapján az összehasonlított férfi és nő azonosnak tekinthető, mégis eltérően végezhetik a munkájukat, ami bérezési különbségekhez vezethet. Ez a mérhető különbség pedig utalhat munkaerőpiaci diszkrimináció jelenlétére. Emiatt általában külön kutatásokban – a bérezéstől függetlenül – vizsgálják a férfiak és a nők eltérő viselkedését, illetve tulajdonságaikat (Lovász–Telegdy (2009)), amelyek végső soron hathatnak a bérekre. A továbbiakban célom a férfiak és a nők eltérő bérezésének okait feltárni, azonban ezt csak a rendelkezésre álló adatok tükrében tudom megtenni és mikor nemi bérkülönbség létezéséről beszélek, akkor az nem egyenlő azzal, hogy diszkrimináló viselkedést tételezek fel a munkáltatók részéről.

A Magyarországra vonatkozó kutatások rámutattak, hogy a nemi bérkülönbségek a rendszerváltást követően csökkentek (Brainerd (2000), Newell–Reilly (2001)), amely csökkenés legnagyobb része a rendszerváltás utáni néhány évben zajlott le. (Galasi (2000), Varga (2018)) Köllő (2000a) Kertesi–Köllő (1996) tanulmányára hivatkozva három tényezőre vezette vissza a nemi bérkülönbségek alakulását a rendszerváltás időszakában: 1) a szellemi foglalkoztatásúak relatív kereseteinek növekedésére, 2) a nemi bérkülönbségek csökkenésére az alacsony bérezésű ágazatokban, valamint 3) a különbségek mérséklődésére a 35-55 éves korosztályban. Továbbá Kertesi–Köllő (1996) kimutatták, hogy a magasan képzett szellemi foglalkozásban dolgozók esetében a nemi bérkülönbségek nőttek. Emellett a magas munkanélküliségű régiókban az általános bércsökkenés jobban hatott a nőkre, mint a férfiakra.

Becker (1971) elmélete szerint a verseny növekedése csökkenti a diszkriminációt, mivel a megkülönböztetés a vállalatok nem hatékony viselkedéséhez vezet, így végső soron ezek a vállalatok vagy felhagynak a diszkrimináló viselkedéssel vagy kiszorulnak a piacról. Ezzel a verseny növekedését hangsúlyozó elmélettel magyarázza Lovász (2009) és Jolliffe–Campos (2005) a nemek közti bérkülönbség csökkenését a rendszerváltást követő társadalmi és munkaerőpiaci átrendeződés időszakában Magyarországon. Lovász (2009) számításai alapján

²További mérési problémákról lásd Reilly (2001) tanulmányát.

a nemi bérkülönbségek csökkenését csak kis mértékben magyarázta a verseny növekedése.

Csillag (2006) szerint a szocialista Magyarországon a férfiak és a nők foglalkozás szerinti megoszlásában jelentősebbek voltak a különbségek, mint a nyugati országokban. „A szocialista rendszerben hivatalos törekvés volt arra, hogy a nők a „számukra alkalmas” foglalkozásokban álljanak munkába. Mivel azonban a szellemi munka a fizikai munkához képest kevésbé értékesnek számított, s ez a sorrend a központosítottnak alakított bértarifarendszerben is megjelent, a tipikusan nőknek fenntartott állásokat alacsony kereset jellemezte. Vagyis a nemek közötti bérkülönbségek jórészt a foglalkozási szegregációban testet öltött diszkriminációnak volt betudható.” (Csillag (2006), p. 100)

Később, a 90-es évek közepétől a női és a férfi szakmák elkülönülése csökkent, a nők egyre inkább vállaltak munkát „férfi” szakmákban. Továbbá Csillag (2006) kimutatta, hogy a nemi bérkülönbségek Magyarországon leginkább az azonos munkakörben dolgozó férfiak és nők bére közötti eltérésekből ered, mintsem a munkakörök közti bérkülönbségekből.

Newell–Reilly (2001) és Lovász (2013) is kimutatta, hogy Magyarországon a bérelosztás felső részében magasabb a nők bérhátránya. Továbbá Lovász (2013) a köz- és a magánszféra összehasonlításával rámutatott, hogy ez a megállapítás mindkét szektorban igaz, azonban a közszférában mért bérhátrány minden kvantilisban alacsonyabb, mint a magánszférában. Ezzel összefüggésben Lovász (2013) igazolta, hogy a nők előléptetési esélyei nagyobbak a közszférában és itt vélhetően az erősebb szabályozottság révén a két nem tagjai egyenlőbb bánásmódban részesülnek, mint a magánszférában.

A nemek közti eltérő preferenciákat és társadalmi szokásokat mutatja, hogy a háztartási tevékenységek és a gyermekneveléssel kapcsolatos teendők általában a nőkre hárulnak, ami befolyásolja a nők munkaerőpiaci döntéseit, illetve a munkahelyi teljesítményüket. (Becker (1985)) A fizetett és nem fizetett munka közti megoszlással kapcsolatban Szabó-Morvai (2018) kimutatta, hogy Magyarországon a nők átlagosan 1-2 órával többet dolgoznak nem fizetett munkában, ami magába foglalja a gyermekneveléssel és a háztartási teendőkkel töltött időt. Ezzel szemben a férfiak a nőkhöz képest átlagosan egy órával több időt töltenek fizetett munkával. A téma kutatása kapcsán Galasi (2002a,b) is hasonló

eredményekre jutott.

Galasi (2002c) kimutatta, hogy a nőknél a gyerekek jelenléte jobban korlátozza a munkaerőpiaci részvételt, mint a férfiaknál. A szerző szerint ennek oka lehet önkéntes kínálati döntés kisebb – 0-3 éves – gyermekeknél vagy keresletoldali diszkrimináció idősebb – 7-14 éves – gyerekeknél. A gyermek jelenléte a nők esetében akadályozza a munkaerőpiaci részvételt, míg a férfiak esetében növeli azt. A munkaerőpiacon megjelenő gyerekes nők pedig jelentős bérhátránnyal szembesülnek, mellyel kapcsolatban Cukrowska-Lovász (2014) kimutatták, hogy a nyers bérkülönbség legnagyobb részét a gyerekes férfiak és nők közti bérkülönbség adja.

A preferenciákkal kapcsolatban Lovász (2013) azt találta, hogy a nők jobban szeretnek a közszférában dolgozni, mivel itt jobbak az előrejutási esélyek, a bértábla miatt kiszámíthatóbbak a bérek, illetve a kötött munkaidő kedvez a szülőknek. A közszféra alacsonyabb bérei miatt ez a munkaerőpiaci döntés megjelenik a nyers bérkülönbségben. Azonban az, hogy valaki a köz- vagy a magánszférában helyezkedik el, egy megfigyelhető munkaerőpiaci döntés és általában kontrollálnak rá a nemi bérkülönbséggel foglalkozó kutatásokban.

A nem kognitív készségek azok, amelyek felelnek a motiváltságért, az önbi-zalomért, a versenyszellemért vagy a másokkal szembeni bizalom meglétéért. Ezek a képességek pedig hatással vannak az iskolai és a munkahelyi teljesítményre, így végső soron befolyásolják a bérek alakulását is. Ezek a különbségek pedig már gyermekkorban is jelen vannak (Feingold (1994), Jacob (2002), Declercq-Varga (2018)) és idővel növelik a különbségeket a két nem között. (DiPrete-Jennings (2012)) Például a kutatások nagy része egyetért abban, hogy a férfiak versengőbbek és kockázatvállalóbbak, mint a nők. A nagyobb versenyt és kockázatvállalást igénylő munkakörök pedig általában magasabb béreket ígérnek. (Drucker et al. (2018)) A különbségek részben evolúciós, részben társadalmi és kulturális okokra vezethetők vissza. (Fazekas (2018)) A nemek közti eltérő kockázatvállalási kedv megjelenhet olyan munkaerőpiaci döntésekben, melyeket általában figyelembe tudnak venni a bérkülönbségek mérésénél, mint a foglalkozás vagy ágazati besorolás. Azonban a konkrét munkahelyi viselkedés általában nem megfigyelhető. A németországi bérkülönbségeket vizsgálva Brenzel-Laible (2016) bevonták a nem kognitív képességeket a kutatásukba és azt találták, hogy a nyers bérkülönbség megfigyelt jellemzők-

kel nem magyarázható része ugyan csökkent, azonban a különbség nagy részét továbbra sem sikerült magyarázniuk.

Lovász–Telegdy (2009) szerint a nem kognitív képességek közötti különbség lehet a termelékenységi és így a bérek közötti különbség oka is. Lovász–Rigó (2009) termelési függvény bevezetésével igyekeztek megragadni a férfiak és a nők közti termelékenységi különbségeket. A szerzők arra a következtetésre jutottak, hogy nincs termelékenységi különbség a nemek között Magyarországon, emiatt a kimutatott bérkülönbség sem magyarázható ezzel.

Összefoglalva elmondható, hogy Magyarországon az átlagos nemi bérkülönbség legnagyobb része nem a két nem közti eltérő foglalkozási, illetve ágazati eloszlással magyarázható, hanem az azonos munkaerőpiaci jellemzőkkel rendelkező férfiak és nők közti kereseti különbségből származik. A bérezési különbség okát az eddigi kutatásokban csak részben sikerült megmagyarázni.

2.3. Bértarifa adatbázis

A 2.1. és 2.2. alfejezetekben bemutatott kutatások jelentős részében a Nemzeti Munkaügyi Hivatal által gyűjtött Bértarifa-felvételből származó (továbbiakban Bértarifa) adatokat használták. A disszertációmban én is ezzel az adatbázissal dolgozom, emiatt ebben a fejezetben egyben mutatom be az adatbázist.³ A saját kutatási eredményeimet tartalmazó 4., 5. és 6. fejezetekben az adatbázis tárgyalásánál már csak azokra a módosításokra és megfontolásokra térek ki, amelyek eltérnek az alábbiaktól.

A Bértarifa 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 logaritmusáé. Az

³A Bértarifa adatok az adattulajdonos tulajdonát képezik. A használt adatokat a Közgazdaság- és Regionális Tudományi Kutatóközpont (KRTK) Adatbankja dolgozta fel, azokat Vincze János, DSc PhD hallgatójaként használom.

átlagos havi kereset magába foglalja az alapilletményt és az egyéb, havi szintre vetített juttatásokat is.

Az elemzéseknél el akartam kerülni a rövidebb munkaidőben foglalkoztatottak alacsonyabb keresetéből adódó torzítást. Emiatt az elemzésekben csak a teljes munkaidőben foglalkoztatottakat vizsgálom. A köz- és a magánszférára bérezési stratégiája is jelentősen különbözik Magyarországon, emiatt csak a magánszektorban foglalkoztatottak béreit 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 bér adatok nem megbízhatóak.

A 4. és az 5. fejezetekben kizárólag a 2016-os adatokat használom. A 6. fejezetben pedig a nemi bérkülönbségek időbeli alakulásának vizsgálatánál a 2008 és 2016 közti adatokra támaszkodom. A 4., 5. és 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 (adatbázisban szereplő név)	Mértékegység	Fejezet		
		4.	5.	6.
Nem (gender)	0: nő, 1: férfi	X		
Iskolai végzettség (educ)	9 kategória	X	X	X
Életkor (age)	Évek	X	X	X
Szolgálati idő (tenure)	Hónapok	X	X	X
Foglalkozás (isco_1)	Egyjegyű FEOR	X		
Foglalkozás (isco_2)	Kétjegyű FEOR		X	X
Külföldi részarány (foreignown)	4 kategória	X	X	X
Állami (önkormányzati) részarány (stateown)	4 kategória	X	X	X
Vállalati létszám (size)	Foglalkoztatottak száma	X	X	X
Régió (region)	NUTS 2 kategóriák	X	X	X
Településtípus (settlement)	1: Bp., 2: város, 3: egyéb	X	X	X
Ágazat (sector)	Egyjegyű TEÁOR	X	X	X
Vállalati kollektív szerződés (kol)	0: nincs, 1: van	X	X	X
Ágazati kollektív szerződés (kag)	0: nincs, 1: van	X	X	
Több munkáltatóra kiterjedő szerződés (ksz)	0: nincs, 1: van	X	X	

Forrás: Bértarifa

Az 1. táblázatban szereplő változóhoz kapcsolódó bővebb leírás az F.1. Függelékben található. Az 1. táblázatban a szolgálati idő az adott munkáltatónál eltöltött hónapokat takarja. A foglalkozási – egyjegyű FEOR szerinti – főcsoportokat tárgyalja a Központi Statisztikai Hivatal (2010) kiadvány, illetve az F.2. Függelék tartalmazza az egyes főcsoportokhoz tartozó elvárt iskolai végzettségeket. Az ágazatok közül a kevés számú megfigyelés miatt a közigazgatást, védelmet és kötelező társadalombiztosítást tartalmazó O ágazatot kizárom. A nemi bérkülönbségeket tárgyaló 5. és 6. fejezetben a „nem” változója a becslésekben nem szerepel, azonban eszerint a változó alapján különböztetem meg a férfiakat és a nőket a tanuló- és a tesztmintában. 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.

3. fejezet

Gépi tanuló algoritmusok

Athey–Imbens (2019) cikke számos aspektusból, összefoglalóan mutatja be a hagyományos ökonometriai módszerek és a tanuló algoritmusok közti különbségeket. A szerzők hangsúlyozzák, hogy a két módszertan eltérő célokat fogalmaz meg. A hagyományos ökonometriai módszerek lényege, hogy az adatok együttes eloszlásából egy becslést készítsenek. Ez a becslés leginkább egy paramétert jelent, amely információt szolgáltat a változók eloszlásáról. A leggyakrabban alkalmazott eljárás a regressziós elemzés, melyben a kimeneti változó és bemeneti változók közti a strukturális, illetve oksági összefüggések feltárása az elsődleges cél. Ezzel szemben a tanuló algoritmusok igyekeznek a lehető leghasznosabb függvényformákat alkalmazni és céljuk a lehető legpontosabb előrejelzések készítése. Az ökonometriai modellek a becslésből eredő torzítás minimalizálására törekednek. A tanuló algoritmusoknál azonban a cél a torzítás és a becslési variancia együttes minimalizálása. Emiatt ezek a módszerek számos esetben megengedik a torzítás növekedését az empirikus illeszkedés javításának érdekében. (Shmueli (2010))

A tanuló algoritmusok számos célra használhatóak. Például alkalmazhatóak újonnan létrejött nagy adatbázisokban létező komplex összefüggések feltárására. Az összefüggések megértése pedig további segítséget nyújthat új hipotézisek felállításában, illetve a meglévő elméletek kiegészítésében. Továbbá a tanuló algoritmusok kevésbé ragaszkodnak az elméleti megfontolásokhoz, így szolgálhatnak ellenőrzésként a már meglévő modellekhez, illetve tesztelhetik annak előrejelző képességét azáltal, hogy adnak egy viszonyítási alapot. Ez-

által pedig ösztönözhetik az alkalmazókat arra, hogy a már létező magyarázó modelleket további változókkal egészítsék ki, új mérési módszereket, illetve új elméleteket dolgozzanak ki. Tehát a kétféle megközelítést – bár eltérő célokat fogalmaznak meg – nem szükségszerű egymás ellentéteiként kezelni, hanem érdemesebb egymás kiegészítéseiként tekinteni rájuk.

A bérezési irodalomhoz kapcsolódóan a bérfüggvényeket leginkább regressziókkal becsülték. A becslések célja a bér és a vizsgálatba bevont személyes, vállalati és piaci változók közötti strukturális vagy oksági összefüggések vizsgálata volt. A bérfüggvény becslés egyre inkább a feltételes várhatóérték függvény minél jobb közelítését célozta, ami viszont további teret ad a tanuló algoritmusok használatának. A disszertációm 4. fejezetében tanuló algoritmussal becslem meg a magyarországi bérfüggvényt. Céлом a bérek és az azt magyarázó változók kapcsolatának vizsgálata, ahol a magyarázóváltozók körének megválasztásán kívül semmilyen feltevessel nem élek az adatgeneráló folyamattal kapcsolatban. A tanuló algoritmusok azon tulajdonságát használok ki, hogy képesek a bármilyen függvényformát alkalmazni és rátanulni az adatokban rejlő összefüggésekre. A fejezetben bemutatott munka tekinthető akár előzetes adatelemzésnek is.

Az 5. és a 6. fejezetben, a nemi bérkülönbségek vizsgálatánál, a gépi tanuló algoritmus használatát az előrejelzések pontosítása szempontjából tartom lényegesnek. A pontosabb előrejelzések készítése a Blinder-Oaxaca dekompozícióban segíthet a magyarázott és a nem magyarázott rész nagyságának pontosabb meghatározásában.

A 4. fejezetben a bérelőrejelzések készítéséhez döntési fákat használok. Számos más tanuló algoritmus közül esett a választásom a fákra. Gyakran használt tanuló algoritmus még a neurális hálók, a k -legközelebbi szomszéd (k -nearest neighbors, k -NN), illetve a támogatási vektoros gép (support vector machines, SVM) eljárások. Ezek az algoritmusok azonban „fekete dobozként” működnek, mivel a bemeneti és a kimeneti változók közötti kapcsolat rejtve marad. (Archer–Kimes (2008)) Ezzel szemben a döntési fák jól interpretálható eredményeket adnak, valamint a fa „végignézetésével” a változók közötti interakciók is feltárhatóak. Emellett a döntési fák esetében nem kell előzetes feltételezéseket tenni a változók közötti kapcsolatokról, nem úgy mint a regressziók esetében. Így elmozdulva a hagyományos ökonometriai eljárásoktól, előzetes feltevések nél-

kül, adatvezérelt módon becslem meg a bérfüggvényt, miközben kapcsolódva a hagyományos módszerekhez, a fák „végignézetésével” igyekszem a változók közti strukturális összefüggéseket elemezni.

A döntési fák esetében a kiindulópontot az N darab egyenként i -vel indexált megfigyelésekből álló adathalmaz jelenti. Ezeket az adatokat a gyökér jeleníti meg, amely a fa tetején látható. Ez az adathalmaz tartalmazza a magyarázóváltozókat és a függőváltozót. A fa minden lépésnél bináris vágásokat végez¹, vagyis a sokaságot pontosan két diszjunkt halmazra osztja, melyekhez tartozóan kiátlagolja az eredményváltozót. Ezek a csoportokhoz rendelt átlagok jelentik az abban a csoportban szereplő megfigyelésekhez tartozó előrejelzést. A kiinduló adathalmaz, a gyökér csomóponttá válik és meghatározódik az első vágási pont. A fa megjelenítésénél a gyökérből leágazik két csoport, úgynevezett levelek. A leveleken lévő csoportok a kezdeti adathalmazzal összehasonlítva homogénebb csoportokat jelentenek.

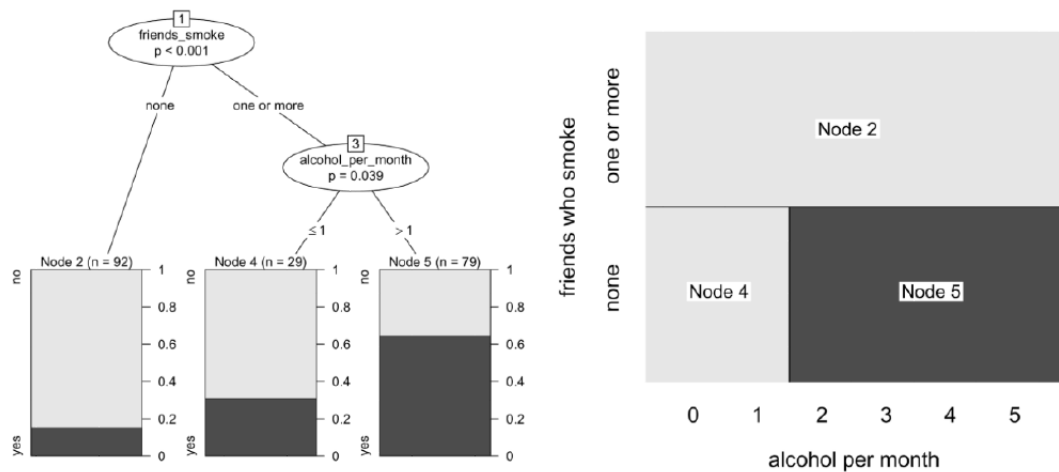
A következő vágás meghatározásához az eljárás a két levelet függetlenül kezeli és az egyik levelet további két csoportra osztja az egyik változó – úgynevezett vágóváltozó – szerint. Ezzel az érintett levél csomóponttá válik és két újabb levél keletkezik. A grafikus megjelenítésénél már három végpont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

A fákat és a változótér felosztását szemlélteti az 1. ábra, mely Strobl és szerzőtársai (2009) tanulmányából származik. A szerzők a dohányzási szokásokat vizsgálták. A kimeneti változó a dohányzást jelölő bináris változó volt, mely egy értéket vett fel, ha az illető dohányzott, egyébként nullát. A tanulmányban használt algoritmus a sokaságot először az alapján különböztette meg, hogy van-e olyan barátja a vizsgált személynek, aki dohányzik, majd az alapján, hogy a vizsgált személy hányszor fogyaszt alkoholt egy hónapban. A fa formában kapott eredmény látható az 1. ábra bal oldalán. Az ábra jobb oldala a változótér felosztását szemlélteti a felhasznált két magyarázóváltozó esetében, ahol a függőleges tengely az első, a vízszintes tengely pedig a második

¹Léteznek olyan eljárások, amelyek több részre is oszthatják a sokaságot, mint a CHAID, FACT vagy CRUISE algoritmusok. Ezekkel a módszerekkel azonban nem foglalkozom a disszertációmban, így a döntési fák működési elvét is bináris vágásokkal mutatom be.

magyarázóváltozót jelöli.

1. ábra. Döntési fa szemléltetése

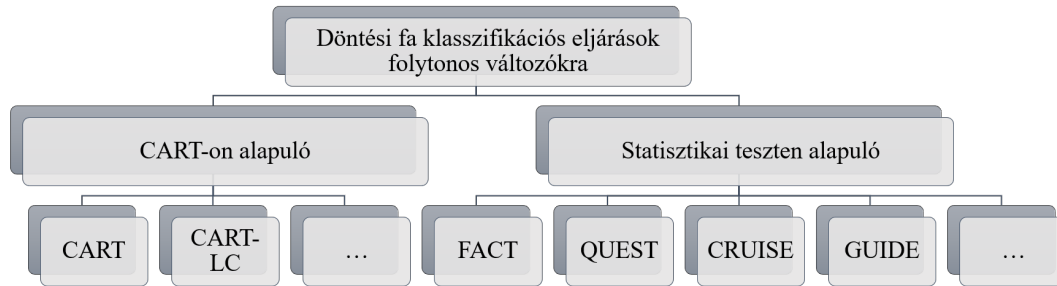


Forrás: Strobl et al. (2009), p. 30

Az 1. ábra alapján látható, hogy a fák célja a sokaság felosztása és nem az egyes változók hatásainak vizsgálata, ami jól példázza a hagyományos ökonometriai módszerek és a tanuló algoritmusok közti különbséget. Nem véletlen, hogy a fák használata elsősorban olyan tudományterületeken terjed el, ahol a változók között gyakoriak lehetnek a kereszthatások (Strobl et al. (2009), Levshina (2020)), illetve ahol az adatbázisban sok a magyarázóváltozó, azonban kevés a megfigyelés, mint például a genetikában (Strobl et al. (2009)) vagy a bioinformatika területén (Strobl et al. (2007)). Ezeken a területeken gyakran használják a döntési fákat és erdőket változószelekcióra. A kutatókat ugyanis elsősorban az érdekli, hogy mely változók függnek össze a kimeneti változóval és nem az egyes változók egyedi hatásaira koncentrálnak.

Számos különböző döntési fa eljárás létezik, melyek általános leírásáért és összehasonlításáért bővebben lásd Loh (2014) tanulmányát. Ezek az algoritmusok két nagy csoportba sorolhatók: a klasszifikációs és regressziós fán (Classification and Regression Tree), röviden CART-on alapulóba és a statisztikai teszten alapulóba. A megkülönböztetést szemlélteti a 2. ábra.

2. ábra. Döntési fa eljárások csoportosítása



Forrás: Jiao et al. (2020), p. 1 (saját fordítás)

A CART-on alapuló döntési fáknál az egyes vágásokat az alapján határozza meg az algoritmus, hogy melyik vágás eredményezi a legnagyobb a szórás² csökkenést. Emiatt tekinthetők ezek az eljárások mohónak. Ezzel szemben a statisztikai tesztek használó algoritmusoknál 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 vágási ponttal rendelkező változókat. A túlillesztés jelensége abból fakad, hogy az algoritmus könnyen rátanulhat az adatbázisban lévő sajátosságokra, ami ronthatja az általánosítási képességet. (James et al. (2013), Levshina (2020)) Hothorn és szerzőtársai (2006) szerint egy feltételes következtetési fa – mely a statisztikai tesztek használó eljárások közé tartozik – és egy optimálisan visszametszett CART teljesítőképessége 10 százalékos szignifikancia szinten megegyezik. Loh (2014), valamint Fernandez-Delgado és szerzőtársai (2014) statisztikai teszt nélkül kimutatták, hogy általában rosszabbul teljesít a következtetési fa, mint a CART. Összességében a torzítás-variancia együttes minimalizálásánál a következtetési fa esetében inkább a torzítás, a CART-nál pedig a variancia csökkentése kerül előtérbe. Az előbbiekben bemutatott problémák elkerülése érdekében a 4. fejezetben egy statisztikai teszten alapuló

²A szórás mérésére a GINI indexet vagy az átlagos négyzetes eltérést használják. (Jiao et al. (2020))

döntési fát használok.

A döntési fák hátránya, hogy magas az előrejelzési hibájuk. Emellett a kiinduló adathalmaz kis módosításával is nagyon különböző fákat kaphatunk eredményül: az eredmények robusztussága gyenge. 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))

Baggingnél a fák alapjául szolgáló adathalmaz bootstrap mintavétellel kerül meghatározásra. Ezek alapján a különböző mintákon más-más fák jönnek létre, az eljárás pedig ezeknek a különböző fának az eredményeit aggregálja. A boosting során a kiinduló adathalmazból az eljárás szekvenciális fákat határoz meg, ahol a következő fa mindig függ az előzőtől. (Breiman (1998), Breiman (2001), Hothorn–Lausen (2003))

A véletlen erdő – akárcsak a bagging – bootstrap mintákon dolgozik. Ezen a mintákon készíti el a döntési fákat. A vágási pontokban az algoritmus nem az összes változót, hanem csak azoknak egy részét tekinti potenciális vágóváltozónak, ami miatt az erdőben szereplő fák különbözőek lesznek. A bagging esetében, ha van egy az adatgeneráló folyamat szempontjából lényeges változó, akkor vélhetően ez a változó mindegyik fában megjelenik. Ezzel szemben a véletlen erdőnél, ha ez a lényeges változó kimarad a potenciális vágóváltozók közül, akkor az így elkészített fa különbözni fog a többitől, ami csökkenti a fák közti korrelációt. Minél kisebb a potenciális vágóváltozók száma, annál inkább csökken a korreláció, ami pedig összességében csökkenti az előrejelzések varianciáját és megbízhatóbb becslésekhez vezet. (Hastie et al. (2009))

A disszertációm 4. fejezetében döntési fát építék, hogy a bérek és a felhasznált magyarázóváltozók közti kapcsolatokat vizsgáljam. Az elérhető algoritmusok közül a feltételes következtetési fákat (Conditional 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 véletlen erdőkkel becslem meg. Emellett a 4. fejezetben a CTree eredményeinek robusztusságát az ezeket a fákat aggregáló feltételes véletlen erdő (Conditional Random Forest, CForest) eljárással ellenőrzöm.

Az 5. és a 6. fejezetben az Athey és szerzőtársai (2019) munkásságán alapuló általánosított véletlen erdők (Generalized Random Forest, GRF) közül a regressziós erdőket (Regression Forest) használok. A módszertani váltást a CForest és a GRF között az indokolta, hogy a CForest eljárással kapott becslőfüggvény ugyan torzítatlan, de nem konzisztens és az abból származó eredmények eloszlása sem aszimptotikusan normális. Az általánosított véletlen erdőkből származó eredmények azonban már bírnak ezekkel a kedvező statisztikai tulajdonságokkal.

A 3.1. alfejezetben a feltételes következtetési fákat (CTree), a 3.2. alfejezetben pedig feltételes véletlen erdőket (CForest) tárgyalom. A 3.4. alfejezetben az általánosított véletlen erdők (GRF) közül a regressziós erdők módszertanát mutatom be. A 4. fejezetben a Hothorn és szerzőtársai (2006) által kidolgozott partykit programcsomag ctree és cforest parancsait használok, melyek részletes leírásáért lásd Hothorn és szerzőtársai (2006, 2015) munkáit. Az általános véletlen erdővel készült becslések pedig a grf programcsomag regression_forest parancsával készülnek, melynek leírása szerepel Athey és szerzőtársai (2019) tanulmányában.

3.1. Feltételes következtetési fa – CTree

A CTree – mivel statisztikai tesztet használó döntési fa – 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 -értéket³. Amennyiben az algoritmus talál olyan magyarázóváltozót vagy magyarázóváltozókat, ahol az adott változó függetlenségére vonatkozó nullhipotézist el tudja vetni, akkor azok közül kiválasztja a

³A tesztstatisztika számolásának elméleti összefoglalásáért lásd Strasser–Weber (1999) munkáját, a CTreeben való alkalmazásáért pedig Hothorn és szerzőtársai (2006, 2008) tanulmányait.

legkisebb p -értékkel rendelkezőt. A legkisebb p -érték mutatja, hogy ennek a változónak a legszorosabb a kapcsolata az eredményváltozóval. Ezt követően 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 a következő kilépési kritériumok egyike nem teljesül:

- A CTree az összes magyarázóváltozó és a kimeneti változó együttes függetlenségére vonatkozó globális nullhipotézist el nem fogadja.
- Az algoritmus nem talál olyan magyarázóváltozót, amelynél a p -érték meghalad egy előre meghatározott α szintet. Ebben az esetben α értelmezhető úgy, mint a hipotézisvizsgálatoknál használt megbízhatósági szint.
- Legalább egy levél számossága egy előre meghatározott küszöbérték felett van. Amennyiben egy levélen található megfigyelések száma ennél a küszöbnél kisebb, akkor az algoritmus nem veszi figyelembe a változót mint lehetséges vágóváltozót.
- A vágás során keletkező levelek számossága egy előre meghatározott minimum küszöböt nem ér el. Amennyiben az egyik újonnan létrejövő levél nagysága a küszöbérték alá kerülne, akkor a létrejövő levél „túl kicsi” lenne, így a vizsgált levelet nem bontja tovább az algoritmus.
- A fa nagysága elér egy előre meghatározott mélységet. (Hothorn et al. (2006), Levshina (2020))

A CTree számos úgynevezett hiperparaméter előzetes meghatározását követeli meg. A legfontosabb paraméterek a partykit programcsomag `ctree_control` parancsánál a következők:

- Tesztstatisztika: a változó kiválasztásához használt statisztika fajtája.
- Vágási statisztika: a vágási pont kiválasztásához használt statisztika fajtája.
- Teszt típusa: a tesztstatisztika teszteléséhez használt eloszlás típusa.

- α : szignifikancia szint a változó kiválasztásához.
- Minimum kritérium: érték, melyet a tesztstatisztikának vagy az 1-p-értéknek meg kell haladnia, hogy létrejöjjön a vágás.
- Csomópont számosságának minimuma: minimum érték egy csomóponton. Amennyiben a csomóponton lévő megfigyelések száma ennél kevesebb, akkor az algoritmus nem bontja tovább. Ekkor ez a csomópont levélnek tekinthető.
- Levelek számosságának minimuma: minimum érték egy levélen. Amennyiben egy vágás után az újonnan keletkező valamelyik levélen lévő megfigyelések száma ennél kevesebb lenne, akkor az algoritmus nem hajtja végre ezt a vágást.
- Maximális szurrogátumok száma: a meghatározandó szurrogátumok (helyettesítők) maximális száma.
- Maximális mélység: a fa szintjeit meghatározó érték.⁴
(Hothorn et al. (2015), Hothorn et al. (2021))

3.2. Feltételes véletlen erdő – CForest

A CTree algoritmussal készített fákat összegző erdő a CForest. Ez az erdő a hagyományos CART-okra épülő véletlen erdőkkel ellentétben nem bootstrap mintavételt használ, hanem az adatok visszatevés nélküli almintájain épít CTree eljárással fákat.

Az előrejelzések készítése is eltér a CForest esetében a CART-okétól. A CART-ra épülő véletlen erdő eljárásoknál az előrejelzés ezeknek a levelekhez tartozó átlagoknak az átlaga. (Breiman (2001)) Azonban a CForest esetében az előrejelzések azoknak a megfigyeléseknek az átlagaiként adódnak, amelyekkel az adott megfigyelés minden fán egy levélre esett. (Levshina (2020))

⁴Ezt a hiperparamétert használom a 4. fejezetben a kis és közepes fa elkészítéséhez.

A CForest esetében hiperparaméterként meghatározhatóak a 3.1. fejezetben bemutatott változók egy fára vonatkozóan és a következő paraméterek az erdő tulajdonságait határozzák meg:

- Fák száma: az erdőben lévő döntési fák száma.
- Perturbáció: a fák alapjául szolgáló véletlen minta tulajdonságainak listája:

Visszatevés: visszatevéssel vagy visszatevés nélküli minta készüljön.

Arány: a teljes minta mekkora része szerepeljen a véletlen mintában.

- Potenciális vágóváltozók száma: a vágásoknál a véletlenül kiválasztott inputváltozók száma, melyek alapján vághat az algoritmus adott pontban. (Hothorn et al. (2021))

A 4. fejezetben használt CTree és CForest hiperparaméter-beállításait az ugyanebben a fejezetben szereplő 3. és a 4. táblázatok tartalmazzák.

3.3. Feltételes változófontossági érték

A 4. fejezetben a döntési fák elemzése mellett úgynevezett változófontossági mértékeket is vizsgálok. Ezek az értékek egy mennyiségi skálán fejezik ki, hogy az adott modellben az inputváltozók milyen mértékben hatnak az eredményváltozóra. (Inglis et al. (2022)) Ennek ismeretében könnyebben el lehet dönteni, hogy az eredményváltozó szempontjából melyek a lényeges és a kevésbé lényeges változók anélkül, hogy az egész fát részletesen meg kellene vizsgálni. Többféle módszer létezik a változófontosságok mérésére⁵, melyek közül én a disszertációmban is alkalmazott, Strobl és szerzőtársai (2008) által kidolgozott feltételes permutációra épülő fontosságokat tárgyalom.

⁵A különböző változófontossági mértékeket összefoglalóan bemutatja Strobl és szerzőtársai (2007) tanulmánya.

A permutációt használó változófontossági mértéknél az algoritmus egy adott változó értékeit keveri össze és az ebből számolt előrejelzéseket veti össze az eredeti modell előrejelzéseivel. Az ötlet a módszertan mögött, hogy a permutációval a kapcsolat megszűnik a permutált és a kimeneti változó között, így biztosan romlik a modell előrejelző képessége. Az eredeti modellből, illetve a permutációval kapott eredmények különbsége ragadja meg a változó fontosságát. (Strobl et al. (2008))

Formálisan megfogalmazva, legyen X_k a permutált változó, a fák sorszáma $t \in \{1, \dots, ntree\}$ és $\beta^{(t)}$ az out-of-bag (OOB) minta, amelyen az előrejelzések készülnek. Ekkor az X_k -ra számolt változófontosság a t . fán, $VI_{(k)}^{(t)}$, ami a következőképpen adható meg:

$$VI_{(k)}^{(t)} = \sum_{i \in \beta^{(t)}} \frac{\left(\hat{y}_{i(k)}^{(t)} - y_i\right)^2}{|\beta^{(t)}|} - \sum_{i \in \beta^{(t)}} \frac{\left(\hat{y}_i^{(t)} - y_i\right)^2}{|\beta^{(t)}|}, \quad (3.1)$$

ahol $\hat{y}_{i(k)}^{(t)}$ az i -edik megfigyelésre vonatkozó permutáció utáni, míg $\hat{y}_i^{(t)}$ a permutáció előtti predikció. y_i jelöli az i -edik megfigyeléshez tartozó eredményváltozó értékét. A véletlen erdőben a változófontosságot az egyes fákból számolt fontosságok számtani átlaga adja meg:

$$VI_{(k)} = \frac{\sum_{t=1}^{ntree} VI_{(k)}^{(t)}}{ntree} \quad (3.2)$$

A permutációval számolt változófontossági mértékek azonban érzékenyek arra, ha az inputváltozók korrelálnak. Amennyiben a permutált változó korrelál egy másik változóval, akkor a permutációval megváltozik a bemeneti és az eredményváltozók együttes eloszlása, ami megváltoztatja az előrejelzés pontosságát és megnöveli a változófontossági mértéket. (Strobl et al. (2008)) Nicodemus és szerzőtársai (2010) kimutatták, hogy ha a permutált változó független a kimeneti változótól, azonban korrelál a többi változóval, akkor a permutált változó fontossága pozitív értéket vesz fel.

Ennek a problémának a megoldására Strobl és szerzőtársai (2008) bevezetik a feltételes permutációs módszert, melyben felteszik, hogy a permutált változó (X_k) független a kimeneti változótól (Y) és a többi inputváltozó halmazától

$(Z_{(-k)})$. Formálisan ez a következőt jelenti:

$$X_k \perp Y \wedge X_k \perp Z_{(-k)} \quad (3.3)$$

Ezt a feltételt pedig úgy lehet teljesíteni, ha a permutációt a többi input-változó csoportjain belül hajtjuk végre, ezzel érintetlenül hagyva a korrelációs struktúrát X_k és a $Z_{(-k)}$ között. Ezt a permutációs módszert szemlélteti a 3. ábra.

3. ábra. Feltételes permutációs módszer illusztrációja

Y	X_k	$Z_{(-k)}$	Y	X_k	$Z_{(-k)}$
y_1	$x_{p(1)k}$	$z_{(-k)1}$	y_1	$x_{p(1)k}$	$z_{(-k)} = a$
y_2	$x_{p(2)k}$	$z_{(-k)2}$	y_3	$x_{p(3)k}$	$z_{(-k)} = a$
y_3	$x_{p(3)k}$	$z_{(-k)32}$	y_{27}	$x_{p(27)k}$	$z_{(-k)} = a$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{i-1}	$x_{p(i-1)k}$	$z_{(-k)i-1}$	y_6	$x_{p(6)k}$	$z_{(-k)} = b$
y_i	$x_{p(i)k}$	$z_{(-k)i}$	y_{14}	$x_{p(14)k}$	$z_{(-k)} = b$
y_{i+1}	$x_{p(i+1)k}$	$z_{(-k)i+1}$	y_{21}	$x_{p(21)k}$	$z_{(-k)} = b$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

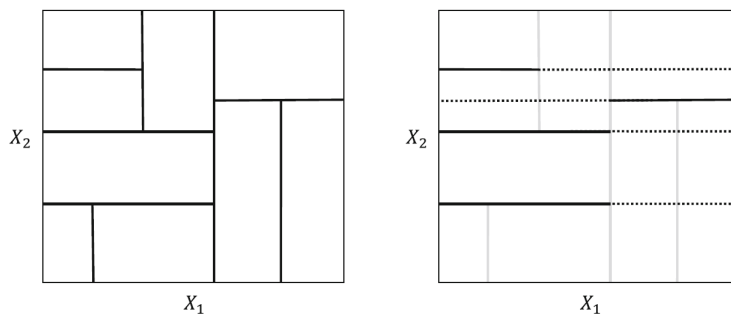
Forrás: Debeer–Strobl (2020), p. 6

A 3. ábra intuitívan mutatja a módszertan lényegét és viszonylag könnyen értelmezhető, ha a $Z_{(-k)}$ -ban kategória változók szerepelnek. Azonban $Z_{(-k)}$ bármilyen típusú változókat tartalmazhat, ami bonyolítja a kérdést, hogy adott struktúra mellett hogyan permutálhatóak X_k értékei.

Ennek a problémának a megoldására Strobl és szerzőtársai (2008) kidolgoztak egy módszertant, melyben asszociációs teszttel határozzák meg, hogy melyek azok a változók, amelyek statisztikai értelemben szignifikáns kapcsolatban vannak a permutált változóval, X_k -val. Ezekből a változókból jön létre $Z_{(-k)}$ egy részhalmaza, $Z_{(-k)}^{(s)}$. A változótér a $Z_{(-k)}^{(s)}$ -ban szereplő változók szerinti felosztása pedig az adott döntési fából származik. Amennyiben a $Z_{(-k)}^{(s)}$ -ban lévő változó szerinti vágás található a fában, akkor azt a két csoportot külön kezeli a módszertan, melyeken belül lehet X_k -t permutálni. A felosztást szemlélteti a

4. ábra, ahol a bal oldalon a döntési fa eredménye látható, míg a jobb oldalon a feltételes változófontosságokhoz használt változótér-felosztás szerepel. A 4. ábrán $X_k = X_1$ és $Z_{(-k)}^{(s)} = X_2$

4. ábra. Változótér-felosztás illusztrálása a feltételes változófontosságoknál



Forrás: Debeer–Strobl (2020), p. 8

A feltételes változófontosságok meghatározásánál az eljárás egy OOB mintán készíti el az előrejelzéseket az X_k eredeti és az $X_k|Z_{(-k)}^{(s)}$ permutált értékei mellett. Majd (3.1) és (3.2) alapján határozza meg a döntési fában és a véletlen erdőben a feltételes változófontosságokat. (Strobl et al. (2008), Debeer–Strobl (2020))

3.4. Általánosított véletlen erdő – regressziós erdő

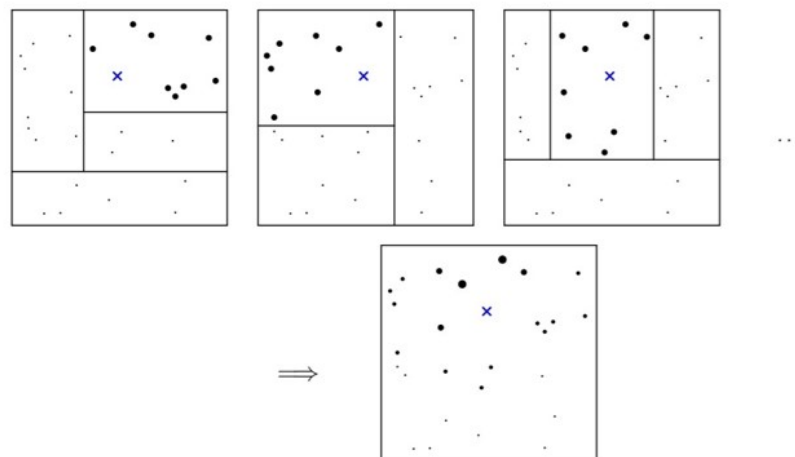
Az általánosított véletlen erdőt Athey és szerzőtársai (2019) alkották meg. Ez az algoritmus felhasználja a korábban tárgyalt véletlen erdő néhány tulajdonságát úgymint a fa építésénél az egymást követő bináris vágásokat, a mintavételt és a lehetséges vágóváltozók számának korlátozását a vágási pontokban. Azonban az algoritmus a fák építésénél még egy lényeges tulajdonságot, a „becsületesség” (honesty) feltételét is figyelembe veszi. Emellett a

fák eredményeinek aggregálása sem egyszerű átlagolással történik, hanem az eljárás súlyozza a megfigyeléseket.

Az általánosított véletlen erdők esetében a becslésség feltételezése úgy teljesül, hogy az algoritmus a tanulómintát két részre osztja és az egyik felén végzi a vágásokat, viszont a levelekhez tartozó becslések a másik mintarészből származnak. A regressziós erdőknél a minta visszatevés nélküli mintavételt jelent, szemben a CART-oknál használt bootstrap mintavétellel. A regressziós erdőnél minden megfigyelés belekerül valamelyik csoportba, így mindegyik megfigyelést felhasználja az algoritmus a fáknál. Wager–Athey (2018) megállapították, hogy ez a felbontása a tanulómintának nemcsak a torzítás csökkenését és a statisztikai mérőszámok használatát tette lehetővé, hanem – az átlagos négyzetes hiba mérséklésén keresztül – javította is az előrejelzési képességet.

Az általánosított véletlen erdők közül a regressziós erdőben az előrejelzések készítésénél azokat a megfigyeléseket átlagoljuk ki, amelyek az adott megfigyeléssel egy levélre esnek. Így egy megfigyelés, amely a változótérben közel áll a vizsgált megfigyeléshez, akár többször is beleszámolódhat az átlagba. Az átlagolási módszertan intuitív megértését segíti az 5. ábra.

5. ábra. Illusztráció a regressziós erdő fájnak aggregálására



Forrás: Athey et al. (2019), p. 1154

Az 5. ábrán az x-szel jelölt pontot vizsgáljuk. A felső sorban lévő panelek

az egyes fákat jelölik és az x körül szereplő nagyobb pontok azok, amelyek valamelyik fában egy levélre esnek a vizsgált megfigyeléssel. Ezeket a „pontokat” a regressziós erdő súlyozottan átlagolja ki: az x -hez tartozó súlyfüggvényben a hozzá közel eső pontok súlya nagyobb az átlagban. Ezeket a pontokat és az átlagban lévő súlyukat szemléltetik a más-más méretű pontok az alsó pane-
len. (Wager–Athey (2018)), Athey et al. (2019)) Az így kapott eredményeknél biztosított a konzisztencia és az aszimptotikusan normalitás feltételezése.⁶

Az általánosított véletlen erdők esetében, ahogyan a CForestnél is szükséges a hiperparaméterek megadása, melyek jelen esetben a következők:

- Fák száma: az erdőben lévő döntési fák száma.
- Potenciális vágóváltozók száma: a vágásoknál a véletlenül kiválasztott inputváltozók száma, melyek alapján vághat az algoritmus adott pontban.
- Minimum levélméret: a levelek méretének minimuma.
- Becsületesség: 0-1 skálán határozza meg, hogy az adatok hány százaléka kerüljön abba a csoportba, amelyen a fa építése történik.
- α : 0 és 1 közti paraméter, amely meghatározza, hogy a két létrejövő levélen legalább a megfigyelések α -szorosának kell lennie.
- Kiegyensúlyozatlansági büntetés: a büntetés mértéke, ha a két létrejövő levél különböző méretű.⁷ (Tibshirani et al. (2022a))

Regressziós erdőket használok az 5. és a 6. fejezetekben, amelyekben a hiperparaméter-beállításokat külön tárgyalom. Ezeknél az erdőknél a fák számán kívül az összes hiperparamétert optimalizálok. A paraméterek optimalizálása a következő képpen történik: az általánosított véletlen erdő meghatároz véletlenszerű paraméterkombinációkat. Mindegyik paraméterbeállítással az algoritmus készít egy erdőt és meghatározza az OOB hibáját az erdőnek. Ebben az esetben a fák száma az erdőben nem túl magas, ami miatt ez a hiba torzított

⁶Egyéb a konzisztenciával és az aszimptotikus normalitással kapcsolatos leírásért lásd Wager–Athey (2018) és Athey és szerzőtársai (2019) cikkeiben.

⁷A minimum levélméret egy abszolút korlátot jelent a metszésekénél, míg az α és a kiegyensúlyozatlansági büntetés a nagyon különböző méretű levelek létrejöttét akadályozza meg.

becslése a végleges erdő hibájának. A torzítás kiküszöbölését variancia dekompozícióval végzi el a GRF, majd az algoritmus ezekre a hibákra alkalmaz egy simítófüggvényt. Utolsó lépésben az algoritmus további paraméterkombinációkat határoz meg, melyek közül az lesz az optimális, ahol a simítófüggvénnyel számolt hiba a legkisebb.⁸ (Tibshirani et al. (2022b))

A szakirodalomban a kezelési hatások mérése egyre népszerűbb témája az empirikus kutatásoknak (Athey–Wager (2019), Dorie et al. (2019)), melynek vizsgálatára gyakran alkalmaznak oksági erdőket (causal forest) (Kreif et al. (2022)). Például munkaerőpiaci alkalmazásokért lásd Davis–Heller (2017), illetve Miller (2020) tanulmányait. Amennyiben a nemi hovatartozást tekintjük a kezelésnek, akkor oksági erdőkkel is vizsgálható lenne a nemi bérkülönbségek kérdése. Azonban ezek a típusú erdők megkövetelik azt, hogy a kezelt csoport tagjainak kiválasztása véletlenül történjen, így biztosítva az inputváltozók és a kezelés tényének függetlenségét.⁹ Ehhez azonban az kell, hogy a magyarázóváltozókat pontosan meg tudjuk figyelni és ne legyen olyan, amely lényeges a bérmeghatározás szempontjából, de kihagyjuk vagy rosszul mérjük. A Bértarifa adatbázisban azonban ez a feltételezés nem teljesül, mivel az adatbázisban csak a potenciális munkatapasztalat szerepel, ami a dolgozó életkora és az iskolai végzettségéhez tartozó életkor különbsége. Tehát a nők esetében éppen az a lényeges információ nem szerepel a munkatapasztalatnál, hogy mennyi időt töltöttek távol a munkaerőpiactól gyereknevelés miatt. Mivel a saját kutatásomban a nemek eloszlása nem tekinthető függetlennek a többi magyarázóváltozótól, így az oksági erdők nem alkalmazhatóak. Emiatt a férfiak és a nők bérfüggvényeit külön-külön becslem meg regressziós erdőkkel.

⁸A paraméterek optimalizálását a `grf` csomag `regression_forest` parancsán belül végzem el. A paraméterek optimalizálásánál hiperparaméter a fák száma, a kezdeti és a továbbiakban meghatározott paraméterkombinációk száma. Ezek alapbeállításai 50 fa, 100 kezdeti és 1000 további kombináció.

⁹A feltételezés angol megnevezése *unconfoundedness*, magyar fordítást nem találtam a kifejezésre.

3.5. 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. Emellett a szerzők egy olyan sokszoros ismételt mintavételen és mintafelbontáson (sample splitting) alapuló becslési módszert is javasolnak, amely lehetővé teszi az extrém csoportok átlagos karakterisztikáinak különbségeihez konfidenciaintervallumok készítését is.¹⁰ (Chernozhukov et al. (2018))

¹⁰A módszertan leírását és egy alkalmazást lásd Chernozhukov és szerzőtársai (2018) munkájában, O’Neill–Weeks (2018) egy másik alkalmazást mutat be.

4. fejezet

Bérfüggvény-becslés feltételes következtetési fákkal

4.1. Bevezetés

A bérezési kérdésekkel foglalkozó kutatások kiindulópontja Mincer (1958) tanulmányához köthető, amelyben a formális – iskolában eltöltött évekkal közelíthető –, az informális – munkahelyi körülmények közötti – tanulás és az életkor jövedelemre gyakorolt hatását vizsgálta. Később Mincer (1974) Becker (1964), valamint Becker–Chiswick (1966) munkásságára alapozva átdolgozta korábbi modelljét (Heckman et al. (2003), Heckman et al. (2005)) és bevette a munkatapasztalatot a magyarázóváltozók közé. A tapasztalat szerepeltetésével a humán tőke modell magyarázóereje nőtt, míg az iskolai végzettség együtthatója csökkent, azonban a Mincer (1958) korábbi tanulmányának eredményei továbbra is érvényesek maradtak.

A 70-es évektől Mincer humán tőke modellje az egyre több információt tartalmazó mikroadatbázisok létrejöttével és a számítási kapacitás növekedésével egyre fontosabb helyet foglalt el az irodalomban. (Chiswick (2003)) A bérezési kérdések további vizsgálatához a humán tőke modellbe újabb személyes jellemzőket vezettek be, mint a nem, a rassz, a házassági és veterán státusz. A személyes jellemzők mellett a későbbiekben az empirikus elemzésekben a magyarázó-

változók körét kiegészítették az ágazattal (Krueger–Summers (1986a,b)), a foglalkozással (Krueger–Summers (1986a), Wachtel–Betsey (1972)), a vállalatmérettel (Brown–Medoff (1989)), valamint a szakszervezeti tagsággal (Dickens–Katz (1986, 1987)). Blanchflower–Oswald (1989) és Blanchflower és szerzőtársai (1992) felhasználták a területi munkanélküliségi rátákat is a becsléseikben.

A szakirodalom másik része a bérmeghatározás kapcsán inkább módszertani oldalról közelített és az újabb változók bevezetése helyett inkább a becslési eljárások pontosításával foglalkoztak. Például instrumentális változóknak használták a szülők és a testvérek végzettségét, hogy a vizsgált személynél a kihagyott képesség változóból eredő torzítást kezeljék. (Ashenfelter–Krueger (1994), Miller et al. (1995))¹ A munkaerőpiacon megjelenő szelekciós torzítás kezelésére Heckman-féle szelekciót alkalmaztak a bérfüggvények felírásánál. (Gronau (1974), Lewis (1974), Heckman (1976, 1979)) Koenker–Bassett (1978, 1982) alapján Buchinsky (1994) és Chamberlain (1991) kvantilis regressziókkal becsülték a béreket. Falk–Leoni (2011) a szokásos magyarázóváltozókat felhasználva spline függvényekkel vizsgálták a területi munkanélküliség hatását a bérek alakulására és arra a következtetésre jutottak, hogy a spline függvénnyel szemben a hagyományos OLS becslés túlbecsüli a munkanélküliség hatását. Szintén spline függvénnyel vizsgálta Wolf (2002) a heti munkaidő órábérre gyakorolt hatását Németországban.

Mind az újabb változók bevonása a vizsgálatokba, mind az újabb statisztikai eljárások használata növelte ismereteinket arról, hogy milyen tényezők határozzák meg a béreket. Ezekben a kutatásokban a bérfüggvényeket lehet úgy értelmezni, mint a bérek feltételes várhatóérték függvényét, melynek becsléséhez olyan módszereket választanak, amelynél a torzítás a lehető legkisebb. Ezen kutatások célja – ahogyan a 2.1. alfejezet is mutatja – a bérek és az azt befolyásoló tényezők közti strukturális kapcsolat vizsgálata. A kutatási kérdések megválaszolásának lényeges eszköze az együtthatók szignifikanciáinak elemzése és a fennálló hipotézisek tesztelése.

Disszertációmban a bérfüggvényt más kritériumok alapján, más eszközökkel becslém meg. Ebben a fejezetben a bérek feltételes várhatóérték függvényét CTree eljárással közelítem. A fa használatával – mint általában a gépi tanu-

¹További az iskolai végzettséghez kötődő instrumentális változókat és az azokhoz kapcsolódó problémákat részletesen tárgyalja Card (1999).

ló algoritmusokkal – jobb illeszkedés érhető el az adatokon egy kevésbé kötött függvényforma alkalmazásával. Azonban itt a célom nem a legpontosabb előrejelzés készítése, hanem egy alternatív módszer használatával a bérekre vonatkozó feltételes várhatóérték függvény becslése és a változók közti kapcsolatok vizsgálata.

A 2016-os magyar adatokra illesztett fában egyrészt elemzem a kapott fa struktúrát, ami magába foglalja a vágások és a leveleken található alcsoportok kiértékelését. Másrészt megvizsgálom a változófontossági mértékeket, ami százalékosan mutatja, hogy adott változó a fa építése során mennyire lényeges. Harmadrészt a változók közötti kapcsolatok vizsgálatánál elemzem az egyes változók szurrogátum (helyettesítő) jellegét. Ez a szurrogátum tulajdonság abból adódik, hogy az algoritmus minden vágásnál meghatározza, hogy ha adott változó szerint nem tudna vágni, akkor mely más változóval kapnánk a változótér hasonló felbontását.² Ennek a helyettesíthetőségnek pedig fontos információtartalma van az magyarázóváltozók közötti kapcsolatokat illetően.

A 4. fejezet felépítése a következő. A 4.2. alfejezet tartalmaz néhány kiegészítést a Bértarifa adatokkal és a fa építésével kapcsolatban. Az eredményeket tartalmazó 4.3. alfejezetben három fa eredményeit tárgyalom. Ezek közül a legkisebb a legkönnyebben interpretálható, a legnagyobb pedig vélhetően a legpontosabb eredményeket adja. A különböző méretű fák, a 4.3.1., a 4.3.2. és a 4.3.3. alfejezetekben, vizsgálom a vágásokat, míg a 4.3.4. alfejezetben a változófontosságokat és a változók szurrogátum jellegét elemzem. A változófontosságok robusztusságát a 4.4. alfejezetben véletlen erdőkkel ellenőrzöm. A 4.5. alfejezetben összefoglalom a tanulságokat.

²A szurrogátumok meghatározásának módszertanát Hothorn és szerzőtársai (2006) munkája tárgyalja.

4.2. Adatbázis és módszertan

A magyarországi bérfüggvény becsléséhez a 2016-os Bértarifa-adatokat használok. A becslésnél figyelembe vett kategóriaváltozók közül néhánynál megfigyelhető a kategóriák természetes rendezettsége. Például az iskolai végzettségnél a megkülönböztetett 9 kategória egyre magasabb képesítéseket jelölnek. A külföldi és az állami részaránynál a 4-4 kategória a külföldi, illetve az állami részarány csökkenése alapján van rendezve. A változók ezen természetes rendezettségét azonban nem vettem figyelembe a fában. Ennek oka, hogy ha a rendezettség lényeges információt hordoz a várhatóérték függvény becslés szempontjából, akkor ennek látszania kell majd a fából.

A célváltozónak, a bruttó havi keresetek logaritmusának leíróstatisztikáját a 2. táblázat mutatja. A magyarázóváltozók leíróstatisztikái a F.3. Függelékben szerepelnek.

2. táblázat. Havi keresetek logaritmusának leíró statisztikája

Mutató	Érték
Átlag	12,43
Szórás	0,60
Minimum	10,67
Maximum	16,32
5. percentilis	11,72
95. percentilis	13,59
Megfigyelések száma	113001

Forrás: Bértarifa

A 2. táblázat alapján látható, hogy a bérek logaritmusának szórása kicsi. Kiugró értékek inkább az adatok felső részében jellemzőek, amit magyaráz a minimálbér létezése.

Három fát készítek, melyeken elemzem a változók közötti kapcsolatokat. A legnagyobb fa tartalmazza a legtöbb információt a bérek és a változók közötti kapcsolatot illetően, azonban több száz levélből áll, melynek egyenkénti

elemzése nehézkes. Emiatt a mélységi paraméter módosításával készítek két kisebb fát, melyeken a változók közötti kapcsolatok már vizuálisan is elemezhetők. A vágások vizsgálata mellett a feltételes változófontosságok elemzésével igyekszem átfogó képet adni arról, hogy mely változók a legmeghatározóbbak a bérek becslésénél.

A fejezetben bemutatásra kerülő fáknál a hiperparaméterek beállításait a 3. táblázat tartalmazza.

3. táblázat. Hiperparaméterek a különböző méretű fáknál

Hiperparaméter	Beállítás
Tesztstatisztika	χ^2 (quadratic)
Vágási statisztika	χ^2 (quadratic)
Teszt típusa	Bonferroni
α	0,05
Minimum kritérium	$1 - \alpha$
Csomópont számosságának minimuma	20
Levelek számosságának minimuma	7
Maximális szurrogátumok száma	optimális fában 0 vagy 5
Maximális mélység	eltérő

Forrás: saját számítás

A CTree a vágási pontokban a változók és a vágási pontok kiválasztásánál a próbastatisztikát χ^2 eloszlással veti össze. A ctree parancs a statisztikai teszteknel használatos $\alpha = 0,05$ -os szignifikancia szintnél Bonferroni korrekciót alkalmaz. A csomópontok és a levelek számosságának alapbeállítását nem változtattam meg. Szurrogátumokat csak a nagy fáknál vettem figyelembe. A szurrogátumok használata nem változtatja meg a fa szerkezetét, csak a változófontosságok meghatározásánál van szerepe. A fák méretét a maximális mélység határozza meg: a legkisebb fa 3, a közepes fa 5 mélységű. Az optimális fáknál nem korlátozom a mélységet.

A fa eredményeit jelentősen befolyásolhatja az adathalmazon történt kis változás is, emiatt a változófontossági mértékek robusztusságát véletlen erdőkkel ellenőrzöm. Mivel az erdő esetében a minta random módon kerül meghatározásra, így az eredmények becslésenként változhatnak. Emiatt a véletlen erdőből számolt változófontosságokat 100 erdő eredményeinek átlagaként hatá-

rozom meg. Az átlagolásnál a szórások elemzésével az eredmények stabilitása is ellenőrizhető. A 4. táblázat tartalmazza cforest parancs hiperparamétereinek beállításait.

4. táblázat. Hiperparaméterek az erdőnél

Hiperparaméter	Beállítás
Tesztstatisztika	χ^2 (quadratic)
Teszt típusa	Bonferroni
Fák száma	100
Perturbáció	nincs
Potenciális vágóváltozók száma	5

Forrás: saját számítás

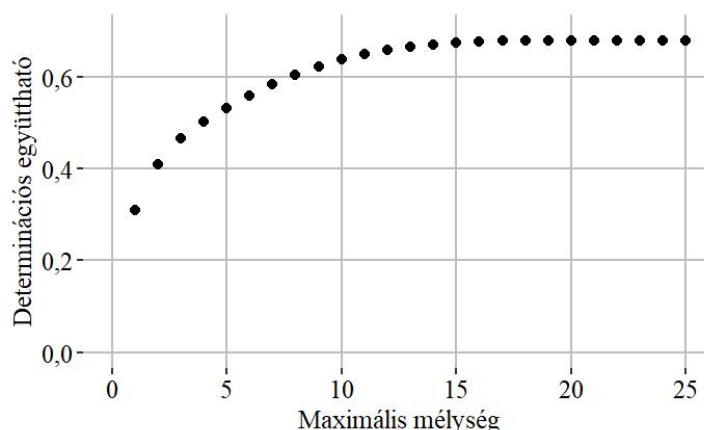
A CForest esetében a tesztstatisztika továbbra is χ^2 eloszlásból származik és Bonferroni korrekciót alkalmaz. Az erdőben szereplő 100 fa visszatevés nélküli mintákon készül³. Alapbeállításként az algoritmus a teljes adathalmaz 63,2 százalékát választja ki a mintába.

4.3. Eredmények

A 6. ábra mutatja a fa mélysége és a determinációs együttható közti kapcsolatot. Minél mélyebb a fa, annál több vágás található benne, ami növeli az összszórásból a fa által magyarázott rész nagyságát. A CTree statisztikai teszttel határozza meg, hogy van-e szignifikáns kapcsolat a eredményváltozó és az inputváltozók között. Amennyiben az algoritmus elfogadja a függetlenségre vonatkozó globális nullhipotézist, akkor a fa építése megáll, így minden fának van egy maximális nagysága. Jelen esetben a legnagyobb létrehozható fa 23 mélységű és 1288 vágással rendelkezik.

³Strobl és szerzőtársai (2007) kimutatták, hogy visszatevés nélküli minta esetében a változófontosságok megbízhatóbban becsülhetők, illetve így azok torzítatlansága is biztosított.

6. ábra. Determinációs együttható a CTree mélységének függvényében



Forrás: Bértarifa, saját számítás

A disszertációmban három fát elemzek, melyekhez tartozó statisztikai leírást az 5. táblázat tartalmazza.

5. táblázat. Különböző nagyságú fák leíróstatistikája

	Optimális fa	Közepes fa	Legkisebb fa
Mélység	23	5	3
Vágások száma	1288	31	7
Felhasznált magyarázóváltozók száma	14	9	3
Determinációs együttható	0,6788	0,5319	0,4654

Forrás: Bértarifa, saját számítás

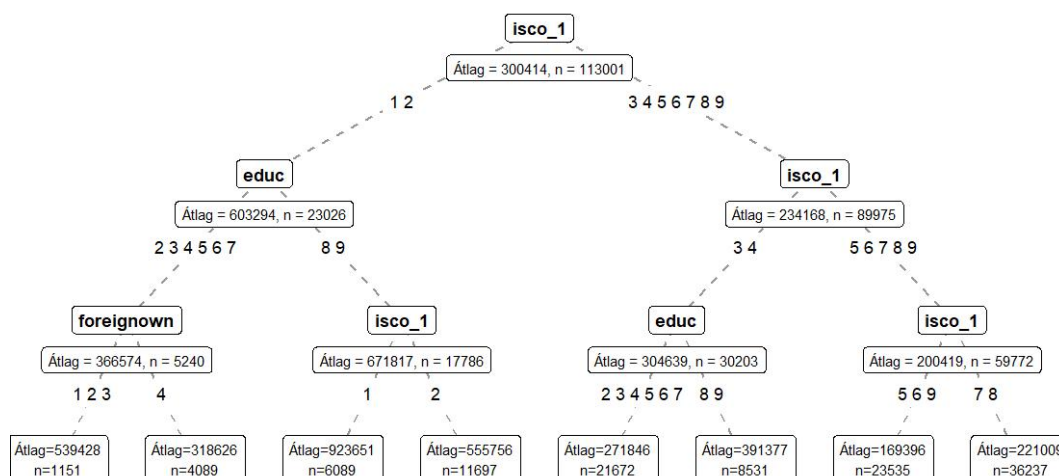
A kis fa bár korlátozottan tartalmaz vágásokat, mégis képes a logbérek szóródásának 46,54 százalékát magyarázni. A nagyobb fák már a logbérek nagyobb szóródását magyarázzák – ahogyan a 5. táblázatban is látható –, azonban az optimális fával sem érhető el 70 százalékos magyarázóerő.

A következő – 4.3.1.-4.3.3. – alfejezetekben a fákat egyenként tárgyalom, majd a 4.3.4. alfejezetben összegzem, hogy fontossági szempontból, mely változók meghatározóak a modellek alapján és elemzem a változók szurrogátum jellegét is. A 4.4. alfejezetben pedig a változófontossági mértékek robusztuságának ellenőrzésére szolgáló véletlen erdőket tárgyalom.

4.3.1. Legkisebb fa

A legkisebb fában szereplő vágóváltozók az iskolai végzettség, a foglalkozás és a külföldi részarány. A legkisebb fát megjelenítő 7. ábrán látszanak, hogy milyen csoportok különülnek el a fában. Az ábrán a félkövér feliratok jelölik, hogy adott levélen melyik változó szerint történt a vágás és a vonalakon a vágási kritériumok látszanak. Továbbá az ábrán az adott csomópontokhoz tartozóan feltüntettem a havi keresetek átlagát forintban és a megfigyelések számát.

7. ábra. Legkisebb fa



Megjegyzés: Az ábrán a forintba átszámolt összegek szerepelnek.

Forrás: Bértarifa, saját számítás

A legfelső doboz jeleníti meg a gyökeret, melyhez 113001 megfigyelés tartozik 300414 forintos átlagbérrel. Az első lépésben az algoritmus a foglalkozás – egyjegyű FEOR – szerint bontja ketté a mintát. A foglalkozás kódok meghatározásánál a FEOR-08 rendszerben az 1. és 2. főcsoportban foglalkoztatottaktól magas szintű önálló problémamegoldást igényelnek összetett gyakorlati és technikai feladatok megvalósításához, amihez főiskolai/egyetemi szintű kép-

zettség vagy érettségire épülő szakképzésben szerzett képesítés az elvart.⁴ A nagyobb számmal jelölt főcsoportokban elvégzendő feladatok egyre rutinszerűbbek és egyre alacsonyabb iskolai végzettséget várnak el a munkavállalóktól. (Központi Statisztikai Hivatal (2010)) Így az látszik már a legelső vágásnál, hogy a felelősségi körök meghatározóak a bérezés szempontjából, ugyanis ezen a levélen – a bal oldalon – nagyobb az átlagos bér, mint a magasabb FEOR-ral rendelkezőket tömörítő – jobb oldali – csomópontban. A jobb oldali levelet szintén a foglalkozás szerint bontja tovább a CTree és a 3. és 4. főcsoportban dolgozók kerül egy csoportba, míg az 5-9. főcsoportokban foglalkoztatottak a másik csoportba. Ennél a vágásnál is igaz, hogy az alacsonyabb főcsoportban dolgozók átlagosan magasabb bérrel rendelkeznek. Ezen az oldalon az 5-9. főcsoporthoz tartozókat az algoritmus a foglalkozás szerint további két csoportra osztja: az egyik csoportba kerülnek az 5., 6. és 9., a másikba a 7. és 8. főcsoportban dolgozók kerülnek. Az előbbi csoport átlagbére alacsonyabb, mint az utóbbié. Ez a felosztás már nemlinearitásra utal az adatokban.

A CTree az 1. és a 2. FEOR-ban dolgozókat az iskolai végzettség szerint osztja szét. Ekkor az algoritmus a felsőfokú – főiskolai/egyetemi – végzettséggel rendelkezőket különbözteti meg a legfeljebb érettségivel rendelkezőktől. Hasonló elkülönítés figyelhető meg a 3. és 4. FEOR-ban dolgozók esetében is. Mindkét vágásnál igaz, hogy a magasabb iskolai végzettség magasabb bérrel párosul.

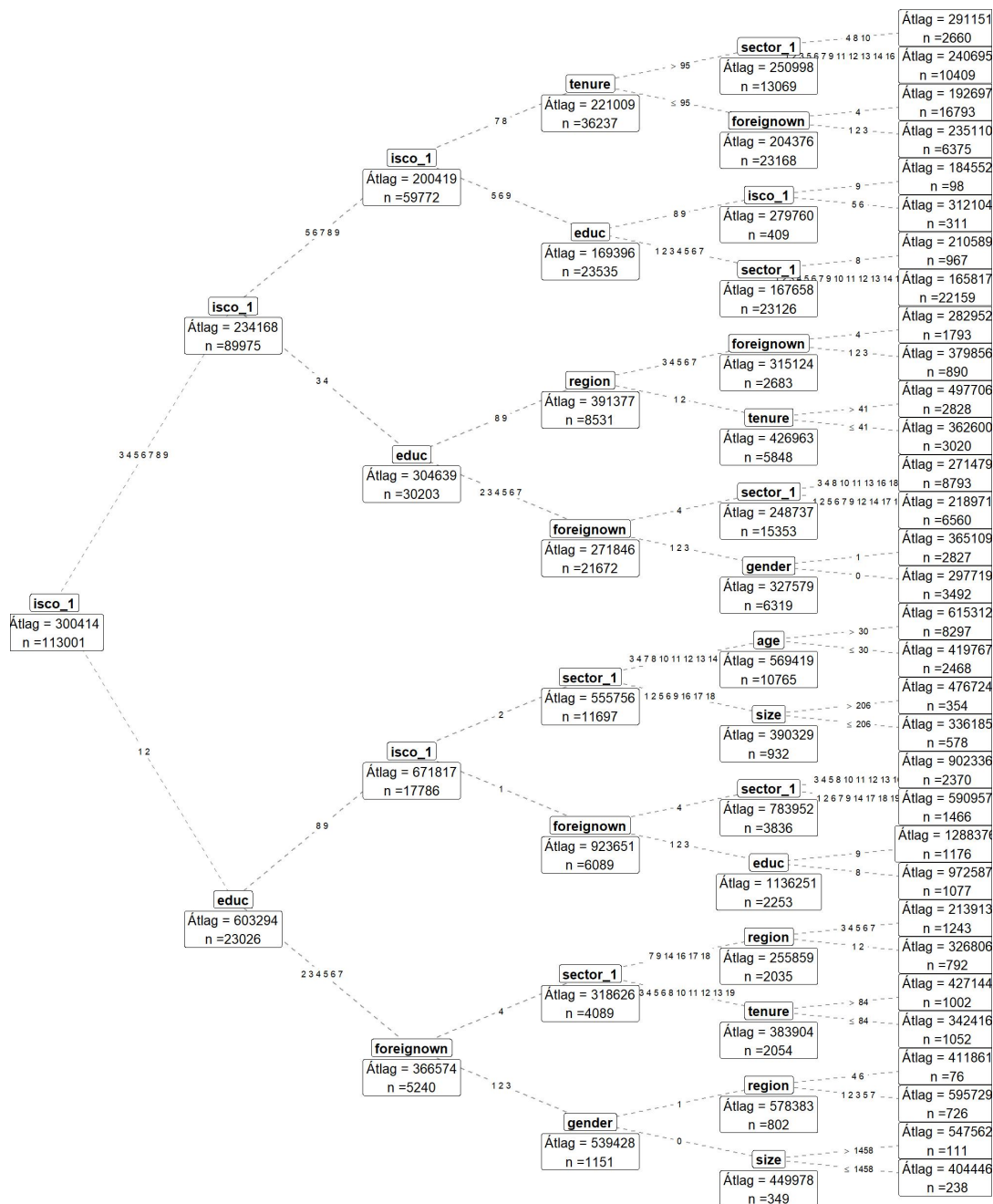
A külföldi tulajdon szerinti vágásnál az algoritmus elkülönítette a teljes mértékben hazai vállalatban dolgozókat (4. kategória) a többi tulajdoni formától. Ennél a vágásnál a hazai vállalatban dolgozókat tartalmazó levélen az átlagbér kisebb, mint a másik levélen.

4.3.2. Közepes fa

A közepes fában, a 8. ábrán, 9 változó szerint 31 vágás található.

⁴A foglalkozási – FEOR – főcsoportokat és a hozzájuk tartozó elvart iskolai végzettségeket összefoglaló táblázat az F.2. Függelékben található.

8. ábra. Közepes fa



Megjegyzés: Az ábrán a forintba átszámolt összegek szerepelnek.
A közepes fa táblázatos formában szerepel az F.4. Függelékben.

Forrás: Bértarifa, saját számítás

A kis fához képest a foglalkozás szerint még egy vágás jelenik meg a közepes fában, míg az végzettség szerint további kettő. Az új foglalkozás szerinti vágásnál továbbra is igaz, hogy a magasabb FEOR kód alacsonyabb bérrel párosul. Az iskolai végzettség esetében a további két vágásnál is megállja a helyét a korábbi megfigyelés, miszerint a magasabb végzettség nagyobb bért jelent. A külföldi tulajdon szerint a kis fához képest további 4 vágás jelenik meg a közepes fában és a CTree mindegyik esetben a – 4-essel jelölt – teljes mértékben hazai vállalatokat különíti el a többi tulajdoni formától. Ezeknél a vágásoknál az látható, hogy a hazai vállalatok átlagosan kisebb bért fizetnek, mint a bármilyen mértékben külföldi tulajdonú cégek.

Új vágóváltozó a közepes fában az ágazat. Eszerint a változó szerint hatszor vág az algoritmus. A 6. táblázat tartalmazza, hogy melyik ágazat hányszor tartozott a nagyobb, illetve a kisebb átlagbérrrel rendelkező csoporthoz.

6. táblázat. Ágazatok felosztása a közepes fában

Ágazat	Kisebb	Nagyobb
(A) Mezőgazdaság, erdőgazdálkodás, halászat	5	1
(B) Bányászat, kőfejtés	5	1
(C) Feldolgozóipar	2	4
(D) Villamosenergia-, gáz-, gőzellátás, légkondicionálás	1	5
(E) Vízellátás; szennyvíz gyűjtése, kezelése. . .	4	2
(F) Építőipar	5	1
(G) Kereskedelem, gépjárműjavítás	5	1
(H) Szállítás, raktározás	0	6
(I) Szálláshely-szolgáltatás, vendéglátás	6	0
(J) Információ, kommunikáció	1	5
(K) Pénzügyi, biztosítási tevékenység	2	4
(L) Ingatlanügyletek	3	3
(M) Szakmai, tudományos, műszaki tevékenység	2	4
(N) Adminisztratív és szolgáltatást támogató tevékenység	5	1
(P) Oktatás	4	2
(Q) Humán-egészségügyi, szociális ellátás	6	0
(R) Művészet, szórakoztatás, szabadidő	5	1
(S) Egyéb szolgáltatás	4	2

Forrás: Bértarifa, saját számítás

A 6. táblázat alapján az ágazatok között megállapítható, hogy a mindig rosszabbul fizető ágazat az I és a Q, míg az általában rosszabbul fizetők a A,

B, E, F, G, N, P, R, S ágazatok. Azonban nem látható monoton kapcsolat az ágazatok és a bér között: nem mondhatjuk, hogy a magasabb ágazati besorolás magasabb bért jelentene. Még a terméket és szolgáltatásokat előállító ágazatok közti megkülönböztetés sem határozza meg egyértelműen, hogy mely ágazatokban keresnek jobban a munkavállalók.

A régió szerinti vágásoknál a CTree a Közép-magyarországi és a Közép-dunántúli régiókat választotta el a többi, rosszabbul fizető földrajzi területtől. A szolgálati idő alapján a hosszabb vállalatnál töltött idő magasabb bérrel jár. Pozitív kapcsolat figyelhető meg a bér és a létszám esetében is. A fában a nemmel kapcsolatban megfigyelhető, hogy a férfiak jobban keresnek, mint a nők. A leírt összefüggések már sejtetik, hogy milyen kapcsolat lehet a magyarázóváltozók és az eredményváltozó között, azonban változónként kevés vágás található a fában, emiatt ezeket a feltárt összefüggéseket érdemes óvatosan kezelni.

4.3.3. Optimális fa

A CTree az optimális fában a foglalkozás szerint 71 alkalommal vág. Ahogyan a kicsi és a közepes fában, úgy a legnagyobb fában is látható a FEOR szerinti vágásoknál a foglalkozáskód és a bérek közti monoton kapcsolat, amit vélhetően az adott főcsoporthoz kapcsolódó felelősség és önálló problémamegoldási képességre való igényvel magyarázható. A CTree az optimális fában a mélyebben lévő vágásoknál általában egymás melletti kategóriákat hasonlít össze: az egyest a kettes, a hármast a négyes főcsoporttal, ami arra utal, hogy további jelentős különbségek lehetnek az egymás melletti főcsoportok esetében.

Az iskolai végzettség 95-ször vágóváltozó az optimális fában. A 7. táblázat tartalmazza, hogy az iskolai végzettségi kategóriák a kisebb vagy a nagyobb átlagbérral rendelkező csomópontokhoz, illetve levelekhez tartanak-e az optimális fában.

7. táblázat. Legmagasabb iskolai végzettség felosztása az optimális fában

Iskolai végzettség	Kisebb	Nagyobb
(1) Általános iskola 0-7 osztály	4	1
(2) Általános iskola 8 osztály	56	14
(3) Szakiskola	48	22
(4) Szakmunkásképző iskola	56	25
(5) Szakközépiskola	32	43
(6) Gimnázium	25	47
(7) Technikum	14	50
(8) Főiskola, Alapfokozat	21	31
(9) Egyetem, Mesterfokozat	4	28

Forrás: Bértarifa, saját számítás

A kis és közepes fában látszódó pozitív kapcsolat a végzettség és a bér között az optimális fában is látható. A 7. táblázat megerősíti, hogy a magasabb iskolai végzettség általában nagyobb bért jelent.

A régió szerint történik a második legtöbb, összesen 189 vágás az optimális fában. A régiókat és azt, hogy hányszor kerülnek a kisebb, illetve a nagyobb átlagbérrrel rendelkező csoportba a 8. táblázat szemlélteti.

8. táblázat. Régiók felosztása a közepes fában

Ágazat	Kisebb	Nagyobb
Közép-Magyarország	13	41
Közép-Dunántúl	23	30
Nyugat-Dunántúl	24	28
Dél-Dunántúl	38	16
Észak-Magyarország	37	16
Észak-Alföld	42	8
Dél-Alföld	36	18

Forrás: Bértarifa, saját számítás

A 8. táblázat rámutat, hogy Közép-Magyarország inkább a jobban fizető régiók közé tartozik, míg a Dél-Dunántúl, Észak-Magyarország, Észak- és Dél-Alföld a rosszabbul kereső régiók. A Közép- és Nyugat-Dunántúl esetében kiegyenlített, hogy merre tartanak.

A leggyakrabban használt vágóváltozó a nagy fában az ágazat, eszerint 210 esetben osztja fel a CTree a megfigyeléseket. A 9. táblázat mutatja, hogy hány esetben tart az adott ágazat a magasabb, illetve az alacsonyabb átlagbérrel rendelkező csomóponthoz, illetve levélhez.

9. táblázat. Ágazatok felosztása az optimális fában

Ágazat	Kisebb	Nagyobb
(A) Mezőgazdaság, erdőgazdálkodás, halászat	37	52
(B) Bányászat, kőfejtés	29	23
(C) Feldolgozóipar	79	79
(D) Villamosenergia-, gáz-, gőzellátás, légkondicionálás	69	24
(E) Vízellátás; szennyvíz gyűjtése, kezelése. . .	48	49
(F) Építőipar	55	48
(G) Kereskedelem, gépjárműjavítás	67	70
(H) Szállítás, raktározás	46	49
(I) Szálláshely-szolgáltatás, vendéglátás	16	76
(J) Információ, kommunikáció	61	47
(K) Pénzügyi, biztosítási tevékenység	42	38
(L) Ingatlanügyletek	41	41
(M) Szakmai, tudományos, műszaki tevékenység	67	36
(N) Adminisztratív és szolgáltatást támogató tevékenység	54	67
(O) Közigazgatás, védelem; kötelező társadalombiztosítás	0	0
(P) Oktatás	18	24
(Q) Humán-egészségügyi, szociális ellátás	14	39
(R) Művészet, szórakoztatás, szabadidő	27	34
(S) Egyéb szolgáltatás	31	53

Forrás: Bértarifa, saját számítás

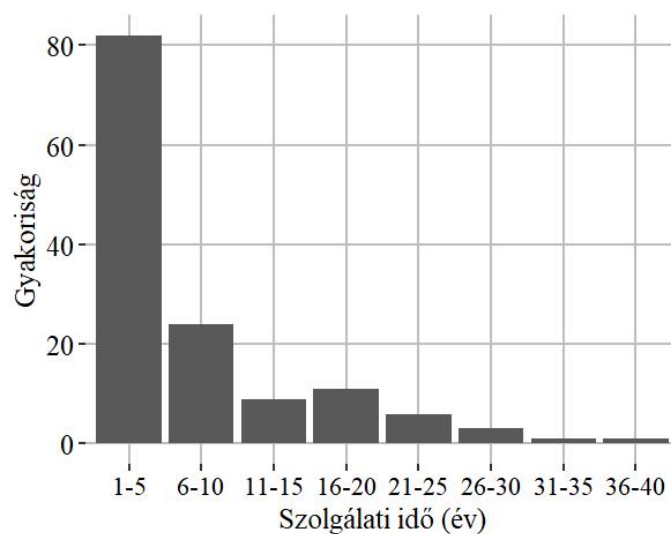
A 9. táblázatban összefoglalt eredmények arra mutatnak, hogy nincs olyan ágazat, amely minden helyzetben – minden vágási pontban – kevesebbet fizetne. Azonban többször kerül a kevesebbet fizető csomópontba az A, az I, a Q és az S ágak. Általában jobban fizető ágnak tekinthető a D, a J és az M. A többi ágazatnál a 9. táblázatban a rosszabbul, illetve a jobban fizető csoportokhoz kerülési értékek közelebb estek egymáshoz, így ezekben az esetekben egyéb tényezők is meghatározóak lehetnek a bérezés szempontjából.

A külföldi tulajdon szerint az algoritmus 51-szer vág és a vágásoknál általánosan igaz, hogy a nagyobb külföldi tulajdonnal rendelkező vállalatoknál általában magasabb béreket fizetnek. A közepes fában és az optimális fában

is a felsőbb szinteken – 9 mélységig – számos olyan vágás található a fában, amely a 4-es kategóriát, a nulla százalékos külföldi tulajdonú – vállalatokat különíti el a többitől. Ez a kategória szinte mindig a kisebb bért jelent a vágási pontokban.

A szolgálati idő változó mindegyik fában havi szinten mutatja az adott munkahelyen eltöltött időt. A 9. ábrán a könnyebb értelmezhetőség kedvéért években tüntettem fel a szolgálati időket.

9. ábra. Szolgálati idő szerinti vágások megoszlása az optimális fában



Forrás: Bértarifa, saját számítás

A legtöbb vágás 1 és 5 év között található. Ezek közül is a legtöbbször az első évben vág az algoritmus, vagyis megkülönböztetett helyzetben lehetnek a frissen érkezett munkavállalók azokhoz képest, akik régebb óta az adott munkahelyen dolgoznak. Emellett a vágások elemzésénél megállapítható, hogy a nagyobb szolgálati idő nagyobb bért jelent. Az életkornál mint vágóváltozónál általában a magasabb életkor nagyobb bérrel párosul. Azok a vágások, ahol a magasabb életkor kisebb bért jelent, inkább a fa alsóbb részein – 12. mélységtől – jelennek meg.

A vállalatméret alapján 163-szor vág az algoritmus. Ezekből a vágásokból 126 esetben a kisebb létszám kisebb bért is jelent. A legtöbb vágás a 20 és 49

fő közötti kategóriában történik, azonban érdemes megjegyezni, hogy ezekből a vállalatokból van a legtöbb a mintában.

A nem változónál a nők egy kivétellel mindig a rosszabbul fizető csoport-hoz tartoznak. A településtípus szerinti vágásoknál látható, hogy Budapesten átlagosan jobban fizetnek, mint a másik két településtípusban. A fa felső részein az algoritmus Budapestet különbözteti meg a másik két településtípustól. Emellett számos esetben kerül Budapest mellé az „egyéb” településtípus. Nem állapítható meg egyértelműen, hogy a városokban – Budapesten kívül – jobban fizetnének, mint az egyéb településtípusokon.

A kollektív szerződések közül a leggyakrabban használt változó a vállalati kollektív szerződés (kol), mely a vágások 80 százalékában nagyobb bért jelent. Az ágazati szintű (kag) és a több munkáltatóra kiterjedő (ksz) kollektív szerződéseknél is igaz, hogy a szerződés megléte általában nagyobb bérrel jár.

4.3.4. Feltételes változófontossági mértékek

A 10. táblázat az egyes megfigyelésekhez tartozó feltételes változófontossági mértékeket tartalmazza.⁵ Mivel a CTree permutációval határozza meg a változófontosságokat, így azok becslésenként eltérhetnek. Emiatt a 10. táblázatban 100 futtatás eredményét foglalom össze. Az algoritmus csak akkor tulajdonít jelentőséget egy változónak, ha aszerint vágott a fában. Például a legkisebb fában a 7 vágás 3 változó szerint történik, emiatt a 10. táblázatban csak 3 változó mellett szerepelnek értékek.

⁵A táblázatban szereplő változófontosságoknál nem veszem figyelembe a szurrogátum jellegét.

10. táblázat. Feltételes változófontossági mértékek (százalék) és vágások száma (darab) a különböző fánál

Változók	Optimális fa		Közepes fa		Legkisebb fa	
	Változó fontosság (%)	Vágások száma	Változó fontosság (%)	Vágások száma	Változó fontosság (%)	Vágások száma
Foglalkozás	43,63 (0,16)	71	68,27 (0,22)	5	80,87 (0,19)	4
Iskolai végzettség	9,62 (0,11)	95	13,09 (0,16)	4	17,36 (0,2)	2
Ágazat	9,37 (0,09)	210	4,59 (0,09)	6		
Külföldi részarány	9,36 (0,09)	51	7,33 (0,13)	5	1,77 (0,06)	1
Régió	6,68 (0,07)	189	1,44 (0,05)	3		
Szolgálati idő	4,48 (0,05)	137	3,11 (0,06)	3		
Vállalati létszám	4,21 (0,06)	163	0,25 (0,02)	2		
Nem	3,65 (0,06)	58	0,49 (0,03)	2		
Településtípus	3,01 (0,06)	99				
Vállalati szerződés	2,48 (0,04)	56				
Életkor	1,68 (0,04)	55	1,43 (0,04)	1		
Állami részarány	1,11 (0,03)	50				
Ágazati szerződés	0,59 (0,02)	42				
Több munkáltató sz.	0,11 (0,01)	12				

Megjegyzés: A változó fontosságoknál a 100 futtatás átlagai, illetve alattuk zárójelekben a szórások szerepelnek.

Forrás: Bértarifa, saját számítás

Az eredmények robusztusságával kapcsolatban lényeges megjegyezni, hogy a 10. táblázatban a szóródások igen alacsonyak, így a változófontossági mértékekre kapott eredmények robusztusnak tekinthetők.

A kis fában a legfontosabb változó a foglalkozás kód, amely a fa méretének növelésével egyre inkább veszít a jelentőségéből. A közepes és az optimális fa esetében a változófontossági mértékek közelítettek egymáshoz. Érdekes, hogy az ágazat és a külföldi részarány változók közül a közepes fában a külföldi

részaránynak van nagyobb jelentősége, addig az optimális fában a változófontosságok kiegyenlítődnek. Vélhetően a tulajdon és az ágazat hasonló mértékű információt hordoz a bérezéssel kapcsolatban. Az ágazatra és a külföldi részarányra számolt és tesztelt Cramer mutató alapján a két változó nem független, azonban a kapcsolat gyenge köztük.

Az algoritmus a kis fán többször vág a foglalkozás szerint és a változófontosság alapján ez a legfontosabb változó. Ennek a lehetséges magyarázata, hogy a FEOR kódok egyszerre tartalmaznak információt a munkakörben elvégzendő feladat összetettségéről és az elvárt iskolai végzettségről. Így az iskolai végzettség és a foglalkozási besorolás összefügg. Az iskolai végzettség a második legfontosabb változó, azonban a változófontossági mértéke jelentősen elmarad a foglalkozásától.

Már a közepes fában is látható, hogy a CTree inkább az adott munkahe-lyen eltöltött szolgálati időt és nem az életkort használja vágóváltozónak. Az optimális fában mind a szolgálati idő, mind az életkor százalékban mért fontos-sága megemelkedik, azonban a szolgálati időé nagyobb mértékben. A földrajzi változókat, a régiót és a településtípust tekintve, az előbbi a fontosabb. A vállalati szintű kollektív megállapodásoknak van a legnagyobb hatása a kollektív szerződések közül. Az ágazati és a több munkáltatóra kiterjedő kollektív szerződések a legkevésbé fontos tényezők a bérezés szempontjából. Összességében azonban a kollektív szerződések kevésbé lényegesek a bérmeghatározás szempontjából.

A szurrogátumokkal és az azok nélkül számolt feltételes változófontossági mértékeket tartalmazza a 11. táblázat. A CTree minden változóhoz legfeljebb 5 szurrogátumot rendel. Amikor ezeket a szurrogátumokat vizsgálom, akkor figyelembe veszem, hogy hány esetben lenne adott változó a valódi vágóváltozó helyettesítője, valamint hogy a kiválasztott 5 szurrogátum közül hanyadik helyen szerepel az adott változó. Az F.5. Függelékben részletes leírás szerepel az egyes vágóváltozókról és az azokhoz tartozó legjobb szurrogátumok sorrendjéről.

11. táblázat. A szurrogátumok nélkül és az azokkal számolt feltételes változófontossági mértékek (százalék) az optimális fában

Változók	Szurrogátumok nélkül	Szurrogátumok- kal
Foglalkozás	43,63	43,87
Iskolai végzettség	9,62	9,51
Ágazat	9,37	9,81
Külföldi részarány	9,36	9,05
Régió	6,68	6,70
Szolgálati idő	4,48	4,43
Vállalati létszám	4,21	4,19
Nem	3,65	3,58
Településtípus	3,01	2,95
Vállalati szerződés	2,48	2,48
Életkor	1,68	1,65
Állami részarány	1,11	1,09
Ágazati szerződés	0,59	0,59
Több munkáltató sz.	0,11	0,11

Forrás: Bértarifa, saját számítás

A szurrogátumokkal és az azok nélkül számolt feltételes változófontossági mértékek az ágazatot és a külföldi részarányt kivéve hasonlóan alakulnak. Az ágazat fontossága emelkedik, ha szurrogátumokat is használ a CTree, ami azzal magyarázható, hogy ez a leggyakrabban használt szurrogátum és minden változónak jó szurrogátuma⁶. Az ágazatnak jó szurrogátuma az állami részarány, a vállalatméret, a településtípus és a régió. A régió és a településtípus szurrogátum jellege azt mutatja, hogy az ágazatok területi megosztottsága nem egyenletes.

A második legtöbbet használt szurrogátum a vállalati létszám, amely az iskolai végzettségen és a külföldi részarányon kívül a többi változónak jó szurrogátuma. A létszám legjobb szurrogátuma az ágazat, a régió és az iskolai végzettség. Az iskolai végzettség mint szurrogátum a 7. szinttől jelenik meg az optimális fában, ami arra utal, hogy csak nagyon speciális csoportoknál jelenik meg a két változó közti kapcsolat.

⁶A szurrogátumokat akkor tekintem jónak, ha a CTree az esetek legalább 10 százalékban használja az adott változót szurrogátumként.

A harmadik legtöbbet használt szurrogátum a régió, amely nem jó szurrogátuma a foglalkozásnak, a szolgálati időnek, a nemnek és az életkornak. A régiónak az ágazat legjobb szurrogátuma, melyet a településtípus követ. A településtípusnál általában Budapest és az egyéb településtípus esik egy csoportba és ez legtöbbször a magasabb jövedelmet jelölő csoportosulás.

A foglalkozásnak erős szurrogátuma az iskolai végzettség, az ágazat, a vállalati létszám és az életkor. Az iskolai végzettség nem meglepő, mivel a foglalkozási szintek általában elvárnak bizonyos képesítéseket/iskolai végzettséget. A CTree az életkort inkább a fa mélyebb rétegeiben használja 3. és 4. szurrogátumként. Tehát a kapcsolat a két változó között csak speciális esetekben jelenik meg.

Az iskolai végzettség legjobb helyettesítői az ágazat, az életkor és a régió. A fa felsőbb szintjein – 4 mélységig – a CTree a felsőfokú végzettségeket jelölő 8. és 9. kategóriákat különíti el a többitől. Ebben a néhány vágásban vegyesen alakul, hogy melyik a legjobb szurrogátum. Érdekes, hogy a foglalkozási besorolás nem szerepel a legjobb szurrogátumok között. Tehát az iskolai végzettség feltehetően hat a foglalkozási besorolásra – ami a FEOR módszertanból következik –, azonban visszafelé nem igaz az összefüggés. A régió gyakran használt 1., 2. vagy 3. szurrogátum, ami arra enged következtetni, hogy különböző iskolai végzettséggel rendelkező munkavállalók területi eloszlása nem egyenletes.

A szolgálati idő és az életkor egymás legjobb szurrogátumai és mindkettőnek további jó szurrogátumai az iskolai végzettség, az ágazat és a vállalati létszám. A nem változónak jó szurrogátuma az ágazat, az iskolai végzettség és a vállalati létszám, melyek közül az első kettő szerepel inkább az 1. és a 2. helyen. A kapcsolat arra utal, hogy a végzettségi és ágazati megoszlása a férfiaknak és a nőknek nem egyenletes. Érdekes, hogy a foglalkozást az esetek 5,52 százalékában használja szurrogátumnak a CTree. Tehát a nemek ágazati megoszlottsága erősebb lehet, mint a foglalkozásbeli.

A településtípus legerősebb szurrogátuma a régió, majd következik az ágazat és a vállalati létszám. Amikor a régió a szurrogátum, akkor általában a CTree a Közép-magyarországi régiót különíti el a többitől. Azonban a településtípus szerinti vágásoknál nem feltétlenül igaz, hogy Budapest elkülönül a

többi várostól és az egyéb településtípusoktól. Általában Budapest valamelyik másik településtípussal együtt alkotja a magasabb bérű csoportot.

A külföldi részarány legjobb szurrogátuma az ágazat, melyet az állami részarány követ. Azonban a külföldi tulajdon egy másik változónak sem jó helyettesítője. Az állami részarány legjobb szurrogátumai az ágazat, a vállalati létszám, a régió, valamint az ágazati kollektív szerződések. Azoknál a vágásoknál, ahol az ágazati szerződés bekerült a szurrogátumok közé, ott a teljes mértékben állami tulajdonban lévő vállalatok (1. kategória) általában rendelkeznek ágazati szerződéssel és a jobban fizetettek közé tartoznak. A kollektív szerződéseket tovább vizsgálva látható, hogy a vállalati kollektív szerződésnek jó szurrogátuma az ágazati kollektív megállapodás, azonban fordítva nem így van.

4.4. Robusztusságvizsgálat

A robusztusságvizsgálat során a CTree alapján számolt változófontosságok érzékenységét vizsgáltam CForest eljárással. A CForesttel számolt változófontossági mérték nem veszi figyelembe egy változó szurrogátum jellegét. Emiatt az összehasonlításnál, a 12. táblázatban a CTree esetében a szurrogátumok nélkül számolt, illetve a CForesttel kapott változófontossági mértékek százra normalizált értékeit tüntettem fel. A CForest esetében a véletlen szerepet játszik mind a fákhhoz használt minták, mind a potenciális vágóváltozók kiválasztásában, így a 12. táblázatban 100 darab véletlen erdővel számolt változófontosság átlagát tüntettem fel. A 100 erdővel számolt változófontosságok leíróstatisztikáit az F.6. Függelék tartalmazza.

12. táblázat. Változófontossági mértékek a CTree-nél és a CForest-nél (százalék), illetve azok különbsége (százalékpont)

Változók	CTree	CForest	Különbség
Foglalkozás	43,63	36,89	6,74
Iskolai végzettség	9,62	19,46	-9,84
Ágazat	9,37	8,60	0,77
Külföldi részarány	9,36	9,37	-0,01
Régió	6,68	6,08	0,60
Szolgálati idő	4,48	3,80	0,68
Vállalati létszám	4,21	3,58	0,63
Nem	3,65	3,71	-0,06
Településtípus	3,01	2,21	0,8
Vállalati kollektív szerződés	2,48	2,47	0,01
Életkor	1,68	2,09	-0,41
Állami részarány	1,11	1,03	0,08
Ágazati kollektív szerződés	0,59	0,65	-0,06
Több munkáltatóra kiterjedő szerződés	0,11	0,06	0,05

Megjegyzés: A különbség a CTree-hez képesti eltérést mutatja.

Forrás: Bértarifa, saját számítás

A 12. táblázat alapján a CForest eljárásból származó változófontossági sorrend nagyságrendileg megegyezik a CTree algoritmussal készítettével. Azonban a százalékban mért változófontosságnál a foglalkozás hatása jelentősen csökken, míg az iskolai végzettségé megnő. A foglalkozás hatásának csökkenését az magyarázza, hogy a véletlen erdők csak a változók egy részét veszik figyelembe az adott vágási pontban. Emiatt ha ez a változó nem szerepel a potenciális vágóváltozók között, akkor annak szerepét egy másik változó veszi át. Emiatt az adott változó fontossága nő, míg a foglalkozásé csökken. A CTree és a CForest alapján számolt változófontosságok esetében a nagy fontossággal bíró változóknál a fontosság csökkenése az erdőben egy általánosan megfigyelhető jelenség. (Levshina (2020))

4.5. Összefoglalás

A CTree magyar béradatakon történő alkalmazásával egyrészt lehetőség nyílt a változók közti kapcsolatok vizsgálatára a fák elemzésével. Másrészt a változófontossági táblázat összefoglalóan tartalmazott információt az egyes változók bérezésben betöltött szerepéről. A szurrogátumok vizsgálata pedig a változók közti kapcsolatokról nyújtott információkat.

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 is csökkennek. Tehát monoton kapcsolat látszik a FEOR kódok és a bérek között. A következő 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. Továbbá érdekes eredménye a CTree használatának, hogy nem elsősorban a humán tőkét meghatározó iskolai végzettség a keresetek szempontjából legfontosabb változó, hanem a foglalkozás. A háttérben az állhat, hogy a foglalkozási kategóriákhoz tartozik egy elvárt végzettségi szint, emiatt a két változó erősen összefügg, viszont a FEOR kódolás több információt hordoz a felelősségi körök megosztása kapcsán. Például a 4-8 FEOR kódokhoz azonos végzettségi szinteket várnak el, azonban a felelősségi körök különböznek.

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 miszerint 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 a magasabb béreket fizetnek, mint a nulla százalék külföldi részesedéssel rendelkezők, ami alátámasztja Kőrösi (2006) azonos megál-

lapítását. Az állami részarány hatása sokkal kisebb, mint a külföldi tulajdoné, azonban meg kell jegyezni, hogy a két változó definíció szerint összefügg, ami a szurrogátumok vizsgálatánál látszik is: a külföldi az állami részarány jó szurrogátuma. A külföldi részarány természetes sorrendje „kijön” a fából, azonban az állami részarányé nem.

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 meghatározóbb, szemben az általános munkatapasztalattal, melyet ebben a kutatásban 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 inkább 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. Azonban 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 létszám és a bérek között sejthető a nem monoton kapcsolat, azonban a kapcsolat jellege nem határozható meg egyértelműen. A nem szerinti megkülönböztetés kimutatható a CTree alapján, mivel egy kivétellel a férfiak mindig a jobban fizető csoportot jelentették. Továbbá igazolva Csillag (2006) eredményeit a férfiak és a nők foglalkozásbeli megosztottsága egyre kevésbé van jelen a magyar munkaerőpiacon, azonban látszik a két nem ágazati megosztottsága.

Rigó (2012) felhívta a figyelmet arra, 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. Emellett Rigó (2012) megállapította, hogy a kollektív szerződések bérekre gyakorolt hatása alacsony, amit a CTree alapján számolt változófontossági táblázat alátámasztott. Így Rigó (2012) mindkét megállapítását a CTree tudta igazolni.

Összességében a CTree használata igazolt néhány kutatási eredményt és további kérdéseket vetett fel ott, ahol nem. Az életkorral kapcsolatos eredmények arra utalnak, hogy a hagyományos regressziós becsléseknél érdemes lehet a továbbiakban a potenciális munkatapasztalatot használni a bérbecslésekben, még akkor is, ha ez sem méri pontosan a ténylegesen a munkaerőpiacon töltött időt. A létszámnál megfontolandó a magasabb rendű tagok bevezetése, ahogyan a külföldi tulajdon esetében az ágazattal képzett keresztszorzat alkalmazása is.

5. fejezet

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

5.1. Bevezetés

A nemi bérkülönbségek folyamatos mérséklődése figyelhető meg a világ számos táján az utolsó néhány évtizedben, azonban a meglévő különbségek mára sem tűntek el teljesen. (OECD (2018), Weichselbaumer–Winter-Ebmer (2005)) Az OECD (2018) tanulmánya szerint 39 százalékos volt a nemi bérkülönbség a férfiak javára 2015-ben. Ezen az átlagos értéken belül, az országok között nagy különbségek figyelhetők meg. Japánban, Koreában, Mexikóban és Chilében mutatták ki a legerőteljesebb különbséget a férfiak és a nők bére között. Az OECD (2018) szerint Európában a német anyanyelvű országokban, a mediterrán térségben, Csehországban és Hollandiában volt a legnagyobb a nemi bérkülönbség. A legkisebb – 30 százaléknál kisebb – bérkülönbség a

Kelet-Európában, a skandináv térségben és Portugáliában volt tapasztalható.

Ezek a számítások átlagosan mutatják a különbség mértékét, azonban az eltérés okát nem magyarázzák. A bérek közötti különbség magyarázható lehet a férfiak és a nők megfigyelt munkaerőpiaci jellemzőinek eltéréseivel, termelékenységi különbséggel, illetve nemi diszkriminációval. Az eltérés okainak megértése lényeges a probléma kezelésének szempontjából, ugyanis a nemi bérkülönbségek csökkentése célzott politikai intézkedések tárgya az Európai Unióban és Magyarországon is. (Európai Bizottság (2015))

A nemi bérkülönbségekkel foglalkozó szakirodalomban leggyakrabban a Blinder-Oaxaca dekompozíciót használják két csoport átlagbérei közti különbség vizsgálatára. A dekompozíció felbontja az átlagok különbségét egy – a megfigyelt jellemzőkkel – magyarázott és egy – a megfigyelt jellemzőkkel – nem magyarázott részre. A felbontáshoz szükség van egy referencia modellre, valamint a két nemre vonatkozó bérfüggvény becslésekre. Gyakran a férfiakat tekintik a referencia csoportnak és így csak a férfiakra és a nőkre vonatkozó bérfüggvények meghatározására van szükség.¹ Ezeket a bérfüggvényeket Blinder (1973) és Oaxaca (1973) valamint későbbi kutatásokban is általában regressziókkal becsülték.

A dekompozícióhoz kapcsolódóan számos általánosítást végeztek, melyek – azáltal, hogy még inkább szem előtt tartották a munkaerőpiac jellemzőit, – alkalmasabb eszközzé tették a Blinder-Oaxaca felbontást a bérkülönbségek vizsgálatára.² A gondolat, hogy a béreloszlás különböző részein más-más mértékben hatnak a megfigyelt jellemzők, az egyéni bérfüggvények becslésénél is már megjelent. (Buchinsky (1994), Chamberlain (1991)) Ennek nyomán Mata–Machado (2005) vezették be a kvantilis regressziók alkalmazását a Blinder-Oaxaca felbontásnál. Reimers (1983), Neuman–Oaxaca (2004) Heckman (1979)-féle szelekciót használtak a munkaerőpiaci részvétellel indokolható torzítások kezelésére a felbontásban használt bérfüggvényekben. Sinning és szerzőtársai (2008) a dekompozíció nemlineáris kiterjesztéseit készítették el.

¹A Blinder-Oaxaca felbontáshoz használt függvényeket részletesen tárgyalja Jann (2008).

²Az általam bemutatásra kerülő általánosítások főként a bérkülönbségek elemzésénél jelennek meg. A Blinder-Oaxaca dekompozíció további általánosításairól pedig lásd Fortin és szerzőtársai (2011) tanulmányát.

Ezek az általánosítások azonban továbbra is meglehetősen kötött függvényformákkal dolgoztak. Azonban számos Magyarországra vonatkozó – a 2. fejezetben tárgyalt – kutatási eredmény arra enged következtetni, hogy a bérfüggvények becsléséhez használt magyarázóváltozók feltehetően nem függetlenek, az interakciós tagok pedig csak korlátozottan jelennek meg a becslésekben. Példaként említhető Fazekas (2005) vagy Kőrösi (2006) eredményei. Fazekas (2005) szerint a betelepülő külföldi vállalatok elhelyezkedésére jelentősen hatott a nyugati határtól való távolság a rendszerváltás után. Ezzel szemben a magyar vállalatok földrajzi helyzetét a nyugati határhoz viszonyított távolság nem befolyásolta. Az ipari hagyományok és a magasabb iskolázottságú területek helyzete azonban mindkét tulajdonosi körnek számítottak. Kőrösi (2006) kimutatta, hogy erős a kapcsolat az ágazatok és a külföldi tulajdon között. Ezeket a magyarázóváltozók közötti kapcsolatokat azonban általában nem veszik figyelembe a Blinder-Oaxaca dekompozícióban használt bérfüggvények becslésénél.

Általánosságban a gépi tanuló algoritmusok képesek jobban rátanulni az adatokban rejlő nemlineáris összefüggésekre és ezáltal jobb előrejelzéseket adni (James et al. (2013)), amit kihasználok a férfi és női bérfüggvények meghatározásánál. Az előrejelzések pontosságának – a 4. fejezethez képest – már nagyobb szerep jut ebben a fejezetben. Azonban a döntési fák előrejelzési képessége és robusztussága gyenge, emiatt ebben a fejezetben a bérfüggvényeket véletlen erdővel becslem meg. Pontosabban a 3.4. alfejezetben bemutatott regressziós erdőt használom ebben a fejezetben.

A véletlen erdővel készített Blinder-Oaxaca felbontás általános leírását Takács–Vincze (2019a) részletesen tartalmazza. Magyar adatokon a Breiman és szerzőtársai (1984) által készített véletlen erdő algoritmus – OLS-sel szembeni – jobb előrejelzési képességét Takács–Vincze (2019b) több évre vonatkozóan igazolták. Céлом, hogy – Takács–Vincze (2019a) és Takács–Vincze (2019b) alapján – összehasonlítsam a regressziós erdővel és az OLS-sel becsült bérfüggvényekkel készült Blinder-Oaxaca felbontást. Az összehasonlítás megmutatja, hogy a bérek feltételes várhatóérték függvényének a közelítése egy hagyományos ökonometriai módszerrel és egy tanuló algoritmussal végső soron milyen mértékben befolyásolja a Blinder-Oaxaca dekompozícióból származó eredményeket. Az összehasonlítás több lépésben történik, ugyanis összehasonlítom az

átlagos magyarázott és nem magyarázott részt. Majd ezt követően megvizsgálom, hogy a két módszerrel egyéni szinten számszerűsített nem magyarázott részek milyen mértékben különböznek.

Az 5.2. alfejezet az elemzéshez használt Bértarifa adatokat tárgyalja. Az 5.3. alfejezetben bemutatom a Blinder-Oaxaca dekompozíció elméleti felírását és gyakorlati megközelítését. Az 5.3.1. alfejezet tartalmazza azt, hogy hogyan írható fel a dekompozíció a férfi és a női bérekre vonatkozó feltételes várhatóérték függvények segítségével. Ez a matematikai felírás segít annak megértésében, hogy milyen tényezők és hogyan határozzák meg a magyarázott és a nem magyarázott részek nagyságát. Az 5.3.2. alfejezet tartalmazza azt, hogy a Blinder-Oaxaca dekompozícióban a feltételes várhatóérték függvény hogyan közelíthető OLS regresszióval és regressziós erdővel. Az erdőhöz kapcsolódóan a hiperparaméterek beállítását az 5.3.3. alfejezet tartalmazza. Az 5.4. alfejezetben három lépésben vetem össze az OLS-sel és az erdővel készült eredményeket: 1) a becsült bérfüggvényekből származó előrejelzéseket vetem össze; 2) a két módszer esetében a Blinder-Oaxaca dekompozíció eredményeként kapott magyarázott és nem magyarázott rész nagyságát hasonlítom össze; 3) az egyéni szinten számolt nem magyarázott részt elemzem leíró statisztikai módszerekkel. Az 5.5. alfejezet összefoglalja a tanulságokat.

5.2. Adatbázis

A 4. fejezet eredményei alapján néhány változó másként szerepel ebben a fejezetben. Az iskolai végzettséget, a külföldi tulajdont és az állami részarányt rendezett változóként kezelem ebben a fejezetben. Továbbá az egyjegyű helyett, kétjegyű FEOR besorolást használok a bérfüggvények becsléséhez, melyet faktorváltozónak tekintek.

Az adattisztítás után az adatbázisban 112996 megfigyelés maradt³, melyek-

³A 4. fejezethez képest kevesebb megfigyelés alkotja az adatbázist ebben a fejezetben, ugyanis a tesztmintán való illesztésnél gondot okoztak azok a kétjegyű FEOR kódok, amelyekben kevesen dolgoztak. Emiatt ezeket az egyéneket eltávolítottam az adatbázisból.

ből 50000 véletlenszerűen kiválasztott megfigyelés alkotja a tanulómintát, a többi pedig a tesztmintában szerepel. A tanuló- és tesztminta megkülönböztetése bevett eljárás a gépi tanuló algoritmusok alkalmazásánál. A tanulómintán építem az erdőt, a tesztmintán pedig ellenőrzöm a teljesítőképességét. Ez az összevetés lehetővé teszi az általánosítási képesség és a túltanulás ellenőrzését az erdőnél. Emellett az OLS regresszió esetében is ugyanígy járok el: a tanulómintán becslem meg és a tesztmintán ellenőrzöm a módszer teljesítőképességét.

A bruttó havi átlagkeresetek logaritmusából számolt nyers bérkülönbség az egész mintában 0,1453. A bérek logaritmusában mért átlagos béreket a férfiaknál és a nőknél a tanuló- és a tesztmintában a 13. táblázat mutatja.

13. táblázat. Női és férfi bérek leíróstatisztikái a tanuló- és a tesztmintán

	Tanulóminta			Tesztminta		
Változók	Nők	Férfiak	Összes	Nők	Férfiak	Összes
Mintaelemszám	19158	30842	50000	24305	38680	62985
Átlag	12,34	12,48	12,43	12,34	12,49	12,43
Szórás	0,55	0,62	0,60	0,55	0,63	0,60
Minimum	10,67	11,46	10,67	11,45	10,72	10,72
Maximum	16,12	16,32	16,32	16,11	16,11	16,11
Bérkülönbség	0,1430			0,1472		

Forrás: Bértarifa

A mintában a nemek aránya nem 50-50 százalék, azonban ettől függetlenül az eredmények robusztusságának szempontjából mindkét csoport nagymintának tekinthető. A 13. táblázat alapján a nők átlagbére alacsonyabb mind a tanuló-, mind a tesztmintában. A nyers bérkülönbség az átlagbérek különbségeként számolódik, ami pozitív a tanuló- és a tesztmintában is. A továbbiakban ezeket az átlagos bérkülönbségeket bontom fel Blinder-Oaxaca dekompozícióval.

5.3. Módszertan

Az 5.3.1. alfejezetben először a Blinder-Oaxaca felbontás várhatóérték függvényekkel történő felírását tárgyalom. Majd az 5.3.2. alfejezetben megmutatom, hogy a várhatóérték függvény közelítéseként alkalmazott OLS regresszióval és regressziós erdővel hogyan lehet meghatározni a magyarázott és a nem magyarázott rész nagyságát. A módszertani rész végén, az 5.3.3. alfejezetben a regressziós erdőnél használt hiperparaméter-beállításokat foglalom össze.

5.3.1. Blinder-Oaxaca felbontás elméleti oldalról

A nemi bérkülönbségek vizsgálatánál a Blinder-Oaxaca dekompozícióban általában a férfiakat tekintik a referencia csoportnak (Jann (2008)), emiatt mindössze két: egy női és egy férfi bérfüggvényt becsülnek. Elméleti megközelítésben a bérfüggvények tekinthetők a bérekre vonatkozó feltételes várhatóérték függvényeknek. Ebben az alfejezetben ez alapján elméleti megközelítésben alapján mutatom be a Blinder-Oaxaca dekompozíció általános felírását.

Jelöljük $E_F(Y)$ and $E_N(Y)$ a férfiak (F) és a nők (N) logbéreinek várható értékét, amíg $E_F(Y | X)$ és $E_N(Y | X)$ a megfelelő feltételes várhatóérték függvények, ahol X a két nemnél megfigyelt változók köre. Így a nyers bérkülönbség a következő formában írható fel:

$$E_F(Y) - E_N(Y) = \int E_F(Y | X) dG_F(X) - \int E_N(Y | X) dG_N(X), \quad (5.1)$$

ahol a $G_F(X)$ és a $G_N(X)$ függvények jelölik a megfigyelt változók eloszlását a két nemnél. Definiáljuk azt a nők bérére vonatkozó elméleti várható értéket, melynél a férfiak várhatóérték függvényét ($E_F(Y)$) a nőkhöz tartozó

megfigyelések eloszlása ($G_N(X)$) felett értékeljük ki:

$$E_{N(F)}(Y) = \int E_F(Y | X) dG_N(X). \quad (5.2)$$

Az (5.1) és (5.2) felhasználásával a következő formában lehet felírni a nyers bérkülönbséget:

$$\begin{aligned} E_F(Y) - E_N(Y) &= \\ (E_F(Y) - E_{N(F)}(Y)) + (E_{N(F)}(Y) - E_N(Y)) &= \\ \left(\int E_F(Y | X) dG_F(X) - \int E_F(Y | X) dG_N(Y) \right) + & \\ \int (E_F(Y | X) - E_N(Y | X)) dG_N(X). & \end{aligned} \quad (5.3)$$

Az X változókkal magyarázott része a nyers különbségnek, mely az összetételhatást ragadja meg:

$$CE = E_F(Y) - E_{N(F)}(Y), \quad (5.4)$$

míg a nem magyarázott része, mely a bérstruktúrahatást jelöli a

$$WS = E_{N(F)}(Y) - E_N(Y). \quad (5.5)$$

Az összetétel- és a bérstruktúrahatás kifejezések Fortin és szerzőtársainak (2011) tanulmányából származnak és jobban megragadják a különbség mibenlétét. Emiatt a továbbiakban ezeket a megnevezéseket használom a magyarázott és a nem magyarázott rész kifejezések helyett.

A bérstruktúrahatás definiálható minden potenciális ($X = x$) tulajdonság szerint és a következőképpen írható fel:

$$HWS(x) = E_F(Y | x) - E_N(Y | x), \quad (5.6)$$

ahol HWS kifejezés jelen esetben az angol szakirodalomban is használt heterogén hatásra utal.

Az elméleti feltételes várhatóérték függvény számszerűsítéséhez becslőfüggvényekre van szükség, melyeket jelöljünk P_F -fel a férfiaknál és P_N -nel a nőknél. Az egyedi megfigyelésekre alkalmazva a becslőfüggvényeket, minden megfigyelt

személyhez hozzárendelhető egy bérelőrejelzés. Ezekből az egyéni előrejelzések közül pedig meghatározható a férfi és a női bérek átlaga. Amennyiben a férfiakat j -vel, a nőket pedig i -vel indexáljuk, akkor a következő kifejezés igaz:

$$\begin{aligned}\overline{P_F} &= \frac{1}{\#N_j} \sum_j P_F(j) \\ \overline{P_N} &= \frac{1}{\#N_i} \sum_i P_N(i),\end{aligned}\tag{5.7}$$

ahol a $\#N_j$ és $\#N_i$ kifejezések a férfiak és a nők számosságát jelölik. Az $E_{N(F)}(Y)$ várhatóértékhez hasonló módon definiálható egy $P_{N(F)}$ függvény, melyhez tartozó átlag a következő:

$$\overline{P_{N(F)}} = \frac{1}{\#N_i} \sum_i P_{N(F)}(i).\tag{5.8}$$

A mintából számolt átlagokra vonatkozóan a következő összefüggéssel ragadható meg a nyers bérkülönbség:

$$\overline{Y_F} - \overline{Y_N} = (\overline{P_F} - \overline{P_{N(F)}}) - (\overline{P_{N(F)}} - \overline{P_N}) + \text{torzítás}.\tag{5.9}$$

Általános felírásban megjelenhet a torzítás, mivel nem biztosított, hogy az előrejelzések átlaga megegyezik a mintaátlaggal. Az (5.9) jobb oldalán szereplő első tag mutatja az összetételhatás nagyságát, vagyis mindazt ami a két nemnél a változók eloszlásbeli különbözőségéből származik. A második tag a bérstruktúrahatás, amely a feltételes várhatóérték függvények közti különbségből adódik. Amennyiben a bérstruktúrahatás különbözik nullától, akkor feltételezhető, hogy a bérek meghatározása eltér a két nem esetében.

A mintán is számszerűsíthetőek a heterogén bérstruktúrahatások és azok átlaga. A Blinder-Oaxaca dekompozícióban a teljes bérstruktúrahatás az (5.3) alapján csak a nőkre számolható. Emiatt a heterogén bérstruktúrahatások becslése is csak rájuk vonatkozik és azt mutatja, hogy adott nőnél mekkora a különbség a férfi és a női bérfüggvénnyel készült előrejelzések között:

$$HWS(i) = P_F(i) - P_N(i).\tag{5.10}$$

Az egyéni szinten számszerűsített heterogén bérstruktúrahatások átlagolásával pedig a dekompozícióban szereplő bérstruktúrahatás is meghatározható a következő módon:

$$\frac{1}{\#N_i} \sum_i HWS(i) = \overline{P_{N(F)}} - \overline{P_N}. \quad (5.11)$$

Az egyéni bérstruktúrahatások átlaga definíció szerint megegyezik az (5.9) jobb oldalának második tagjával. Tehát a Blinder-Oaxaca dekompozícióban szereplő teljes bérstruktúrahatás nem más, mint a heterogén bérstruktúrahatások átlaga.

5.3.2. Blinder-Oaxaca felbontás gyakorlati oldalról

A disszertációmban a feltételes várhatóérték függvényeket közelítem OLS regresszióval és regressziós erdővel. OLS-beclés esetén a következő összefüggés igaz:

$$\overline{Y_F} - \overline{Y_N} = (\overline{X_F} - \overline{X_N}) \beta_F - \overline{X_N} (\beta_F - \beta_N) + \text{torzítás}, \quad (5.12)$$

ahol $\overline{Y_F} - \overline{Y_N}$ a nyers bérkülönbség, $\overline{X_F}$ 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 (5.12) jobb oldalán az első kifejezés az összetételhatást, a második a bérstruktúrahatást ragadja meg. OLS esetében a tanulómintán a torzítás definíció szerint nulla, míg a tesztmintán ettől eltérhet.

Az (5.12) megegyezik Blinder (1973) és Oaxaca (1973) eredeti felírásával, amelynél a szerzők úgy gondolták, hogy a bérstruktúrahatás tökéletesen méri a diszkriminációt. Későbbi tanulmányok azonban kimutatták, hogy ez a tag egyéb statisztikai hibákat is tartalmaz: például a kihagyott változók hatását vagy a munkatapasztalat hibás mérését. (Reilly (2001)) Ennek következtében az átlagos bérstruktúrahatás fontos információt hordoz a munkaerőpiacon ta-

pasztalt nemi bérkülönbség 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:

$$HWS(i) = X_i'(\beta_F - \beta_N), \quad (5.13)$$

A regressziós erdő esetében nem becsülhetők a β együtthatók, hanem csak közvetlenül az előrejelzések. Emiatt az erdőknél használható az (5.9), amely összefüggés lényegében megragadja, hogy a Blinder-Oaxaca dekompozíciót hogyan lehet kiterjeszteni a regressziós erdőkre. Ismétlésként az összefüggés a következő:

$$\overline{Y_F} - \overline{Y_N} = (\overline{P_F} - \overline{P_{N(F)}}) - (\overline{P_{N(F)}} - \overline{P_N}) + \text{torzítás}, \quad (5.14)$$

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 bérstruktúrahatások felírása véletlen erdővel becsült bérfüggvények esetében megegyezik az (5.10)-ben szereplő felírással, ahol a két nemhez tartozó modellekkel számszerűsítjük az adott női munkavállaló lehetséges bérét és vesszük annak különbségét.

Az (5.12) alapján látható, hogy az OLS-sel felírt Blinder-Oaxaca felbontásban az egyes magyarázóváltozók marginális hatásai külön-külön is meghatározhatóak, így ezek hatása is vizsgálható a különbségre vonatkozóan. Regressziós erdő esetében azonban csak becslést tudunk adni a bérekre, de nem mérhetőek az egyes változók egyedi hatásai, ahogyan az (5.14) is mutatja. Emiatt a heterogén bérstruktúrahatások változónkénti elemzésével mérem majd az egyes magyarázóváltozók bérstruktúrára gyakorolt hatását.

Ahogyan az (5.10) és az (5.13) egyenletekből látható: az OLS és az erdő esetében is a heterogén bérstruktúrahatás kizárólag a nőknél van értelmezve. Másképpen megfogalmazva, a bérstruktúrahatás csak a nőknél tér el nullától, a férfiaknál nem. Emiatt a heterogén bérstruktúrahatás elemzését csak a nők esetében végzem el: egyéni szinten az OLS regresszióval és a regressziós erdővel

is megbecslem, hogy a férfi és a női bérfüggvényekkel mekkora lenne az adott nő bére és a kettő különbsége mutatja a heterogén bérstruktúrahatás nagyságát. Amennyiben ez az érték 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. Ez a fajta összevetés korlátozott mértékben, de lehetőséget ad annak elemzésére, hogy a két módszer esetében az egyes változók milyen mértékben hatnak a bérstruktúrahatás nagyságára.

5.3.3. Regressziós erdő hiperparaméterei

A becselőfüggvények elkészítéséhez a regressziós erdőnél a hiperparamétereket optimalizálom, mivel a lehető legjobb előrejelzések készítése kiemelten fontos cél a bérek becslésénél. A regressziós erdőket az R grf csomagjának `regression_forest` parancsával készítem el. Az erdőnél a 3.4. alfejezetben bemutatott hiperparaméterek közül a fák számát adottnak veszem és a többi paramétert optimalizálom. A hiperparaméterek optimalizált értékeit a 14. táblázat foglalja össze.

14. táblázat. Regressziós erdő hiperparaméterei

Paraméter	Nők	Férfiak
Mintanagyság	19158	30842
Fák száma	2000	2000
Minimum levélméret	5	5
Változók száma	74	74
Potenciális vágóváltozók száma	21	19
Mintavétel	Visszatevés nélkül	Visszatevés nélkül
Bootstrap mintanagyság	8578	14710
Alfa	0,0142	0,0028
Kiegyensúlyozatlansági büntetés	0,1717	0,7052

Megjegyzés: A változók továbbra is a 1. táblázatban feltüntetett változók, azonban az általánosított véletlen erdők csak folytonos és két értékű kategória változókat tud kezelni. Emiatt a több kategóriából álló változók dummy változókként kerültek bele a modellbe, ami megnövelte a magyarázóváltozók számát.

Forrás: Bértarifa, saját számítás

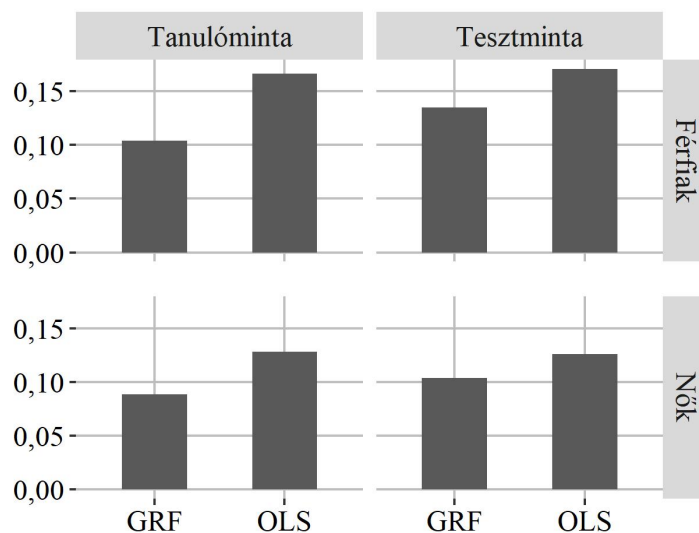
5.4. Eredmények

Az eredményeket négy alfejezetben tárgyalom. Az 5.4.1. alfejezetben az OLS-sel és a regressziós erdővel számolt bérbecslések statisztikai tulajdonságait elemzem. Az 5.4.2. alfejezetben a kétféle becslési módszerrel készített bérfüggvény alapján számolt Blinder-Oaxaca felbontás eredményeit, az összetétel- és a bérstruktúrahatás nagyságát hasonlítom össze. Az utolsó két alfejezetben az OLS-sel és véletlen erdővel számolt heterogén bérstruktúrahatásokat vetem össze. Az 5.4.3. alfejezet a kétfajta módon számolt heterogén bérstruktúrahatások általános összehasonlítását, míg az 5.4.4. alfejezet az egyéni szintű bérstruktúrahatások változónkénti összevetését tartalmazza.

5.4.1. Férfi és női bérfüggvények összehasonlítása

Az OLS-sel és egy regressziós erdővel számolt előrejelzések pontosságát az átlagos négyzetes eltérés (MSE) alapján hasonlítom össze. Az eredményt a 10. ábra mutatja.

10. ábra. Átlagos négyzetes eltérés a regressziós erdővel (GRF) és az OLS regresszióval (OLS) készült nemi bérfüggvények esetében a tanuló- és a tesztmintán



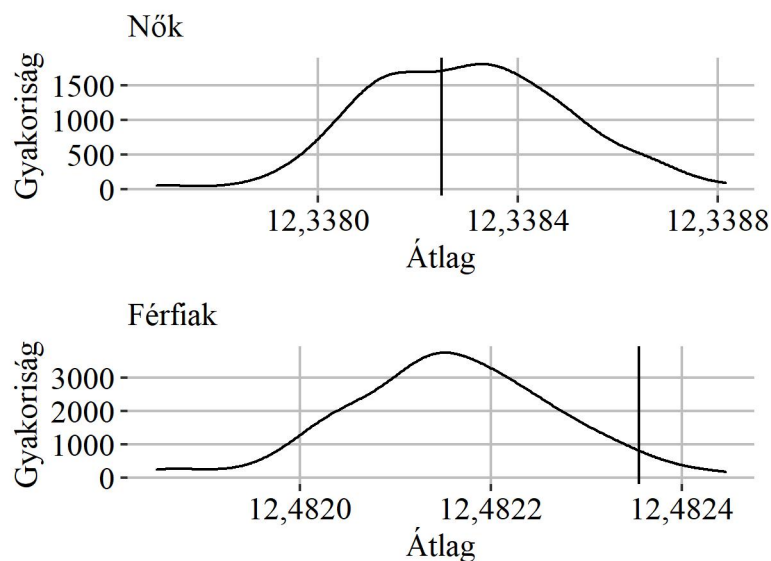
Forrás: Bértarifa, saját számítás

A 10. ábrán látható, hogy az átlagos négyzetes eltérés az erdőknél kisebb, mint az OLS-nél, ami egybevág az általános várakozásokkal egy tanuló algorit-mussal kapcsolatban. Továbbá alátámasztja azt a feltételezést, hogy léteznek nemlineáris összefüggések a magyarázóváltozók között. Az eltérés négyzet-összegek összehasonlítása igazolja, hogy jelen esetben a regressziós erdő jobb előrejelzési képességgel rendelkezik, mint a hagyományos OLS. Azonban az MSE emelkedése a tesztmintán a véletlen erdőnél arra utal, hogy az eljárás rátanul a tanulóminta egyes sajátosságaira, emiatt romlik az általánosítási ké-pessége a tesztmintán. A nőknél az MSE kisebb mindkét mintán, ami arra

vezethető vissza, hogy a nők bére kevésbé szóródik a két almintában, ahogyan a 13. táblázat is mutatja.

A regressziós erdővel kapcsolatban fontos megemlíteni, hogy az eljárás – nem úgy mint az OLS – adott megfigyeléshalmazon való többszöri alkalmazása eltérő eredményekre vezet. Ennek oka, hogy az erdőben szereplő fák elkészítésénél a véletlen fontos szerepet játszik mind a fa építéséhez használt almintá, mind a vágási pontokban a potenciális vágóváltozók meghatározásánál. Annak következtében, hogy az erdők becslésenként változhatnak, száz véletlen erdővel vizsgálom az eredmények robusztusságát. A száz erdőnél a 14. táblázatban szereplő hiperparaméterekkel dolgozom. Minden személynél a saját neméhez tartozó becslőfüggvénnyel határozom meg a bért. Így a száz erdővel száz darab bért becslek minden megfigyelt személyre. Ezt követően az egyes megfigyelésekhez tartozó bérbecslések átlagait, valamint a 10. ábrához tartozó erdővel készített predikciók átlagait tüntetem fel a 11. és a 12. ábrákon.

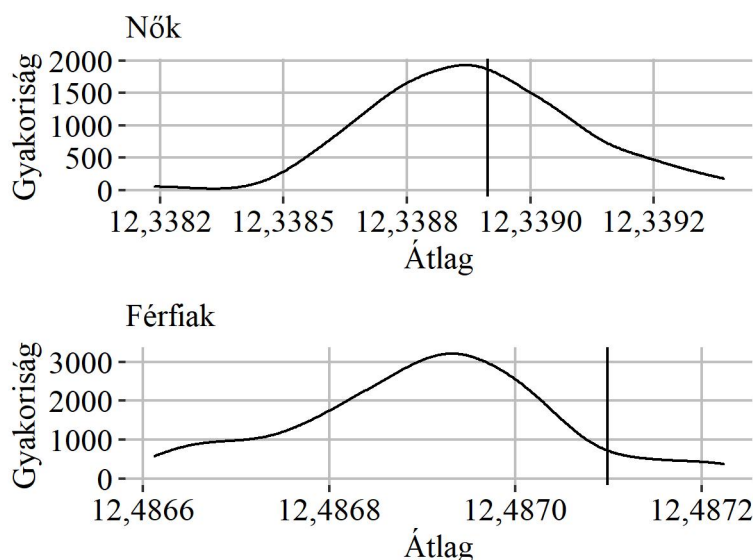
11. ábra. Átlagok megoszlása a 100 erdő esetében a tanulómintán



Megjegyzés: A függőleges egyenesek a korábban használt modellekből számolt átlagokat mutatják.

Forrás: Bértarifa, saját számítás

12. ábra. Átlagok megoszlása a 100 erdő esetében a tesztmintán



Megjegyzés: A függőleges egyenesek a korábban használt modellekből számolt átlagokat mutatják.

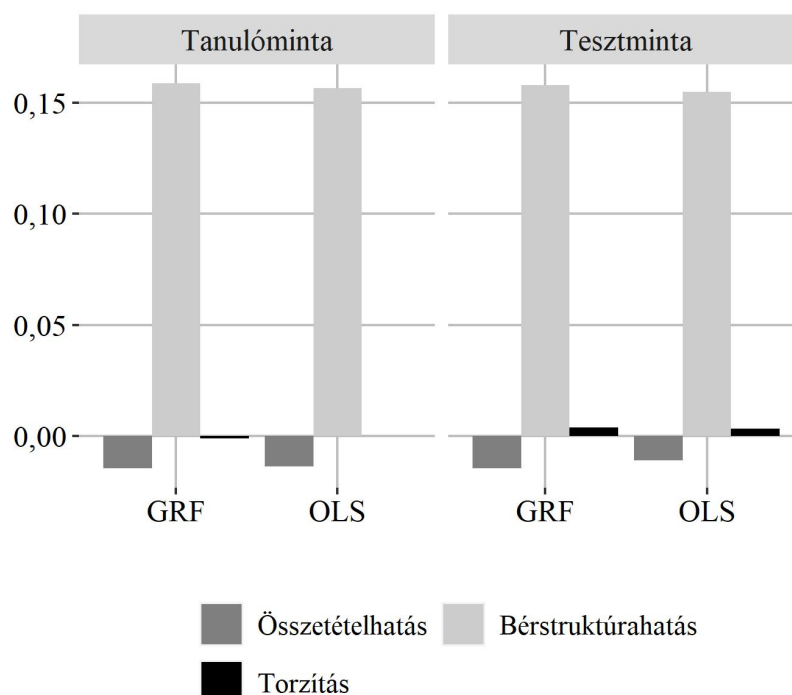
Forrás: Bértarifa, saját számítás

A 11. és 12. ábrák alapján a 100 erdőből számolt átlagok különböznek, azonban szűk tartományon szóródnak: az átlagok általában a harmadik tizedesben különböznek. Az egyenesekkel ábrázolt átlagok torzítottak tűnhetnek, azonban a 100 erdőből számolt átlagok szóródását figyelembevéve a torzítottság mértéke elhanyagolható, az erdő eredményei alapvetően robusztusok. Emiatt a továbbiakban csak egy regressziós erdőből származó eredményeket mutatok be, melyek átlagait a 11. és a 12. ábrákon a függőleges egyenesek szemléltetik.

5.4.2. Blinder-Oaxaca felbontás

A 13. ábra az 5.3.2. alfejezetben bemutatott Blinder-Oaxaca dekompozíció eredményeit mutatja a kétféle becslőfüggvény esetében.

13. ábra. Blinder-Oaxaca felbontás a regressziós erdő (GRF) és az OLS regresszió (OLS) alapján a tanuló- és a tesztmintán



Megjegyzés. A függőleges tengelyen a bérek logaritmusában mért nemi bérkülönbség szerepel.

Forrás: Bértarifa, saját számítás

A két módszer esetében közel azonos eredményeket kapunk a két mintán. Az OLS esetében a tanulómintán definíció szerint nulla a torzítás mértéke. Ezzel szemben a torzítás a regressziós erdő esetében nullától különböző, azonban – ahogyan a 13. ábra is mutatja – alacsony. A tesztmintákon a torzítás már mindkét becslési eljárásnál megjelenik, azonban ezek mértéke elhanyagolhatóan kicsi. Az alacsony torzítás bizonyítja, hogy a regressziós erdő is alkalmas módszer a Blinder-Oaxaca felbontás elkészítéséhez.

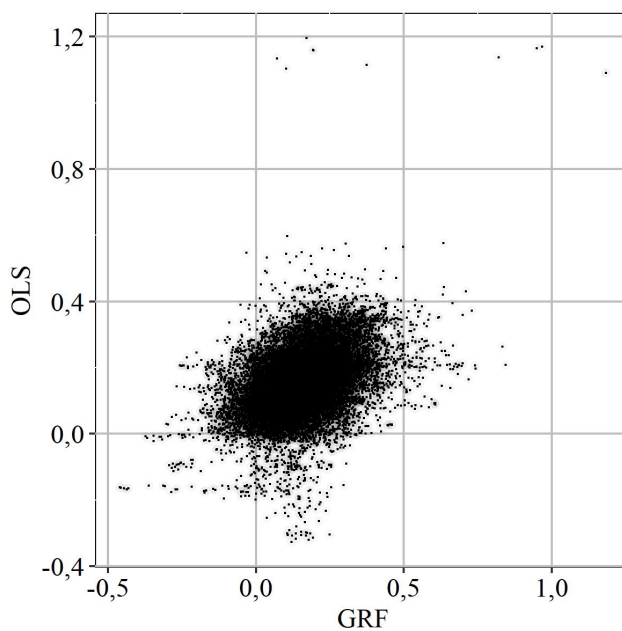
A 13. ábra alapján az összetételhatás abszolút mértéke mindkét mintán magasabb az erdő esetében. Ennek következtében a bérstruktúrahátás viszont alacsonyabb a regressziós erdőnél. Azonban az összetétel- és a bérstruktúrahátások így is hasonlóan alakulnak a két módszerrel. A bérstruktúrahátások az OLS-sel és az erdővel készült bérbecslések esetében is jelentősen magasabb,

mint az összetételhatás. A bérstruktúrahatások mindkét mintán, mindkét módszer esetében pozitívak, tehát átlagosan a nőket alulárazzák a férfiakhoz képest.

5.4.3. Heterogén bérstruktúrahatások

A 14. ábra az OLS-sel és a regressziós erdővel készült bérfüggvények esetében a heterogén bérstruktúrahatásokat mutatja a tanulómintán. Ezeket a heterogén bérstruktúrahatásokat elemzem tovább a következőkben.

14. ábra. Heterogén bérstruktúrahatások az OLS regresszióval (OLS) és a regressziós erdővel (GRF) becsült bérfüggvények esetében a tanulómintán



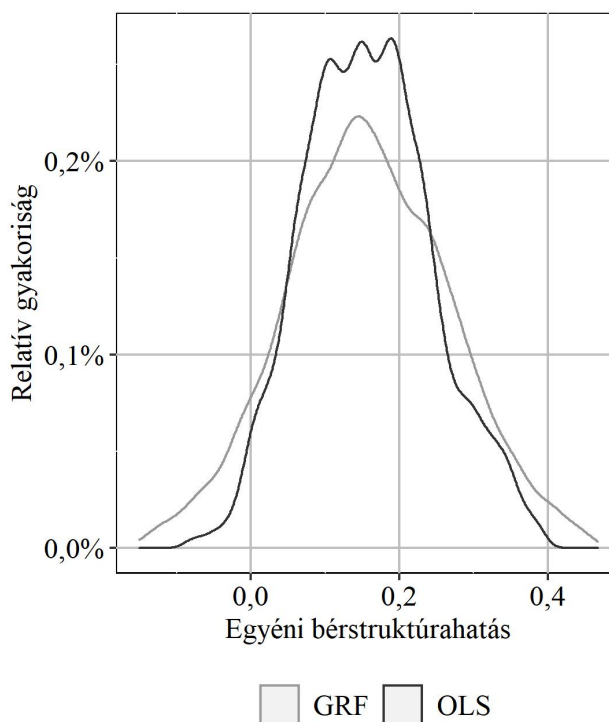
Forrás: Bértarifa, saját számítás

A 14. ábrán látható, hogy a heterogén bérstruktúrahatások az erdőnél szélesebb, az OLS-nél pedig szűkebb sávban szóródnak. Emellett találhatóak kiugró értékek mindkét becslés esetében, melyeket a további vizsgálatokhoz

kivettem a megfigyelések közül.⁴ A kilógó értékek kiszűrése után a tanulóminta kezdeti 19824 megfigyeléséből összesen 18892 marad meg.

A 15. ábra alapján a kilógó értékek nélküli mintán a két módszerrel számolt heterogén bérstruktúrahatások eloszlásai közel normálisak, valamint az átlagok is hasonlóan alakulnak. A heterogén bérstruktúrahatások jobban szóródnak az erdő esetében, ami miatt ennek az eloszlása laposabb. A kapott eloszlásokat Kolmogorov–Szmirnov teszttel hasonlítottam össze, mely szerint a két eloszlás szignifikánsan különbözik.⁵ A 15. táblázat az OLS-sel és az erdővel számolt heterogén bérstruktúrahatások leíró statisztikáit mutatja a tanulómintán és a kilógó értékek nélküli mintában. A két módszerrel készült bérbecslések átlagát mutatja a 16. táblázat.

15. ábra. Heterogén bérstruktúrahatás eloszlása a két módszer szerint a tanulómintán



Forrás: Bértarifa, saját számítás

⁴Azokat az értékeket tekintem kiugrónak és távolítom el a mintából, amelyek a $Q25 - 1,5 * IQR$ és a $Q75 + 1,5 * IQR$ tartományon kívülre esnek. Az IQR az eloszlás középső 50 százalékának a terjedelme.

⁵A kétoldali Kolmogorov–Szmirnov-próba tesztstatisztika értéke 0,13057 és $p < 2,2 * 10^{-16}$. Tehát az eloszlások egyezőségét állító H_0 hipotézist elvetem.

15. táblázat. Heterogén bérstruktúrahatalás leíró statisztikái a tanulómintán és a kilógó értékek (outlierek) nélküli mintán

Változók	Tanulóminta		Outlierek nélkül	
	OLS	GRF	OLS	GRF
Átlag	0,1567	0,1587	0,1576	0,1571
Szórás	0,0986	0,1218	0,0864	0,1104
Minimum	-0,3265	-0,4605	-0,0879	-0,1507
Maximum	1,1954	1,1842	0,3978	0,4671
Korreláció	0,4011		0,3882	

Forrás: Bértarifa, saját szerkesztés

16. táblázat. A bérelőrejelzések átlagai (a bérek logaritmusában mérve) a kilógó értékek nélküli mintán

	OLS	GRF	Különbség
Férfi bérfüggvénnyel	12,4862	12,4864	-0,0002
Női bérfüggvénnyel	12,3286	12,3292	-0,0006
Különbség	0,1576	0,1572	

Forrás: Bértarifa, saját szerkesztés

A 16. táblázat mutatja, hogy a számítások alapján a férfi modellel készített előrejelzések átlaga mindkét módszer esetében magasabb. Az erdő esetében a férfi és a női bérfüggvény is magasabb átlagbért becsül, mint az OLS-sel készített bérfüggvények. Az erdő alapján kicsit alacsonyabb az átlagos bérstruktúrahatalás, mint amit az OLS alapján kapnánk.

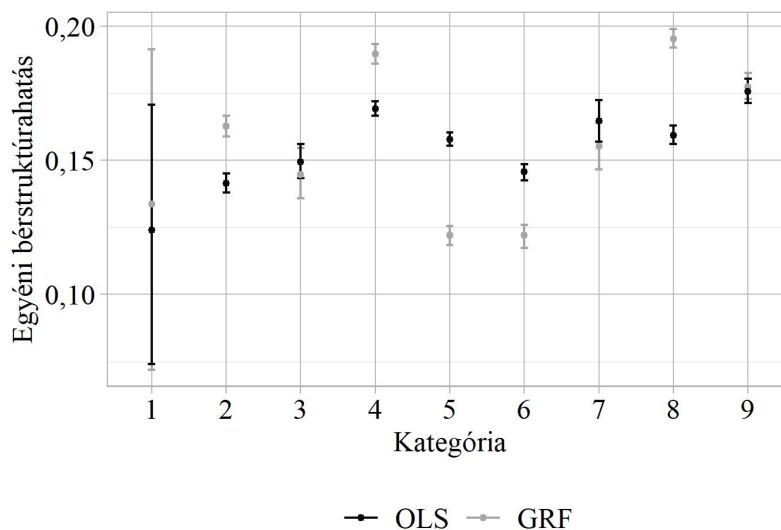
A 15. táblázat alapján a heterogén bérstruktúrahatalások átlagai hasonlóan alakulnak, azonban közepesen erős a korreláció közöttük. Tehát egy megfigyeléshez tartozóan a két módszerrel számolt heterogén bérstruktúrahatalások között lehetnek eltérések. Ezt a különbséget akarom magyarázni a következőkben a bérstruktúrahatalások változónkénti elemzésével.

5.4.4. Heterogén bérstruktúrahatások változonként

Ebben az alfejezetben a heterogén bérstruktúrahatások alakulását változonként vetem össze. Célom annak vizsgálata, hogy az egyes változóknál, azon belül is az egyes kategóriáknál mekkorák az eltérések az átlagos bérstruktúrahatás tekintetében. Az eltérés elemzése megmutatja, hogy melyik változó okozza az OLS-sel és az erdővel számolt heterogén bérstruktúrahatások közti különbséget. Az egyes változók megoszlásai az F.7. Függelékben találhatóak.

Az iskolai végzettség esetében az OLS-sel és az erdővel számolt heterogén bérstruktúrahatások átlagai különbözőképpen alakultak, ahogyan a 16. ábra is mutatja.

16. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az iskolai végzettségnél



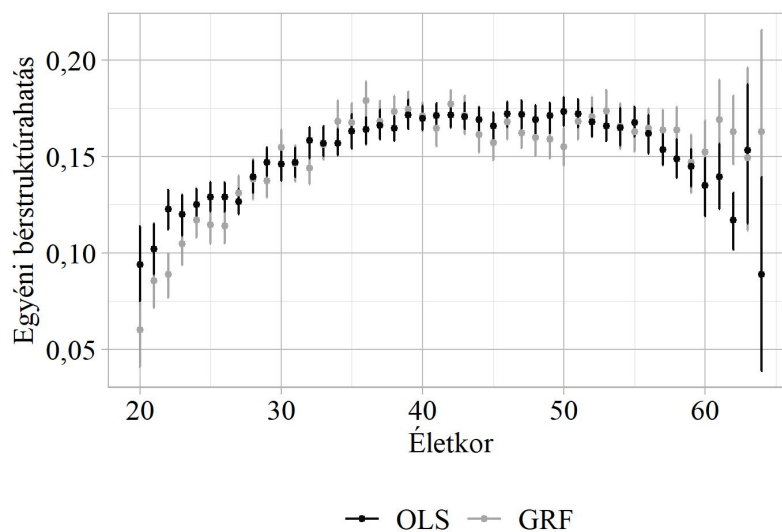
Megjegyzés: Az iskolai végzettség kategóriái: 1: Általános iskola 0-7 osztály, 2: Általános iskola 8 osztály, 3: Szakiskola, 4: Szakmunkásképző iskola, 5: Szakközépiskola, 6: Gimnázium, 7: Technikum, 8: Főiskola, 9: Egyetem. A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

Forrás: Bértarifa, saját számítás

A 16. ábra alapján az OLS és az erdő is alacsony átlagos bérstruktúrahatást becsül a 0-7 osztályt végzetteknel, azonban – tekintve a megoszlásokat – kevés számú megfigyelés tartozik ehhez a kategóriához és a bérstruktúrahatások is erősen szóródnak. A regressziós erdő alapján a legalacsonyabb bérstruktúrahatás a szakközépiskolánál (5. kategória) és a gimnáziumnál (6. kategória) figyelhető meg. Az OLS nagyobb átlagokat becsül ezeknél a végzettségi szinteknél. A legkisebb különbségek a két módszerrel számolt átlagok között a szakiskolát (3. kategória) és az egyetemet végzettek (9. kategória) között figyelhető meg.

A 17. ábrán a 20-64 éves korosztályban az egyes évekhez tartozó megfigyelések átlagos bérstruktúrahatásait tüntetem fel. Ebben a korosztályban az átlagok kevésbé szóródnak, mint ezen kívül, valamint kellő számú megfigyelés áll rendelkezésre minden életkorhoz⁶.

17. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az életkornál



Megjegyzés: A könnyebb áttekinthetőség miatt az ábrán csak a 20 és 64 év közötti életkorokat tüntettem fel. A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

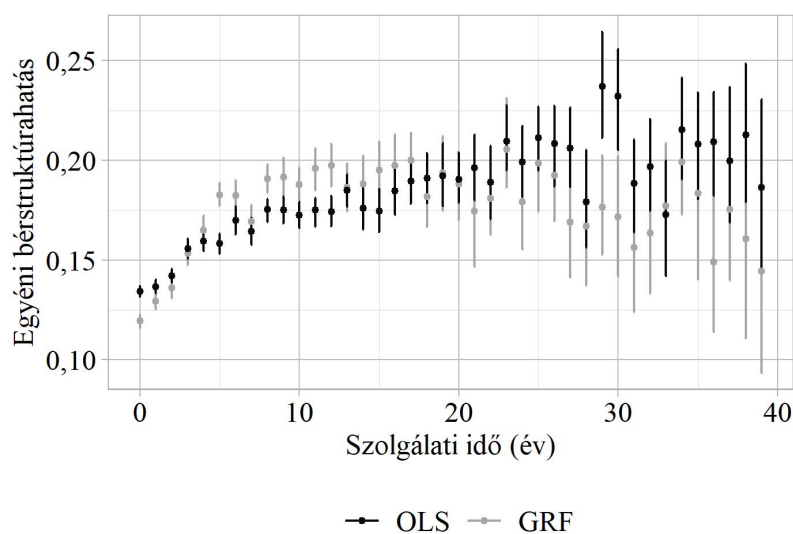
Forrás: Bértarifa, saját számítás

⁶A 20 év alatti és a 64 év feletti korosztályban életkoronként kevesebb, mint 20 megfigyelés található, továbbá a bérstruktúrahatások jelentős szóródást mutatnak.

A 20-64 éves korcsoportban az életkor növekedésével az átlagos bérstruktúrahatás először növekszik, majd az 50. évtől mérséklődik. A két módszer szerinti becslési eredmények meglehetősen közel esnek egymáshoz. 51 éves korig az OLS, később viszont a regressziós erdő becsl nagyobb bérstruktúrahatást. Gábor (2008) szerint a férfiak bérfüggvénye a nőkével összehasonlítva magasabbról indul és a kezdeti erőteljesebb emelkedése következtében konkávabb, ami magyarázhatja a nemi bérkülönbségek kezdeti növekedését a 17. ábrán.

A Blinder-Oaxaca felbontáshoz becsült bérfüggvényekben a szolgálati idő hónapokban szerepel, azonban ezt a változót a 18. ábrán – a könnyebb értelmezhetőség miatt – években szerepeltetem.

18. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai a szolgálati éveknél



Megjegyzés: A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

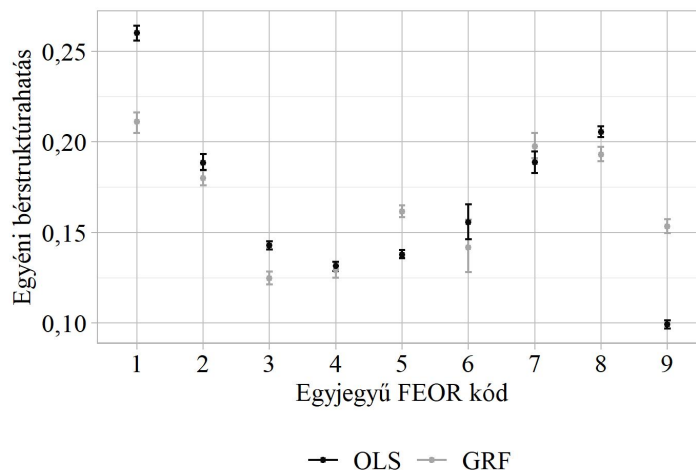
Forrás: Bértarifa, saját számítás

A szolgálati idő alapján megállapítható, hogy a bérstruktúrahatás a vállalathoz való belépéskor – mely a 18. ábrán az első évnél található – sem nulla. Emiatt feltételezhető, hogy a nők már kezdetben kisebb bérért vállalják

a munkát, mint az ugyanolyan megfigyelt munkaerőpiaci jellemzőkkel rendelkező férfi kollégáik. A bérstruktúrahatás a 10. évig emelkedik, ami után további egyértelmű trend nem látható az alakulásában. Az eredmények felhívják a figyelmet arra a tényre, hogy a heterogén bérstruktúrahatások az emelkedő szolgálati időkkel egyre inkább szóródnak, ahogyan az adott szolgálati évhez kapcsolódóan egyre csökkennek a megfigyelésszámok.

A 19. ábra az egy- és a 20. ábra a kétjegyű FEOR-okhoz tartozó átlagos bérstruktúrahatásokat mutatják, melyek közül a Blinder-Oaxaca felbontásban az utóbbit használtam. A 19. ábra azt mutatja, hogy a két módszerrel számolt átlagos bérstruktúrahatások együttlmozognak. Az 1-es – gazdasági, igazgatási, érdek-képviselési vezetők és törvényhozók – és 9-es FEOR-ban – szakképzettséget nem igénylő foglalkozásokban – dolgozók között látszik a legnagyobb eltérés az OLS és az erdő között. A legalacsonyabb bérstruktúrahatás mindkét módszer esetében a 4-es FEOR-ban – irodai és ügyviteli (ügyfélkapcsolati) foglalkozások esetében – figyelhető meg és az átlagok ebben a foglalkozásban esnek a legközelebb egymáshoz.

19. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az egyjegyű FEOR-oknál

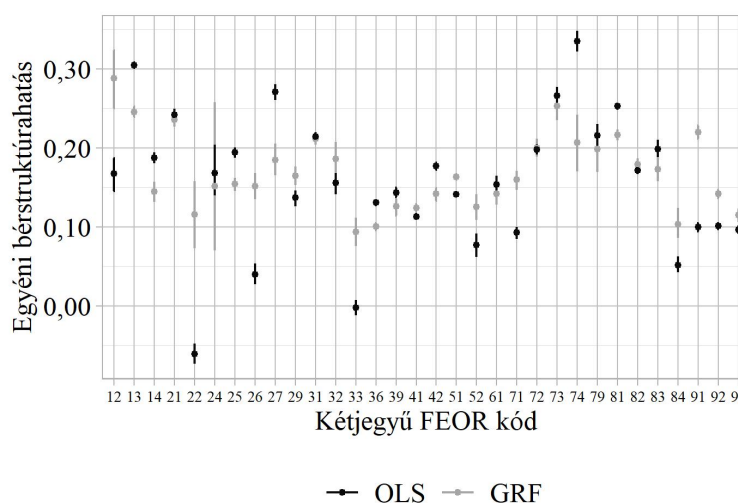


Megjegyzés: A FEOR struktúrát lásd az F.1. Függelékben. A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

Forrás: Bértarifa, saját számítás

A kétjegyű foglalkozáskódokat bemutató 20. ábra azt mutatja, hogy az egyjegyű kategóriákon belül a foglalkozások között lehetnek eltérések a két módszer alapján. Például a 2-es – felsőfokú képzettség önálló alkalmazását igénylő – foglalkozásoknál az OLS-sel és az erdővel becsült átlagos bérstruktúrahátások hasonlóan alakulnak, azonban a kétjegyű FEOR-nál látható, hogy leginkább a 21-es és 29-es FEOR-ban esnek nagyon közel egymáshoz a két módszer szerinti átlagok, míg a többi FEOR-ban nagyobbak a különbségek.

20. ábra. Átlagos bérstruktúrahátások és azok 95 százalékos konfidencia intervallumai a kétjegyű FEOR-oknál



Megjegyzés: A FEOR struktúrát lásd az F.1. Függelékben. A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

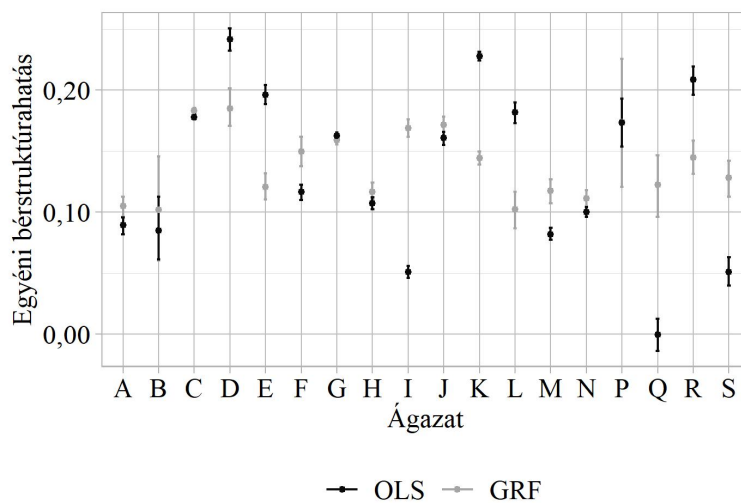
Forrás: Bértarifa, saját számítás

A 21. ábrán a nemzetgazdasági ágak szerinti átlagos bérstruktúrahátások szerepelnek. Az A és B ágazatoknál a regressziós erdő hasonló átlagos bérstruktúrahátásokat mutat. A C-J ágazatoknál az erdővel számolt átlagok közelebb helyezkednek el egymáshoz, mint az OLS-sel kapottak. Az N ágazatban alacsonynak mondható a bérstruktúrahátása, aminek oka, hogy ezekben ágazatokban jelentős a költségvetési szektor aránya⁷ és a költségvetési szférában meghatározott bérszint befolyásolhatja a magánszektorban lévőt is. (Te-

⁷Lényeges megjegyezni, hogy ebben a kutatásban kizárólag a versenyszférában foglalkoztatottakat vizsgálom.

legdy (2013)) Lovász (2013) pedig kimutatta, hogy a költségvetési szektorban kisebbek a nemi bérkülönbségek, ami végső soron az N ágazatban a versenyszektorban is csökkentheti a nemi bérkülönbségeket. Az egészségügyet jelölő Q ágazatban a két módszerrel számolt átlagos hatások jelentősen eltérnek. Az OLS 0 körüli, míg a regressziós erdő 12 százalékos bérstruktúrahatást becsül ebben az ágazatban.

21. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az ágazatoknál



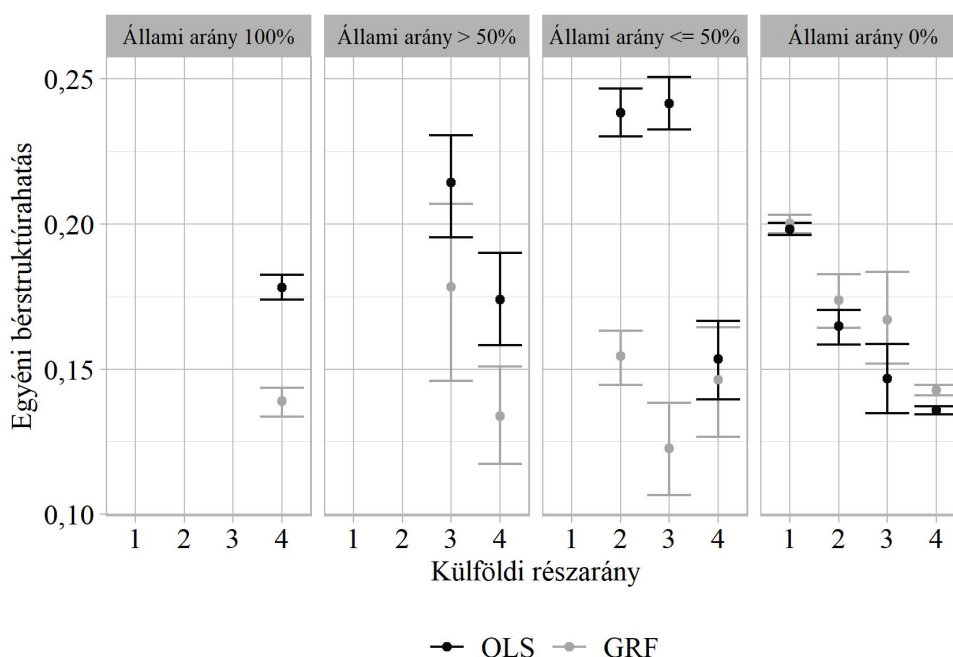
Megjegyzés: Az ágazati struktúrát lásd a F.1. Függelékben.
A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

Forrás: Bértarifa, saját számítás

A 22. ábrán az állami és a külföldi részarány összevetésénél erős módszertani eltérés látható. A teljes mértékben állami tulajdonban lévő vállalatok munkavállalói szerepelnek az első panelen. Az ezen a panelen látható különbség az átlagos bérstruktúrahatások közötti módszertani eltérésre hívja fel a figyelmet. A többségben állami tulajdonú cégeknél – a 2. panelen – is hasonló mértékű az átlagok közti eltérés. A kevesebb, mint 50 százalékos állami részarányú cégeknél – a 3-as panelen – a külföldi részarány 2-es és 3-as kategóriájánál hasonlóan alakulnak az átlagok az OLS-sel becsült bérfüggvények esetében, míg a regressziós erdővel kisebb átlagok adódnak. A külföldi tulajdon 4. kategóri-

ájában a 3. panelen az OLS és az erdő eredményei közelebb esnek egymáshoz. A 0 százalékos állami részaránynál – a 4. panelen – nagyságrendileg megegyeznek az átlagos bérstruktúrahatások a kétféle becslési módszernél. Mindkét eljárás képes megragadni azt az összefüggést, hogy a magasabb külföldi részarány nagyobb bérstruktúrahatast von maga után. Az eredmények egyezősége vélhetően arra vezethető vissza, hogy itt található a legtöbb megfigyelés.

22. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai az állami és külföldi tulajdonú vállalatoknál



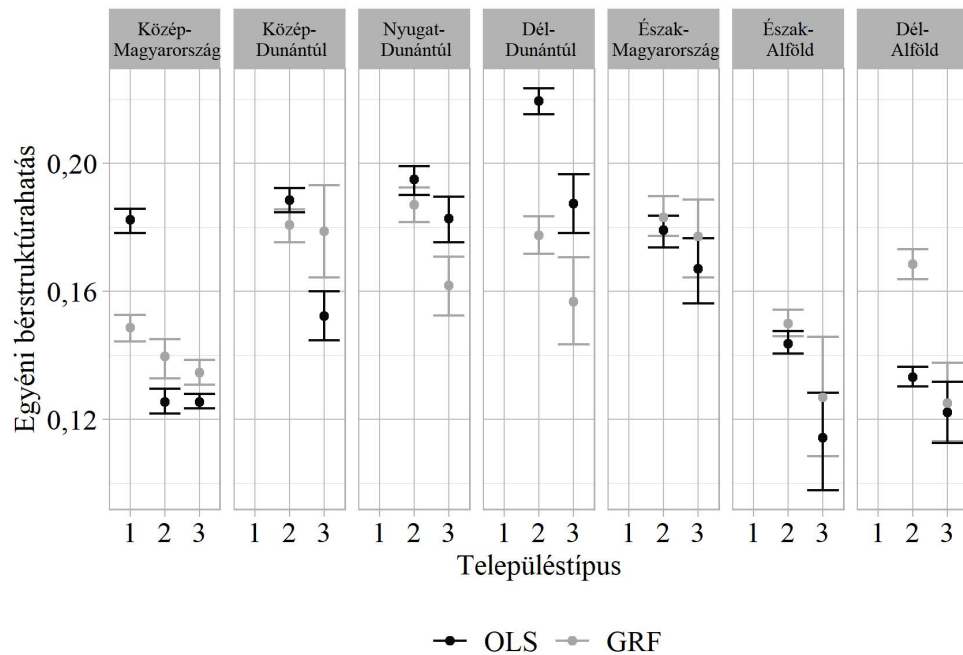
Megjegyzés: A külföldi részarány kategóriák: 1. 100% külföldi tulajdon, 2. >50% külföldi tulajdon, 3. <50% külföldi tulajdon, 4. 0% külföldi tulajdon. A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

Forrás: Bértarifa, saját számítás

A régióra és a településtípusra vonatkozó eredmények a 23. ábrán láthatóak. A két földrajzi változó összevetése alapján Budapesten a többi Középmagyarországi településtípussal összehasonlítva magasabb a bérstruktúrahatast. (Lásd a 23. ábra első paneljén.) A regressziós erdővel számolt átlagok az egyes településtípusoknál hasonlóan alakulnak Közép-Magyarországon, míg az

OLS-sel számoltak között nagyobbak a különbségek. Közép-Magyarországon alacsonyabb a bérstruktúrahatás a városoknál és az egyéb településtípusoknál, mint néhány másik régióban. Az Észak-magyarországi és az alföldi régiókban a regressziós erdő mind a városokban, mind az egyéb településtípusokban nagyobb bérstruktúrahatást mér, mint az OLS. Ezzel szemben a dunántúli régiókban – egy kivételtől eltekintve – az OLS mér magasabb hatást. Ez a kettősség felhívhatja a figyelmet arra, hogy lehetnek kereszthatások az adatokban, amelyet az OLS nem képes megragadni. Általában a bérstruktúrahatás nagyobb a városokban, mint az egyéb településtípusokban.

23. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidencia intervallumai a magyar régiókban a különböző településtípusoknál

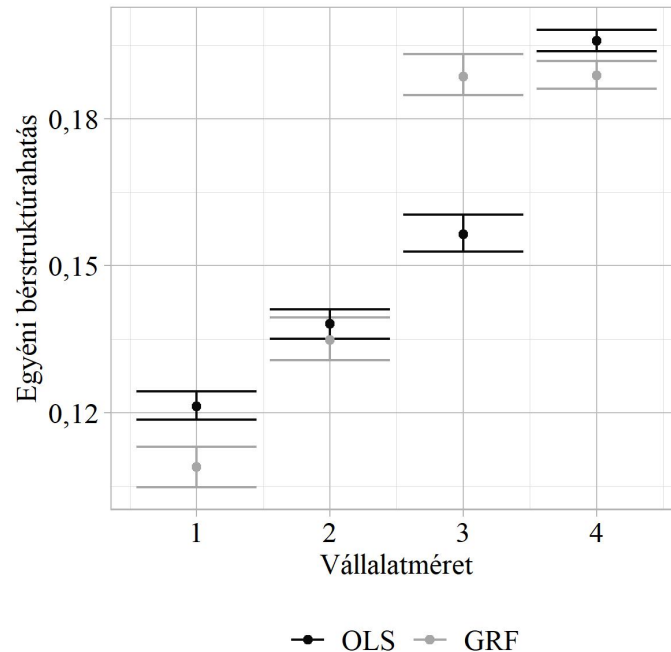


Megjegyzés: Településtípusok kategorizálása: 1. főváros, 2. város, 3. egyéb. A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

Forrás: Bértarifa, saját számítás

A bérfüggvény becslésénél a folytonos skálán mért vállalati létszám szerepel. A 24. ábrán azonban – a könnyebb áttekinthetőség miatt – összevontabb létszám kategóriák láthatóak.

24. ábra. Átlagos bérstruktúrahatás és azok 95 százalékos konfidencia intervallumai a vállalatméret (létszám) függvényében



Megjegyzés: Vállalatméret kategóriák: 1. 20-49 fő, 2. 50-149 fő, 3. 150-499 fő, 4. 500 fő és afelett.

A 95 százalékos konfidencia intervallumok 100 bootstrap mintavétel alapján készültek.

Forrás: Bértarifa, saját számítás

A 24. ábra alapján a vállalatméret növekedésével az OLS-sel számolt bérstruktúrahatás egyre magasabb. Ez a lineáris kapcsolat az 1-3 kategóriáknál az erdőnél is látszik. Azonban az erdőnél 3. és 4. kategóriában a bérstruktúrahatások nagyságrendileg azonosak. A módszerek közti különbség oka, hogy az OLS becslésben a létszám és a bér között lineáris kapcsolatot határoztam meg. A regressziós erdő viszont rámutat, hogy a bérstruktúrahatás és a létszám között feltételezhetően nem lineáris a kapcsolat.

5.5. Összefoglalás

Ebben a fejezetben az OLS-sel és a regressziós erdővel becsült bérfüggvényeket használó Blinder-Oaxaca felbontás eredményeit vetettem össze magyar adatokon. Az összehasonlítás célja annak bemutatása, hogy mennyiben különböznek az eredmények a módszertan függvényében.

Az OLS-sel és a regressziós erdővel kapott becslések összehasonlítása rámutat arra, hogy az erdővel készített bérelőrejelzések pontosabbak. Azonban az összehasonlítás nem igazságos, mivel igazodva korábbi kutatásokhoz (Jolliffe–Campos (2005), Blau–Kahn (2017), Leythienne–Ronkowski (2018)), a regressziós becslésben az életkor négyzetén kívül semmilyen egyéb nemlineáris kapcsolat nem szerepel.⁸ A regressziós erdő – a módszertanából kiindulva – azonban rá tud tanulni az adatban található rejtett összefüggésekre. 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 az OLS-sel és a regressziós erdővel 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 eltéréseért. Inkább a kevesebb megfigyelést tartalmazó csoportok közti átlagos bérstruktúrahatások közötti eltérésre vezethető vissza az eloszlások különbözősége és a becslési

⁸Egy magasabbrendű tagokat, illetve kereszthatásokat is tartalmazó OLS regresszió és egy regressziós erdő összehasonlítását a 6. fejezet tartalmazza.

eredmények közti közepes korreláció.

A megfigyeléseim igazolják Weichselbaumer–Winter-Ebmer (2005) eredményeit, akik a metaanalízisükben többféle módszertant hasonlítottak össze és megállapították, hogy 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.

6. fejezet

Heterogén bérstruktúrahatások időbeli alakulása és a legszélsőségesebb csoportok

6.1. Bevezetés

A 6. fejezetben egy hagyományos paraméteres eljárás és egy gépi tanulási algoritmus eredményeit vetem össze. A fejezet elején a Blinder-Oaxaca felbontáshoz használt férfi és női bérfüggvényeket elkészítem több évre vonatkozóan regressziós erdővel és egy magasabb rendű tagokat, illetve kereszthatásokat is tartalmazó OLS regresszióval¹. Ezek összehasonlítása tekinthető a módszerek közti, illetve a időbeli robusztusság vizsgálatának is. Az időbeli változásokat 2008 és 2016 között vizsgálom.

Ezt követően a heterogén bérstruktúrahatásokat vizsgálom az erdőkkel kapott becslések esetében. Először elemzem a heterogén hatások és néhány változó kapcsolatát. Ez az elemzés elsősorban grafikusan történik és lehetővé teszi az időbeli trendek összevetését és a bérstruktúrahatást befolyásoló munkaerőpiaci változások azonosítása.

¹Az 5. fejezetben az OLS nem tartalmazott kereszthatásokat és csak az életkor négyzete szerepelt benne mint magasabbrendű tag.

A következő lépésben Chernozhukov és szerzőtársai (2018) ajánlását követve klasszifikációs analízist (CLAN) használva meghatározom a legnagyobb és a legkisebb bérstruktúráhatással rendelkező csoportokat és az ezekhez a csoportokhoz tartozó megfigyelések átlagos jellemzőit elemzem. A heterogén bérstruktúráhatás alapján az extrém csoportok azonosítása és a jellemzőik vizsgálata megmutatja, hogy a magyar munkaerőpiacon milyen tulajdonságokkal rendelkeznek azok a nők, akik a legkevésbé vagy a leginkább vannak másként árazva, mint az azonos adottságokkal rendelkező férfi kollégáik.

Ebben a fejezetben az általánosított véletlen erdők közül a – 3.4. alfejezetben tárgyalt – regressziós erdőket használom. A 6.2. alfejezet tárgyalja az adatokat és a regressziós erdőkhöz kapcsolódóan a hiperparaméter-beállításokat. A 6.3. alfejezet összefoglalja az eredményeket: a becslések pontosságának összehasonlítását, az összetétel- és a bérstruktúráhatás alakulását, valamint a heterogén bérstruktúráhatás változónkénti elemzését és a legszélsőségebb csoportok vizsgálatát. A 6.4. alfejezet tartalmazza a diszkussziót.

6.2. Adatbázis és módszertan

Ebben kutatásban a 2008 és 2016 közötti Bértarifa adatokat használom. Minden évben 50000 véletlenszerűen kiválasztott megfigyelés képezi a tanulómintát, a többi megfigyelés pedig a tesztmintában szerepel. A 17. táblázat néhány leíróstatisztikát tartalmaz az adatokat illetően.

17. táblázat. Adatok leíró statisztikái

Év	Nem	Nyers bérkülönbség	Logbérek átlag	Logbérek szórás	Megfigyelések száma
2008	Nők	0,1251	11,93	0,59	41393
	Férfiak		12,06	0,64	64116
2009	Nők	0,0923	12,01	0,61	38266
	Férfiak		12,10	0,67	60346
2010	Nők	0,1279	12,04	0,60	39803
	Férfiak		12,17	0,66	58400
2011	Nők	0,1368	12,07	0,60	39632
	Férfiak		12,21	0,66	58864
2012	Nők	0,1511	12,16	0,56	40481
	Férfiak		12,31	0,63	61069
2013	Nők	0,1410	12,19	0,56	39019
	Férfiak		12,33	0,63	62789
2014	Nők	0,1468	12,20	0,59	40171
	Férfiak		12,34	0,65	66860
2015	Nők	0,1510	12,28	0,56	50851
	Férfiak		12,44	0,63	81108
2016	Nők	0,1457	12,34	0,55	43462
	Férfiak		12,49	0,63	69462

Megjegyzés: A nyers bérkülönbség a férfiak és a nők logaritmusban mért átlagos kereseteinek a különbsége.

Forrás: Bértarifa, saját számítás

A 17. táblázat mutatja, hogy a teljes mintában évente a nők aránya körülbelül 40 százalék. A férfiak logaritmusban mért bére minden évben meghaladja a nők átlagát, ami miatt a nyers különbség minden évben pozitív. Továbbá rávilágít arra, hogy Magyarországon a nemi bérkülönbség a vizsgált időszakban nőtt, míg a nemzetközi tapasztalatok a nemi bérkülönbségek csökkenését dokumentálták. (Blau–Kahn (2017), OECD (2018))

A 3.4. alfejezetben bemutatott hiperparamétereket minden évre vonatkozóan mind a férfi, mind a női bérfüggvények esetében optimalizálom. A hiperparaméterek beállításait a 18. és a 19. táblázatok tartalmazzák.

18. táblázat. Regressziós erdő hiperparamérei a férfiaknál

Paraméterek	2008	2009	2010	2011	2012	2013	2014	2015	2016
Fák száma	2000	2000	2000	2000	2000	2000	2000	2000	2000
Változók száma	68	68	68	71	70	71	71	71	71
Mintanagyság	0,50	0,33	0,50	0,34	0,42	0,36	0,37	0,42	0,34
Potenciális vágóváltozók	29	18	29	16	23	19	19	20	20
Minimális levélméret	5	1	5	5	1	1	2	6	3
Becsületességi arány	0,50	0,65	0,50	0,62	0,70	0,66	0,74	0,69	0,57
α	0,05	0,01	0,05	0,00	0,01	0,04	0,01	0,00	0,00
Kiegyensúlyozatlansági büntetés	0,00	0,24	0,00	0,94	0,47	1,01	0,48	0,27	0,41

Forrás: Bértarifa, saját számítás

19. táblázat. Regressziós erdő hiperparamétereit a nőknél

Paraméterek	2008	2009	2010	2011	2012	2013	2014	2015	2016
Fák száma	2000	2000	2000	2000	2000	2000	2000	2000	2000
Változók száma	68	68	68	71	70	71	71	71	71
Mintanagyság	0,50	0,33	0,50	0,34	0,42	0,50	0,45	0,42	0,45
Potenciális vágóváltozók	29	18	29	16	23	29	21	20	21
Minimális levélméret	5	1	5	5	1	5	2	5	5
Becsületességi arány	0,50	0,65	0,50	0,62	0,70	0,50	0,69	0,69	0,76
α	0,05	0,01	0,05	0,00	0,01	0,05	0,00	0,00	0,01
Kiegyensúlyozatlansági büntetés	0,00	0,24	0,00	0,94	0,47	0,00	1,32	0,27	0,17

Forrás: Bértarifa, saját számítás

A fák száma 2000, ami megfelel a regressziós erdő alapbeállításának és amit nem változtatok meg. Az egyes erdőknél a felhasznált változók számának változását az évek között két tényező határozta meg. Egyrészt, hogy hány foglalkozási kategóriát lehetett megkülönböztetni, ugyanis történt egy áttérés 2011-ben a FEOR-97-ről FEOR-08-ra, ami miatt megváltozott a FEOR besorolások száma. Másrészt a megfigyelések közül kizárom azokat a személyeket, akik – a kétjegyű FEOR alapján – olyan foglalkozásban dolgoznak, ahol a nők, illetve a férfiak száma kevesebb, mint 10 fő. Amennyiben nem zárom ki ezeket a megfigyeléseket, akkor az erdő, illetve az OLS tesztmintára való illesztését nem minden évben lehet elvégezni.

Az erdővel kapott Blinder-Oaxaca dekompozíció eredményeinek robusztuságát OLS regresszióval való összevetéssel vizsgálom, hasonlóan mint az 5. fejezetben. Ebben a fejezetben az OLS már magasabbrendű tagokat és ke-

reszthatásokat is tartalmaz. A felhasznált változók köre megtekinthető az F.8. Függelékben. Emellett az eredmények robusztusságát a tanuló és a tesztminta eredményeinek összehasonlítása is mutatja. A következőkben egy a férfi almintára és egy a női almintára készített regressziós erdő eredményeit mutatom be.

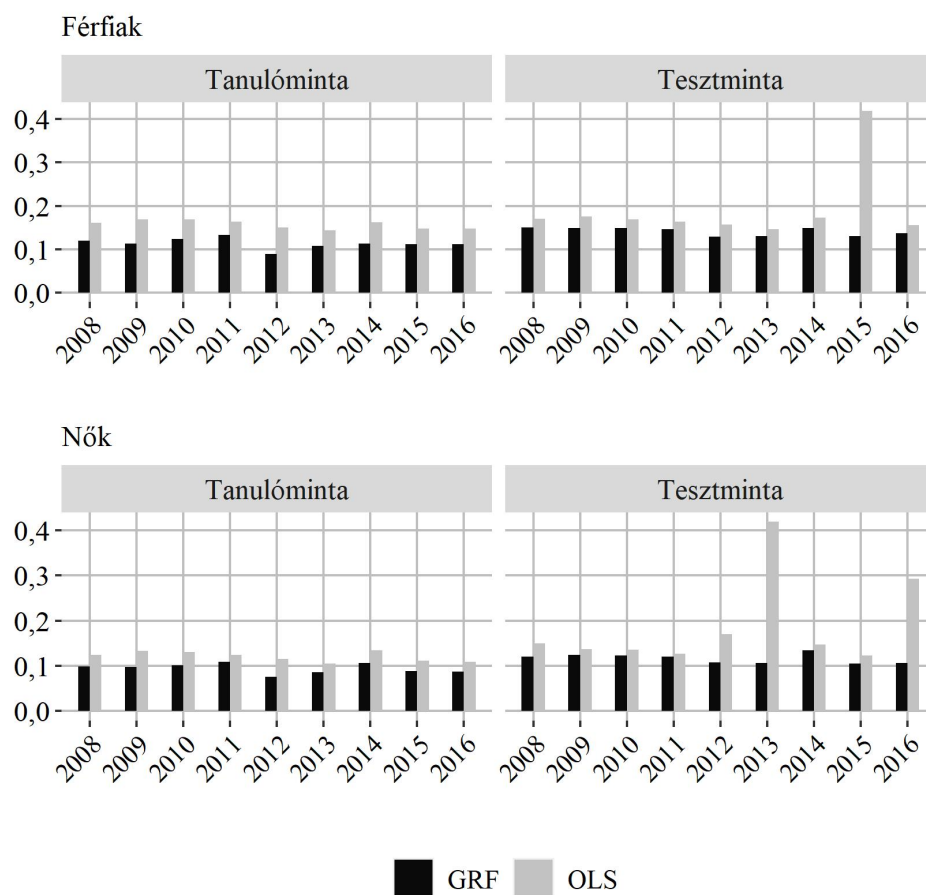
6.3. Eredmények

Az előrejelzési képesség összehasonlítására az átlagos négyzetes hibát használom. Az összevetést a 25. ábra mutatja és megerősíti az 5. fejezet eredményeit, miszerint a regressziós erdő jobban becsli a feltételes várható értéket, mint az OLS mind a tanuló-, mind a tesztmintán.²

A négyzetes hiba nagysága a férfiak esetében nagyobb, ami arra vezethető vissza, hogy az ő bérük jobban szóródik, mint a nőké. Néhány évben az OLS-beclés kiugróan magas értéket vesz fel a tesztmintán, ami arra utal, hogy lehetnek nagyhatású megfigyelések a tesztmintába. Ilyen lehet például, hogy az adott illető egy nagyon nagy – több ezres – vállalatnál dolgozik és tanulómintában nem szerepelt hasonló adat adott ágazatban vagy foglalkozásban. Akkor ennek következtében az adott ágazatban vagy foglalkozásban mért létszám hatása nagy lesz, ami miatt a becsült bér is kiugróan magas értéket vesz fel.

²Az összehasonlítás eredményei megegyeznek az 5. fejezetben látottakkal, azonban ebben a fejezetben az OLS magasabb rendű tagokat és kereszthatásokat is tartalmaz, ami igazolja, hogy az erdővel kapott jobb illeszkedés nem az 5. fejezetben meghatározott egyszerűbb függvényforma eredménye.

25. ábra. Átlagos négyzetes hiba az OLS és a regressziós erdő (GRF) esetében



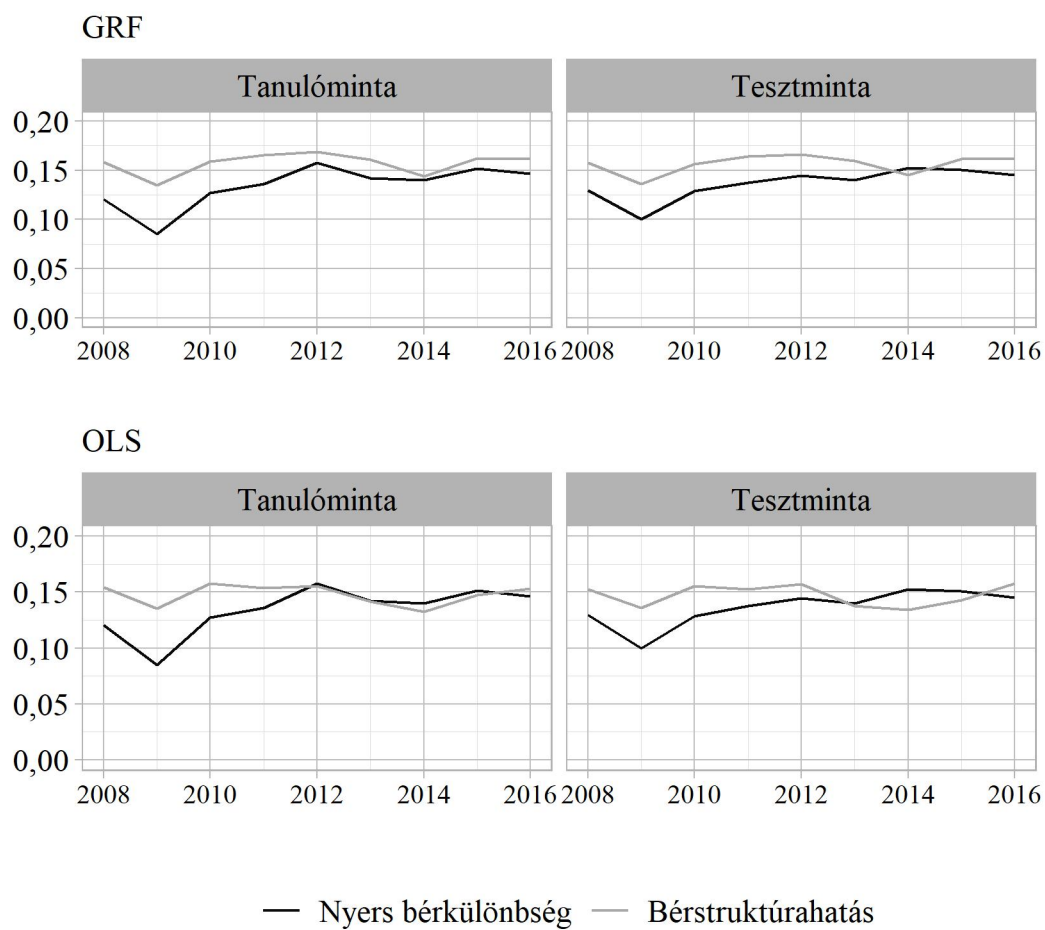
Forrás: Bértarifa, saját számítás

A következő alfejezetben a Blinder-Oaxaca felbontás eredményeit tárgyalom az OLS regresszió és a regressziós erdő esetében. Majd a felbontás alapján számolt heterogén bérstruktúrahatások időbeli alakulását vizsgálom az egyes változóknál. Az utolsó alfejezetben pedig elkülönítem a legnagyobb és legkisebb heterogén bérstruktúrahatással rendelkező csoportokat és ezek tulajdonságait elemzem.

6.3.1. Blinder-Oaxaca felbontás

A Blinder-Oaxaca felbontás eredményét mutatja a 20. táblázat, valamint a 26. és a 27. ábrák a tanulómintán.

26. ábra. Nyers bérkülönbségek és bérstruktúrahatások 2008 és 2016 között



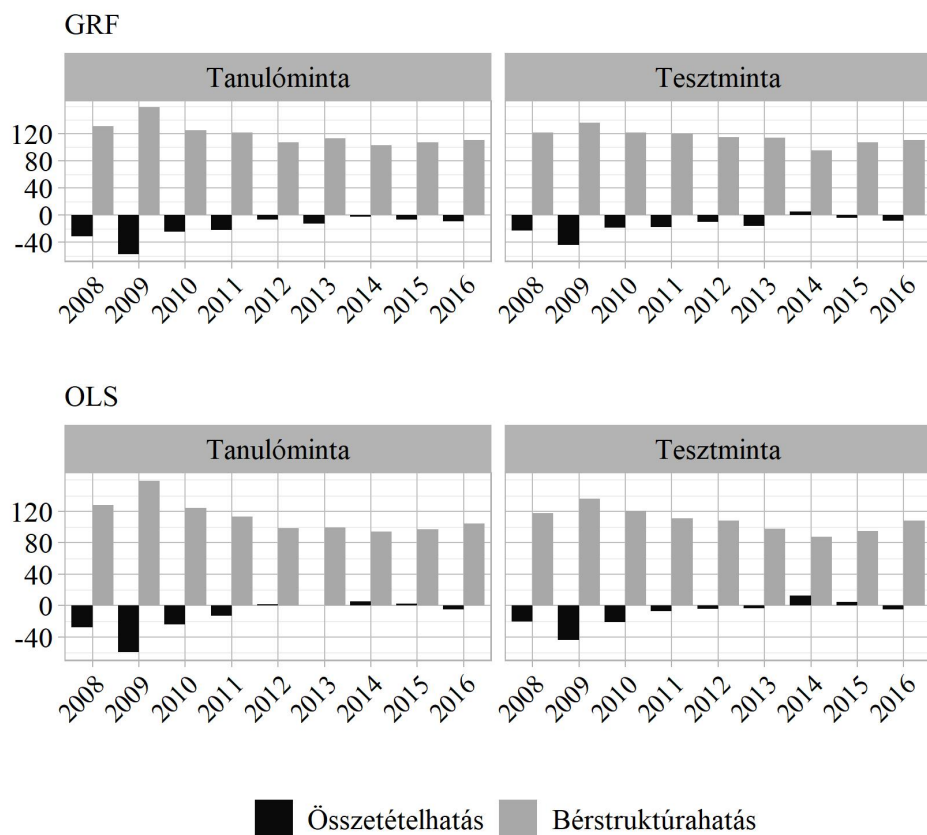
Forrás: Bértarifa, saját számítás

20. táblázat. A nyers bérkülönbség Blinder-Oaxaca felbontása regressziós erdővel (GRF) és OLS regresszióval (OLS) a tanulómintán

Év	Nyers különbség	GRF			OLS		
		Összetétel-hatás	Bérstruktúra-hatás	Torzítás	Összetétel-hatás	Bérstruktúra-hatás	Torzítás
2008	0,1203 (0,1103; 0,1294)	-0,0375 (-0,0447; -0,0300)	0,1580 (0,1556; 0,1603)	-0,0002	-0,0337 (-0,0422; -0,0255)	0,1540 (0,1515; 0,1566)	0
2009	0,0849 (0,0733; 0,0966)	-0,0491 (-0,0579; -0,0407)	0,1347 (0,1325; 0,1368)	-0,0007	-0,0502 (-0,0594; -0,0413)	0,1351 (0,1329; 0,1373)	0
2010	0,1271 (0,1144; 0,1396)	-0,0312 (-0,0401; -0,0212)	0,1589 (0,1563; 0,1614)	-0,0006	-0,0308 (-0,0404; -0,0207)	0,1578 (0,1557; 0,1602)	0
2011	0,136 (0,125; 0,1464)	-0,0296 (-0,0379; -0,0219)	0,1657 (0,1639; 0,1676)	-0,0001	-0,0177 (-0,0277; -0,0086)	0,1537 (0,1516; 0,1559)	0
2012	0,1577 (0,1472; 0,1675)	-0,0106 (-0,0188; -0,0026)	0,1684 (0,1664; 0,1702)	-0,0001	0,0022 (-0,0065; 0,0108)	0,1555 (0,1533; 0,1576)	0
2013	0,1422 (0,1318; 0,1521)	-0,0181 (-0,0262; -0,0111)	0,1611 (0,1593; 0,1629)	-0,0008	0,0004 (-0,0086; 0,0082)	0,1418 (0,1398; 0,1439)	0
2014	0,1402 (0,1307; 0,1516)	-0,0038 (-0,0104; 0,0042)	0,1438 (0,1421; 0,1456)	0,0002	0,0079 (0,0003; 0,0166)	0,1323 (0,1298; 0,1346)	0
2015	0,1516 (0,1411; 0,1603)	-0,0104 (-0,0182; -0,0035)	0,1623 (0,1607; 0,164)	-0,0003	0,004 (-0,0043; 0,0108)	0,1475 (0,1453; 0,1501)	0
2016	0,1463 (0,1348; 0,1571)	-0,014 (-0,0218; -0,0056)	0,1616 (0,1597; 0,1631)	-0,0013	-0,0069 (-0,0162; 0,0026)	0,1533 (0,1515; 0,1548)	0

Megjegyzés: A zárójelekben a 100 bootstrap mintán készített 95 százalékos konfidenciaintervallumok szerepelnek.

27. ábra. Blinder-Oaxaca felbontás (torzítás nélkül) 2008 és 2016 között
(nyers bérkülönbség százalékában)



Forrás: Bértarifa, saját számítás

2008 és 2016 között a nyers bérkülönbségek növekedése figyelhető meg, míg a bérek logaritmusában számolt bérstrukturahatás a 26. ábra alapján időben stabilnak bizonyult. A 27. ábrán látható, hogy a nyers bérkülönbség százalékában számolt negatív összetételhatás a vizsgált időszakban közeledett a nullához, sőt néhány évben még pozitív is volt. A negatív összetételhatás jelenlétét Magyarországon már korábbi tanulmányok is kimutatták. (Brainerd (2000), Leythienne–Ronkowski (2018), Cukrowska–Lovász (2014)) A mérséklődő összetételhatás következménye, hogy a nyers különbség százalékában mért bérstrukturahatás idővel csökkent.³

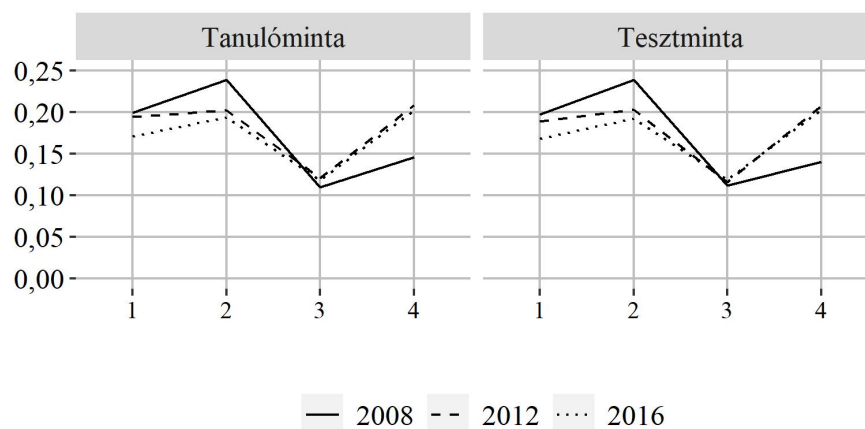
³Az Egyesült Államok adatait vizsgálva Blau–Kahn (2017) hasonló eredményekre jutottak. A szerzők azt találták, hogy a bérstrukturahatás különbségen belüli aránya a 80-as éveket követően csökkent, amit leginkább a nők nem megfigyelt jellemzőinek változásával, mintsem a diszkrimináció csökkenésével magyarázták.

A bérstruktúrahatás 2010 és 2016 között nagyobb volt az OLS-becslés esetében. A 20. táblázatban feltüntetett 95 százalékos konfidenciaintervallumok az erdő esetében általában szűkebbek, ami a becslési módszer megbízhatóságát igazolja.

6.3.2. Heterogén bérstruktúrahatások

Ebben az alfejezetben a heterogén bérstruktúrahatásokat vizsgálom az egyes magyarázóváltozóknál. Az időbeli trendek elemzéséhez három – a 2008-as, a 2012-es és a 2016-os – év adatait használom. A következő ábrákon néhány változó összevontabb formában szerepel, mint ahogyan az 1. táblázatban feltüntettem. A 28. ábra az egyes iskolai végzettségi kategóriákhoz tartozóan mutatja a heterogén bérstruktúrahatás átlagait.

28. ábra. Heterogén bérstruktúrahatás és iskolai végzettség



Megjegyzés: A becslésekben szereplő 9 kategóriát négybe vonom össze:

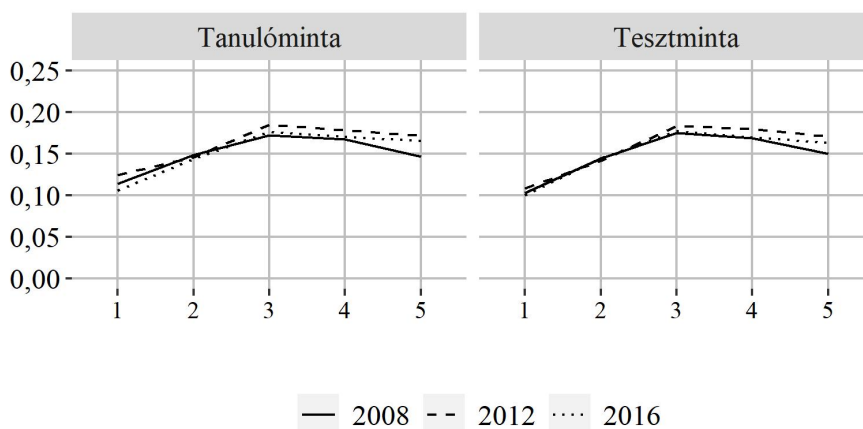
1. legfeljebb 8 osztály, 2. középiskolai végzettség érettségi nélkül,
3. középiskolai végzettség érettséggel, 4. felsőfokú végzettség.

Forrás: Bértarifa, saját számítás

A 28. ábra alapján a kapcsolat nem monoton a heterogén bérstruktúrahata-
 tás és az iskolai végzettség között. A legnagyobb átlagos bérstruktúrahata-
 tás az érettségivel nem rendelkező középiskolát végzetteknél figyelhető meg, azonban
 a hatás folyamatosan csökkent az évek alatt. 2012-re a felsőfokú végzettséggel
 rendelkezőknél a heterogén bérstruktúrahata-
 tás emelkedése látható. Összességé-
 ben az évek alatt a bérstruktúrahata-
 tások közeledését figyelhetjük meg.

A 29. ábra alapján az életkor és az átlagos heterogén bérstruktúrahata-
 tás között fordított U-alakú kapcsolat van. Kezdetben az életkor emelkedésével
 meredekebben emelkedik a bérstruktúrahata-
 tás, majd később enyhén csökken.
 2012-ben és 2016-ban a csökkenés kevésbé meredek, mint 2008-ban.

29. ábra. Heterogén bérstruktúrahata-
 tás és az életkor kategóriák

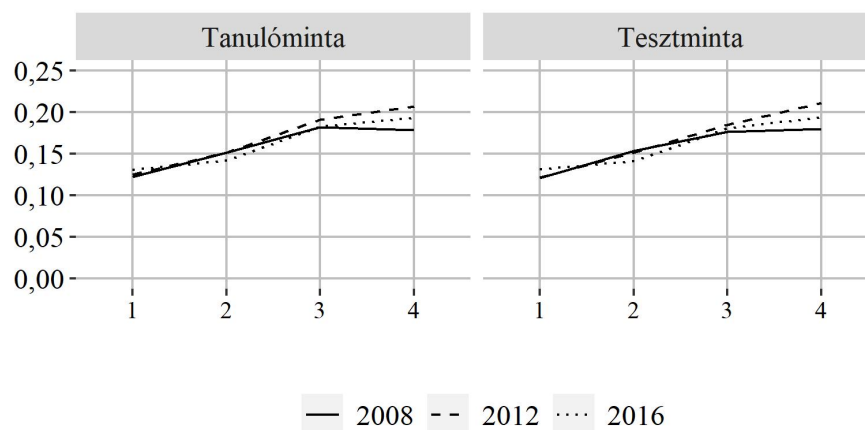


Megjegyzés: Korcsoportok: 1. fiatalok (18-24 év),
 2. fiatal felnőttek (25-34 év), 3. középkorúak (35-44 év),
 4. idősebb felnőttek (45-54 év), 5. idősek (55 év felett).

Forrás: Bértarifa, saját számítás

A 30. ábra szerint pozitív monoton kapcsolat van a szolgálati idő és a
 bérstruktúrahata-
 tás között. Az adott munkaadónál eltöltött idő növekedésével a
 bérstruktúrahata-
 tás is emelkedik. A 2008-as és a 2016-os eredmények meglehe-
 tősen együtt mozognak. 2012-ben a szolgálati idő növekedésével meredekebben
 nőtt a heterogén bérstruktúrahata-
 tás.

30. ábra. Heterogén bérstruktúrahatás és a szolgálati idő

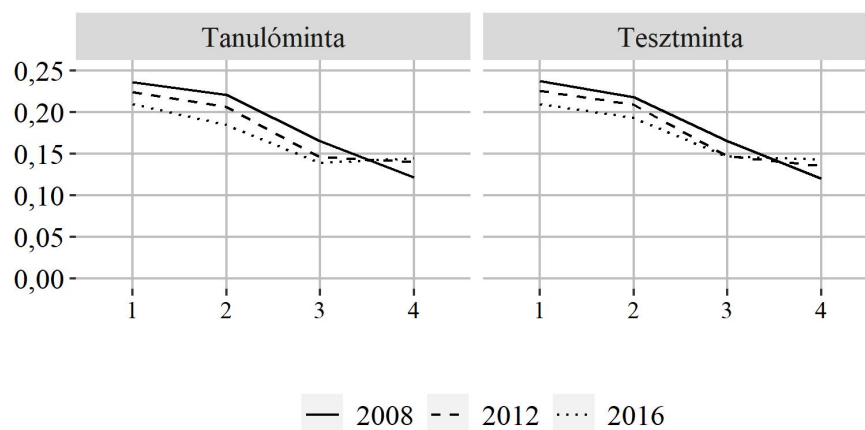


Megjegyzés: A szolgálati idő a kvartilisek alapján kerültek meghatározásra.

Forrás: Bértarifa, saját számítás

Minél alacsonyabb a külföldi részarány, annál kisebb az átlagos bérstruktúrahatás a 31. ábra alapján. Az évek során az 1-3. kategóriákban az átlagos bérstruktúrahatások mérséklődtek.

31. ábra. Heterogén bérstruktúrahatás és a külföldi részarány

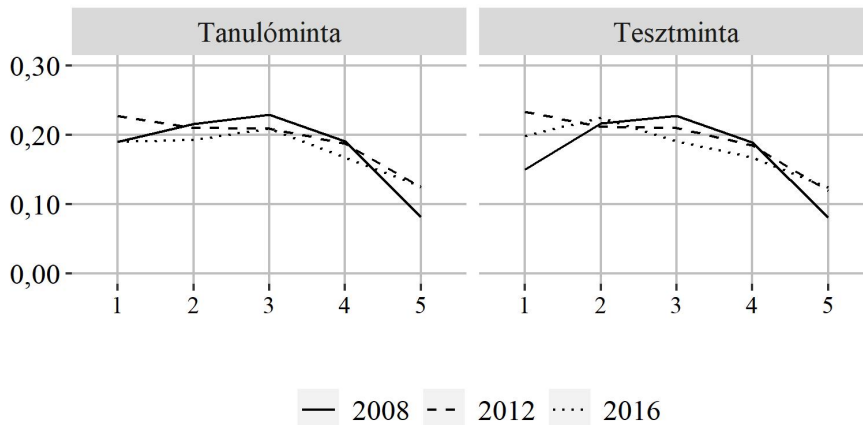


Megjegyzés: A külföldi részarány kategóriák megegyeznek az F.1. Függelékben lévőekkel: 1. 100% , 2. >50%, 3. <50%, 4. 0% tulajdon.

Forrás: Bértarifa, saját számítás

Az adott foglalkozáshoz tartozó női arány alapján, a 32. ábrán azokban a szakmákban, ahol nagyobb a nők aránya, ott az átlagos bérstruktúrahatások kisebbek. 2008-ban a kapcsolat a két tényező között négyzetes volt, később azonban inkább lineárisrá vált.

32. ábra. Egyéni bérstruktúrahatás és a női részarány



Megjegyzés: A női részarány a kétjegyű foglalkozáskódokhoz tartozóan mutatja a nők arányát a teljes Bértarifa-adatbázisban. Az ábrán ennek az arálynak a kvintiliseihez tartozó átlagok szerepelnek.

Forrás: Bértarifa, saját számítás

Összességében a 28-32. ábrák rámutattak arra, hogy milyen nem lineáris összefüggések léteznek a heterogén bérstruktúrahatások és az egyes változók között. Emellett néhány esetben láthatjuk, hogy a kapcsolat jellege is változott az évek során. A következőkben osztályozási elemzéssel vizsgálom a legextrémebb heterogén bérstruktúrahatással rendelkező csoportokat.

6.3.3. Extrém csoportok meghatározása klasszifikációs analízissel

A CLAN elemzésénél a heterogén bérstruktúrahatás eloszlás alsó és felső tíz százalékában lévőket tekintem az extrém csoportoknak. Az előbbi rendelkezik a legkisebb, az utóbbi a legnagyobb bérstruktúrahatással. Feltételezem, hogy az alacsonyabb bérstruktúrahatással rendelkező csoport (továbbiakban A-csoport) az, amelynél a nők közel azonos vagy magasabb béreket kapnának a becsült bérfüggvények alapján, mint az ugyanolyan munkaerőpiaci jellemzőkkel rendelkező férfiak. Az ő jelenlétük az átlagos bérstruktúrahatást lefelé húzza. Ezzel szemben a legmagasabb heterogén bérstruktúrahatással rendelkező extrém csoport (továbbiakban M-csoport) tagjai vannak a leginkább alulárázva – a becsült bérfüggvények alapján – az azonos adottságokkal rendelkező férfiakhoz képest, így az ő jelenlétük a munkaerőpiacon felfelé húzza az átlagot.

Chernozhukov és szerzőtársai (2018) módszertanával lehetséges pont- és intervallumbecslést is készíteni az extrém csoportok átlagos karakterisztikáira. Azonban esetemben előfordul, hogy bizonyos karakterisztikák nem jelentek meg a tanulómintában, így a modell tesztmintára való illesztését nem lehet elvégezni.⁴ Ezen technikai korlát miatt az átlagos karakterisztikákat leíró vektorokat csak a korábban is használt tanuló- és tesztmintákon számszerűsítem és nem csatolok konfidenciaintervallumokat a becslésekhez.

A tanulómintán és a két legextrémebb csoportban néhány változó kategóriánkénti megoszlását vizsgálom. A 21. táblázat tartalmazza az A- és az M-csoport változónkénti megoszlásainak különbségét a tanulómintán. A tanulómintára, illetve az A- és az M-csoportokra vonatkozó megoszlásokat, valamint ugyanezeket a megoszlásokat a tesztmintára vonatkozóan az F.9. Függelék tartalmazza.

⁴Leggyakrabban a mintafelbontásánál bizonyos FEOR kódba tartozók kimaradnak a tanulómintából és csak a tesztmintában jelennek meg.

21. táblázat. A legnagyobb és a legkisebb bérstruktúrahatással rendelkező csoportok (M- és A-csoportok) megoszlásainak különbsége (százalékpont)

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	0,4316	0,4060	0,4355	0,4023	0,4449	0,3953	0,3842	0,3919	0,3764
Életkor	1	-6	-4	-5	-6	-3	-4	-2	-9	-7
	2	-6	-13	-8	-9	-15	-12	-10	-21	-13
	3	8	8	10	7	14	10	6	16	15
	4	5	7	5	7	5	6	4	12	5
	5	-2	1	-1	0	-1	0	1	2	1
Szolgálati idő	1	-19	-14	-20	-14	-19	-14	-6	-29	-17
	2	-6	-9	-1	-6	-11	-4	-9	-13	-16
	3	11	12	15	6	10	8	3	19	13
	4	13	11	6	14	21	10	12	23	20
Iskolai végzettség	1	12	21	13	18	9	8	4	0	8
	2	39	28	31	29	21	40	21	13	21
	3	-43	-38	-45	-62	-46	-66	-42	-30	-55
	4	-8	-11	0	15	16	19	17	17	27
Régió	1	-41	-29	-35	-31	-29	-32	-22	-26	-29
	2	28	23	28	25	21	29	14	14	20
	3	13	5	7	7	7	2	8	12	10
Ágazat	1	5	4	9	-8	-4	1	5	2	5
	2	56	43	44	45	45	43	28	28	24
	3	-1	-2	-2	-1	2	0	-5	-3	-3
	4	-60	-45	-52	-36	-42	-44	-28	-27	-26
Vállalatméret	1	-43	-50	-40	-39	-25	-30	-28	-28	-28
	2	-14	0	-11	-8	-6	-6	-9	-7	-12
	3	20	19	20	24	19	19	20	19	13
	4	37	31	32	23	11	17	17	15	27
Külföldi részarány	1	49	36	48	43	32	21	37	20	30
	2	13	10	9	10	5	-1	-1	5	3
	3	-1	2	1	-4	-1	-5	-3	-1	-3
	4	-61	-49	-59	-49	-36	-16	-33	-24	-31

Megjegyzés: Korcsoportok: 1. fiatalok (18-24 év),

2. fiatal felnőttek (25-34 év), 3. középkorúak (35-44 év),

4. idősebb felnőttek (45-54 év), 5. idősek (55 év felett).

A szolgálati idő kategóriák a kvantilisok adott évben.

Iskolai végzettség kategóriák: 1. legfeljebb 8 osztály,

2. középiskolai végzettség érettségi nélkül,

3. középiskolai végzettség érettségivel, 4. felsőfokú végzettség.

Régió megfeleltethető a 3 NUTS 1 kategóriának:

1. Központi régió: Budapest és Pest megye,

2. Dunántúl: Közép-, Nyugat- és Dél-Dunántúl,

3. Alföld és Észak: Észak-Magyarország, Észak- és Dél-Alföld.

Ágazatok az export kitettség alapján rendezve: 1. kereskedelem: G,

2. erősen kereskedelmi: C, 3. valamennyire kereskedelmi: A, B, D, E,

4. szolgáltatások: F, H, I, J, K, L, M, N, P, Q, R, S.

Vállalatméret csoportok: 1. 20-49 fő, 2. 50-149 fő, 3. 150-499, 4. 500 fő felett.

Külföldi részarány: 1. 100% , 2. >50%, 3. <50%, 4. 0% külföldi tulajdon.

A különbség az M-csoport mínusz az A-csoport képlettel számolódik. A negatív számok jelentése, hogy adott kategóriában nagyobb arányban vannak az A-csoport tagjai, a pozitív számok pedig az M-csoport tagjainak nagyobb arányára utalnak.

Forrás: Bértarifa, saját számítás

A 21. táblázat alapján az A-csoporthoz tartozó személyek általában rövidebb szolgálati idővel rendelkeznek, mint az M-csoport tagjai. Az M-csoportban nagyobb a 35 év felettiek aránya (3-5. korcsoportok). Az A-csoportban dolgozók feltehetően inkább kisebb vállalatoknál dolgoznak. Az M-csoport tagjai leginkább a Dunántúlon a feldolgozóiparban tevékenykednek, míg az A-csoportban inkább olyan személyek vannak, akik a Közép-magyarországi régióban a szolgáltatászektorban dolgoznak. Azonban az ágazatnál megfigyelhető az M-csoport tagjainak a szolgáltatászektorba való áramlása. Emellett a régiókat tekintve nő ezen csoport aránya a központi régióban. Emellett az A-csoport tagjai valószínűleg inkább dolgoznak 0 százalékban külföldi tulajdonú (4. kategória) vállalatoknál, ahová megfigyelhető az M-csoport tagjainak áramlása. A tanulómintában és az M-csoportban a végzettség szerinti megoszlások hasonlóak, azonban az A-csoportnál kiemelkedik az érettségivel rendelkezők aránya. (Lásd az F.9.1-F.9.3. táblázatokban.)

6.4. Diskusszió

A regressziós erdő és a magasabb rendű tagokat is tartalmazó OLS teljesítmőképességének összehasonlítása alapján az erdő jobb előrejelzési képességgel rendelkezik, mint az OLS becslés. A tesztmintán mindkét módszerrel számolt átlagos négyzetes hibák emelkednek, azonban a regressziós erdő hibái így is alacsonyabbak. Az OLS esetében megfigyelhető volt a hiba kiugróan magas értéke a tesztmintában, ami óvatosságra int a használatával kapcsolatban.

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 csökkenő tendenciát mutatott. A nyers bérkülönbség időben emelkedett, a bérstruktúrahatás abszolút értéke 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.

Ahogy a 22. táblázat is mutatja: a férfiak jelentős része rendelkezik középfokú végzettséggel, de érettségivel nem, míg a nőknél az érettségivel rendelkezők vannak a legtöbben. A tendenciák az évek között hasonlóak mindkét nemnél: nő a felsőfokú végzettségűek aránya, mialatt az érettségivel nem rendelkezők aránya csökken. Azonban a 22. táblázat is sugallja, hogy ez az átmenet nagyon lassan megy végbe.

22. táblázat. Iskolai végzettség relatív gyakorisága nemek szerint (%)

Nem	Végzettség	2008	2009	2010	2011	2012	2013	2014	2015	2016
Férfiak	1	14	12	12	11	12	11	10	12	11
	2	41	39	37	38	38	40	37	35	35
	3	28	30	31	31	29	28	30	30	31
	4	18	20	20	21	22	22	24	23	23
Nők	1	16	14	14	13	13	13	13	12	12
	2	21	19	20	21	23	23	21	22	21
	3	45	46	45	45	42	40	42	43	42
	4	18	21	21	21	22	23	24	23	25

Megjegyzés: Végzettségi kategóriák:

1. legfeljebb 8 osztály, 2. középiskolai végzettség érettségi nélkül,
3. középiskolai végzettség érettségivel, 4. felsőfokú végzettség.

Az iskolai végzettség mellett a férfiak és a nők külföldi részarány szerinti megoszlását is érdemes megfigyelni! A nemek külföldi tulajdon szerinti relatív gyakoriságát mutatja az egyes kategóriákban a 23. táblázat.

23. táblázat. Külföldi részarány relatív gyakorisága nemek szerint (százalék)

Nem	Tulajdon	2008	2009	2010	2011	2012	2013	2014	2015	2016
Férfiak	1	19	16	20	22	22	24	23	20	21
	2	6	6	6	6	7	5	5	5	5
	3	3	2	2	2	2	3	2	2	2
	4	72	76	72	70	69	68	69	73	72
Nők	1	25	20	26	27	28	29	29	23	23
	2	7	7	6	7	7	6	5	7	5
	3	2	2	2	2	1	4	3	3	2
	4	66	71	65	64	64	62	63	67	70

Megjegyzés: A külföldi részarány kategóriák megegyeznek az F.1. Függelékben lévőkkel: 1. 100% , 2. >50%, 3. <50%, 4. 0% külföldi tulajdon.

A 23. táblázat alapján a férfiak aránya a 100 százalékban külföldi tulajdonban lévő vállalatoknál növekszik, míg a nőké inkább csökken. Ez az átmenet – mivel a külföldi vállalatok általában magasabb béreket fizetnek (Kőrösi (2006)) – az összetételhatás növekedése irányába hatott.

A heterogén bérstruktúrahatás vizsgálata rámutatott arra, 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) eredményeit⁵. Az átlagos bérstruktúrahatások és a külföldi részarány összevetésénél monoton kapcsolat látszik: a külföldi tulajdon csökkenésével az átlagos hatás is mérséklődik. Köllő (2000c) 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.

⁵Gábor (2008) eredményei szerepelnek a 2.1. alfejezetben.

A két csoport elkülönülését magyarázhatja a monopszonikus munkapiac elmélete. A elmélet klasszikus változata szerint a piac méretéhez képest nagyobb vállalatok piaci erőfölényre tesznek szert. Modern megközelítésben a vállalat erőfölénye a foglalkoztatási szerkezet és/vagy a munkaerőpiaci sűrűlódások létezésének a következménye. (Manning (2003)) Kapcsolódva a nemi bérkülönbségek irodalmához számos tanulmányban kimutatták, hogy a nemek közti eltérő ingázási, illetve keresési költségek, valamint nem pénzügyi juttatások fontossága szintén vezethet a vállalat erőfölényéhez. (Hirsch (2009), Hirsch–Schnabel (2010), Webber (2013), Heinze–Wolf (2010)) ⁶ Összességében a nemek közti eltérő vállalati rugalmasság létezését használhatja ki a munkáltató, amennyiben a versenypiaci helyzet – éppen a verseny hiánya – azt megengedi. Például, ha egy földrajzi területen kevés a munkaadó, a köztük lévő verseny kisebb, akkor ott a nőknek – akár pszichológiai, akár családi okok miatt – nehezebb lehet új állás találniuk, akkor előfordulhat, hogy a nők alacsonyabb munkabérért is hajlandóak dolgozni.

A CLAN elemzés eredményei alátámasztják a monopszonikus munkaerőpiac feltételezését. 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 bérstruktúráhatással 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.

⁶A 2.2. alfejezet tárgyalja a nem pénzügyi juttatások munkaerőpiaci döntésekben játszott szerepét a nőknél, illetve a két nem közti pszichológiai és kulturális eltéréseket Magyarországra vonatkozóan.

7. fejezet

Összefoglalás

A disszertációmban a magyarországi bérfüggvényt becslem meg tanuló algoritmusokkal. A kutatás kiindulópontja, hogy az eddigi szakirodalomban leginkább hagyományos ökonometriai módszerekkel – főként regressziókkal – becsülték a bérfüggvényeket, melyekben a felhasznált változók közti kapcsolatok korlátozott mértékben jelentek csak meg. Általában azok a kereszthatások szerepeltek a becslőfüggvényekben, melyek az adott kutatás/kérdésfelvetés szempontjából lényegesek voltak. Azonban a függvényforma korlátozása befolyásolhatja mind a paraméterek, mind a bérek becslését. Emiatt a disszertációmban a bérfüggvények formájára való korlátozást oldom fel tanuló algoritmusok használatával.

Az értekezés 2. fejezetében összefoglalom a Magyarországra vonatkozó bérezési kérdésekkel foglalkozó kutatási eredményeket. Ebben a fejezetben kitérek az egyéni kereseteket befolyásoló tényezőkre és a nemi bérkülönbségekkel foglalkozó kutatások eredményeire is. A 3. fejezetben a kutatásaimhoz használt gépi tanuló algoritmusokat tárgyalom. Az értekezés 4., 5. és 6. fejezetei tartalmazzák a saját kutatási eredményeket. A 4. fejezetben a célom a bérekre ható tényezők vizsgálata, melyhez döntési fát használok. Az 5. és a 6. fejezetben a lehető legjobb előrejelzéssel a férfiak és a nők béreit jelzem előre és Blinder-Oaxaca dekompozíció segítségével a nemi bérkülönbségeket elemzem.

A 4. fejezetben a 2016-os Bértarifa adatbázis felhasználásával feltételes következtetési fákkal becslem meg béreket. A fák, szemben más gépi tanulási módszerekkel mint a neurális hálók vagy a k-legközelebbi szomszéd osztályo-

zó algoritmusok, könnyen interpretálható eredményeket adnak és természetes módon használhatóak a változók közti interakciók megtalálására. Egy kicsi, egy közepes és egy nagy – az optimális – fa bemutatásával és a bennük található vágások elemzésével kerülök közelebb a bérekre ható tényezők, illetve a változók közti kapcsolatok megértéséhez. Emellett elemzem a legnagyobb fa alapján készült változófontossági mértékeket, valamint a változók szurrogátum jellegét. Ezek a vizsgálatok összességében igazolnak néhány kutatási kérdést és van, ahol további kérdéseket vetnek fel. Nem sikerül például igazolni az életkor bérmeghatározó szerepét. Helyette a szolgálati idő – mint a munkatapasztalat proxy változója – bizonyul fontosabb változónak. A változók közül kiemelkedik a foglalkozás és az iskolai végzettség bérmeghatározó szerepe. A szurrogátumok vizsgálata pedig arra hívja fel a figyelmet, hogy az ágazat bizonyos kontextusban – vágási pontokban – erősen összefügghet a többi változóval. Tehát nem mondható, hogy az ágazat erősen korrelál a többi változóval, azonban egyes csoportoknál ez mégis igaz.

Az 5. fejezetben a célom annak bemutatása, hogy a bérfüggvények becsléséhez használt módszertan milyen mértékben befolyásolja a Blinder-Oaxaca dekompozícióból kapott eredményeket. A nemi bérkülönbségek irodalmában elterjedt a Blinder-Oaxaca dekompozíció használata (Weichselbaumer–Winter-Ebmer (2005)), melyhez a bérfüggvényeket általában regressziókkal becslik meg. Fortin és szerzőtársai (2011) összefoglalóan mutatja be, hogy a munkaerőpiaci jellemzők figyelembevételével milyen feltételek mentén finomították a dekompozícióhoz használt regressziós formát. Alkalmaztak például kvantilis regressziókat (Mata–Machado (2005)) vagy Heckman-szelekciót (Reimers (1983), Neuman–Oaxaca (2004)). Ezek a finomítások általánosságban pontosították a Blinder-Oaxaca dekompozícióból származó eredményeket, azonban a függvényforma továbbra is – az előzetes feltételezések alapján – kötött maradt. Egyrészt ezt a feltételezést oldom fel regressziós erdők használatával. Másrészt általánosan igaz, hogy a tanuló algoritmusokból számolt előrejelzések pontosabbak, mint a hagyományos ökonometriai módszerekből származók.¹ Ebben a fejezetben a bérfüggvények becsléséhez egy OLS regressziót, valamint egy regressziós erdőt használok és az ezekből számolt bérbecsléseket használom fel a dekompozícióban, majd a kapott eredményeket hasonlítom össze.

¹Ezt az általános megfigyelést a saját kutatásom is igazolja.

Az eredmények rámutatnak, hogy a dekompozícióból számolt összetételhatás alacsonyabb az erdő esetében, ami miatt a bérstruktúrahatás viszont magasabb. Azonban az összetétel- és a bérstruktúrahatások így is hasonlóan alakulnak az OLS és az erdő esetében. Mindkét módszer esetében kimutatható, hogy a bérstruktúrahatás jelentősen magasabb, mint az összetételhatás.

Mivel az erdőkkel csak előrejelzések készíthetők a bérekre és nem elemezhetők az egyes változók hatásai a nyers bérkülönbségre – nem úgy, mint a regressziós becsléseknél –, emiatt a heterogén bérstruktúrahatások számszerűsítésével vizsgálom a változók egyedi hatásait. A heterogén bérstruktúrahatás megmutatja, hogy mekkora lenne a különbség egy nő bérében, ha a férfi, illetve a női bérfüggvénnyel áraznák a munkáját. Az OLS-sel és az erdővel számolt heterogén hatások közti közepes korreláció módszertani különbségre hívja fel a figyelmet. Vagyis bár az átlagok hasonlóan alakulnak, azonban az egyéni szinten számolt hatásokat jelentős mértékben befolyásolhatja a módszertan megválasztása. Emellett az OLS-sel és az erdővel számolt változókatégóriánkénti átlagos hatások nagy különbségeket mutatnak ott, ahol kevés számú megfigyelés esik az adott csoportba. Azonban nem lehet egy változót azonosítani, amely egyértelműen felelős lenne a két módszerrel számolt heterogén bérstruktúrahatások közti különbségért. Az eredményeim egybevágóak Weichselbaumer–Winter-Ebmer (2005) eredményeivel, akik kimutatták, hogy a Blinder-Oaxaca dekompozícióban használt módszerek hasonló eredményekre vezetnek, azonban a különböző becslési eljárások másként értékelik a változók nemi bérkülönbségre gyakorolt hatását.

A 6. fejezetben a heterogén bérstruktúrahatások időbeli alakulását vizsgálom és azonosítom azokat a csoportokat, amelyek a leginkább, illetve a legkevésbé járultak hozzá az átlagos bérstruktúrahatás fennmaradásához. Ebben a fejezetben regressziós erdőkkel becslem meg a férfi és a női bérfüggvényeket éves szinten a 2008 és 2016 közti időszakban. Emellett robusztusság vizsgálatként az erdők eredményeit összehasonlítom egy magasabbrendű tagokat és kereszthatásokat is tartalmazó OLS becslés eredményeivel. Az összehasonlítás igazolja, hogy az erdők pontosabb előrejelzéseket adnak, mint az OLS.

Az időbeli trendeket vizsgálva Magyarországon – szemben a nemzetközi tapasztalatokkal (Blau–Kahn (2017), OECD (2018)) – a nyers bérkülönbség időben emelkedett 2008 és 2016 között. A Blinder-Oaxaca felbontás alap-

ján a nyers különbségen belül a negatív összetételhatás idővel csökkent, míg a bérstruktúrahatás időben állandónak bizonyult. A csökkenő összetételhatás vélhetően a rendszerváltást követő munkaerőpiaci átrendeződés hosszútávú eredménye.

A heterogén bérstruktúrahatások elemzése rámutatott, hogy az életkor előrehaladtával a különbség nő a férfiak és a nők bérei között. A jelenség azzal magyarázható, hogy a férfiak bérfüggvénye magasabbról indul és meredekebben emelkedik, mint a nőké, ami megfelel Gábor (2008) eredményeinek. Míg a heterogén bérstruktúrahatás és az életkor között fordított U-alakú kapcsolat látszik, addig a szolgálati időnél lineáris trend található. Mindkét eredmény arra utal, hogy a különbség a férfiak és a nők bérei között kezdetben növekszik. A heterogén bérstruktúrahatás magasabb azoknál a vállalatoknál, amelyek külföldi tulajdonban vannak és a külföldi részarány csökkenésével mérséklődik a hatás. Emellett kimutatható, hogy azokban a szakmákban, ahol magasabb a nők aránya, ott kisebb a bérstruktúrahatás és a kapcsolat a két változó között az évek során egyre inkább lineárisává vált.

Klasszifikációs analízis (CLAN) segítségével vizsgálok azoknak a nőknek a csoportjait, akik a legnagyobb és a legkisebb heterogén bérstruktúrahatással rendelkeztek. Az elkülönített csoportok átlagos jellemzői rámutatnak, hogy azok a nők, akiknél a bérstruktúrahatás alacsony volt, a kisebb hazai szolgáltatászektorban tevékenykedő vállalatoknál dolgoztak a központi régióban. Ezzel szemben a legnagyobb bérstruktúrahatással rendelkező csoport tagjai a nagyobb, külföldi tulajdonú, feldolgozóipari vállalatoknál dolgoztak a Dunántúlon. A két csoport elkülönülése magyarázható a monoposzonikus munkapiac elméletével.

A monoposzonikus munkapiac elmélete alapján a nők vállalati rugalmassága kisebb, mint a férfiaké, ugyanis számukra a keresési, illetve az ingázási költségek nagyobbak lehetnek, valamint a nem pénzbeli juttatások is jobban hathatnak a munkaerőpiaci döntéseikre. (Hirsch (2009), Hirsch–Schnabel (2010), Webber (2013), Heinze–Wolf (2010)) A munkáltató pedig feltehetően kihasználhatja a nemek közti eltérő vállalati rugalmasságot. A CLAN eredményei alátámasztják a monoposzonikus munkapiac elméletét, ugyanis a központi régióban nagyobb lehet a munkáltatók közti verseny, illetve a nők könnyebben juthatnak új állásokhoz, ami miatt a munkaadók piaci erőfölénye kisebb lehet.

Tehát ezen a területen kevésbé valószínű, hogy a fennálló nemi bérkülönbség a vállalat monopszonikus erejéből származzon.

Összefoglalva az értekezés rámutat arra, hogy milyen új lehetőségek nyílnak meg a tanuló algoritmusok alkalmazása révén a munkaerőpiaci döntések vizsgálatára. A disszertációmban a fák vizsgálatára és – az erdők használatával – az előrejelzések pontosítására helyeztem a hangsúlyt. A gépi tanuló algoritmusok emellett használhatóak változószelekcióra. Tehát lehetőségeként felmerül, hogy a fák, illetve az erdők alapján határozzuk meg a felhasznált változók körét a hagyományos regressziós becslésekben. Az erre való igény pedig egyre inkább jelen van a nemi bérkülönbségek irodalmában, ugyanis egyre elterjedtebb a bérkülönbségek becslésénél a post-LASSO használata, mellyel a bérfüggvények becsléséhez felhasználandó változók előzetes kiválasztását végzik el. (Böheim–Stöllinger (2021), Bonaccolto–Töpfer–Briel (2022))

Irodalomjegyzék

- Archer, K. J. – Kimes, R. V. (2008): Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, DOI: <http://dx.doi.org/doi10.1016/j.csda.2007.08.015>.
- Arrosa, M. L. – Gandelman, N. (2016): Happiness decomposition: Female optimism. *Journal of Happiness Studies*, 17(2):731–756, DOI: <http://dx.doi.org/doi10.1007/s10902-015-9618-8>.
- Ashenfelter, O. – Krueger, A. (1994): Estimates of the economic return to schooling from a new sample of twins. *The American Economic Review*, 84(5):1157–1173.
- Athey, S. – Imbens, G. (2019): Machine learning methods economists should know about. *Annual Review of Economics*, 11(1):685–725, DOI: <http://dx.doi.org/doi10.1146/annurev-economics-080217-053433>.
- 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>.
- Athey, S. – Wager, S. (2019): Estimating treatment effects with causal forests: An application. *Observational Studies*, 5:37–51, DOI: <http://dx.doi.org/doi10.1353/obs.2019.0001>.
- 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.

- Becker, G. S. (1985): Human capital, effort, and the sexual division of labor. *Journal of Labor Economics*, 3(1):S33–S58, DOI: <http://dx.doi.org/doi10.1086/298075>.
- Becker, G. S. – Chiswick, B. R. (1966): Education and the distribution of earnings. *The American Economic Review*, 56(1/2):358–369.
- Böheim, R. – Stöllinger, P. (2021): Decomposition of the gender wage gap using the lasso estimator. *Applied Economics Letters*, 28(10):817–828, DOI: <http://dx.doi.org/doi10.1080/13504851.2020.1782332>.
- Blanchflower, D. G. – Oswald, A. J. (1989): The wage curve. Working Paper 3181, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w3181>.
- 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>.
- Blau, F. D. – Kahn, L. M. (2017): The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, 55(3):789–865, DOI: <http://dx.doi.org/doi10.1257/jel.20160995>.
- 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>.
- Bonaccolto-Töpfer, M. – Briel, S. (2022): The gender pay gap revisited: Does machine learning offer new insights? *Labour Economics*, 78:102223, DOI: <http://dx.doi.org/doihttps://doi.org/10.1016/j.labeco.2022.102223>.
- 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>.
- Breiman, L. (1998): Arcing classifiers. *Annals of Statistics*, 26(3):801–823, DOI: <http://dx.doi.org/doi10.1214/aos/1024691079>.
- Breiman, L. (2001): Random forests. *Machine Learning*, 45(1):5–32, DOI: <http://dx.doi.org/doi10.1023/a:1010933404324>.

- Breiman, L. – Friedman, J. H. – Olshen, R. A. – Stone, C. J. (1984): *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.
- Brenzel, H. – Laible, M.-C. (2016): Does personality matter?: the impact of the big five on the migrant and gender wage gaps. IAB Discussion Paper 201626, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].
- Brown, C. – Medoff, J. (1989): The employer size wage effect. NBER Working Paper 2870, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w2870>.
- Buchinsky, M. (1994): Changes in the u.s. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62(2):405–58, DOI: <http://dx.doi.org/doi10.2307/2951618>.
- Card, D. (1999): The causal effect of education on earnings. In Ashenfelter, O. – Card, D., szerk., *Handbook of Labour Economics*, pp. 1802–1863. Elsevier Science, North Holland, DOI: [http://dx.doi.org/doi10.1016/s1573-4463\(99\)03011-4](http://dx.doi.org/doi10.1016/s1573-4463(99)03011-4).
- 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 heterogeneous treatment effects in randomized experiments. NBER Working Paper 24678, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w24678>.
- Chiswick, B. R. (2003): Jacob mincer, experience and the distribution of earnings. *Review of Economics of the Household*, 1:343–361, DOI: http://dx.doi.org/doi10.1007/0-387-29175-x_10.
- Csillag, M. (2006): Női munka és nemek szerinti kereseti különbségek a késő szocializmustól napjainkig. In Fazekas, K. – Kézdi, G., szerk., *Munkaerőpiaci tükrök 2006*, pp. 100–105. MTA Közgazdaságtudományi Intézet.

- Cukrowska, E. – Lovász, A. (2014): Are children driving the gender wage gap? comparative evidence from poland and hungary. Budapest Working Papers on the Labour Market BWP - 2014/4, Hungarian Academy of Sciences, Institute of Economics, Centre for Economic and Regional Studies.
- Davis, J. M. – Heller, S. B. (2017): Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550, DOI: <http://dx.doi.org/doi10.1257/aer.p20171000>.
- Debeer, D. – Strobl, C. (2020): Conditional permutation importance revisited. *BMC Bioinformatics*, 21, DOI: <http://dx.doi.org/doi10.1186/s12859-020-03622-2>.
- Declercq, K. – Varga, J. (2018): Horizontális nemi szegregáció a felsőoktatásban - stem-jelentkezések. In Fazekas, K. – Szabó-Morvai, Á., szerk., *Munkaerőpiaci tükrök 2017*, pp. 114–121. MTA Közgazdaságtudományi Kutatóközpont.
- Dickens, W. T. – Katz, L. F. (1986): Interindustry wage differences and industry characteristics. NBER Working Paper 2014, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w2014>.
- Dickens, W. T. – Katz, L. F. (1987): Inter-industry wage differences and theories of wage determination. Working Paper 2271, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w2271>.
- DiPrete, T. A. – Jennings, J. L. (2012): Social and behavioral skills and the gender gap in early educational achievement. *Social Science Research*, 41(1):1 – 15, DOI: <http://dx.doi.org/doi10.1016/j.ssresearch.2011.09.001>.
- Dorie, V. – Hill, J. – Shalit, U. – Scott, M. – Cervone, D. (2019): Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, DOI: <http://dx.doi.org/doi10.1214/18-STS667>.
- Drucker, L. F. – Horn, D. – Kiss, H. J. (2018): Nemek közötti preferenciakülönbségek a szakirodalomban. In Fazekas, K. – Szabó-Morvai, Á., szerk., *Munkaerőpiaci tükrök 2017*, pp. 132–139. MTA Közgazdaságtudományi Kutatóközpont.

- Earle, J. S. – Telegdy, Á. (2012): A külföldi beruházások hatásai a munkavállalók béreire. In Fazekas, K. – Benczúr, P. – Telegdy, Á., szerk., *Munkaerőpiaci tükrök 2012*, pp. 215–230. MTA Közgazdaságtudományi Kutatóközpont.
- Elek, P. – Scharle, Á. – Szabó, B. – Szabó, P. A. (2009): A bérekhez kapcsolódó adóeltitkolás magyarországon. Közpénzügyi füzetek 23, Pénzügyminisztérium.
- Európai Bizottság (2015): Strategic engagement for gender equality 2016-2019.
- Falk, M. – Leoni, T. (2011): Estimating the wage curve with spatial effects and spline functions. *Economics Bulletin*, 31(1):591–604.
- 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: Munkapiac és regionalitás Magyarországon, Budapest*, pp. 47–74. MTA Közgazdaságtudományi Kutatóközpont.
- Fazekas, K. (2018): A nem kognitív készségek felértékelődésének hatása a nők munkaerőpiaci helyzetére. In Fazekas, K. – Szabó-Morvai, Á., szerk., *Munkaerőpiaci tükrök 2017*, pp. 122–129. MTA Közgazdaságtudományi Kutatóközpont.
- Feingold, A. (1994): Gender differences in personality: a meta-analysis. *Psychological bulletin*, 116(3):429–56, DOI: <http://dx.doi.org/doi10.1037/0033-2909.116.3.429>.
- Fernandez-Delgado, M. – Cernadas, E. – Barro, S. – Amorim, D. (2014): Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181.
- 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](http://dx.doi.org/doi10.1016/s0169-7218(11)00407-2).
- Galasi, P. (2000): *Női-férfi kereseti különbségek Magyarországon 1986–1996*. Országos Munkaügyi Kutató és Módszertani Központ, Budapest.

- Galasi, P. (2002a): A fizetett és nem fizetett munka értéke. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2002*, pp. 101–117. MTA Közgazdaságtudományi Kutatóközpont.
- Galasi, P. (2002b): A nő–férfi munkakínálatti különbségek tényezőkre bontása. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2002*, pp. 127–133. MTA Közgazdaságtudományi Kutatóközpont.
- Galasi, P. (2002c): Női–férfi munkaerőpiaci részvételi különbségek tényezői. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2002*, pp. 127–133. MTA Közgazdaságtudományi Kutatóközpont.
- Gábor, R. I. (2008): A hiányzó láncszem? Életpálya-keresetek és keresetengedezés. *Közgazdasági Szemle*, 55(december):1057–1074.
- Gronau, R. (1974): Wage comparisons—a selectivity bias. *Journal of Political Economy*, 82(6):1119–1143, DOI: <http://dx.doi.org/doi10.3386/w0013>.
- 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. (1976): The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In Berg, S. V., szerk., *Annals of Economic and Social Measurement*. 5(4), pp. 475–492. National Bureau of Economic Research, Inc.
- Heckman, J. J. (1979): Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, DOI: <http://dx.doi.org/doi10.2307/1912352>.
- Heckman, J. J. – Lochner, L. J. – Todd, P. E. (2003): Fifty years of mincer earnings regressions. NBER Working Paper 9732, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w9732>.
- Heckman, J. J. – Lochner, L. J. – Todd, P. E. (2005): Earnings functions, fates of return, and treatment effects: the mincer equation and beyond. NBER Working Paper 11544, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w11544>.

- Hegedüs, J. (2003): A lakáspiac és hatása a települések közötti lakásmobilitásra a kilencvenes években. In Fazekas, K., szerk., *Munkaerőpiaci tükrök 2003*, pp. 88–101. MTA Közgazdaságtudományi Kutatóközpont.
- 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, Boris and 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>.
- Hothorn, T. – Hornik, K. – Van de Wiel, M. – Zeileis, A. (2008): Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28, DOI: <http://dx.doi.org/doi10.18637/jss.v028.i08>.
- Hothorn, T. – Hornik, K. – Zeileis, A. (2006): Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, DOI: <http://dx.doi.org/doi10.1198/106186006X133933>.
- Hothorn, T. – Hornik, K. – Zeileis, A. (2015): ctree: Conditional inference trees. *The comprehensive R archive network*, 8.
- Hothorn, T. – Lausen, B. (2003): Double-bagging: combining classifiers by bootstrap aggregation. *Pattern Recognition*, 36(6):1303–1309, DOI: [http://dx.doi.org/doihttps://doi.org/10.1016/S0031-3203\(02\)00169-3](http://dx.doi.org/doihttps://doi.org/10.1016/S0031-3203(02)00169-3).
- Hothorn, T. – Seibold, H. – Zeileis, A. (2021): Package ‘partykit’: A toolkit for recursive partytioning. <http://partykit.r-forge.r-project.org/partykit/>. Hozzáférés: 2021. október 7.
- Hungerford, T. – Solon, G. (1987): Sheepskin effects in the returns to education. *The Review of Economics and Statistics*, 69(1):175–177, DOI: <http://dx.doi.org/doi10.2307/1937919>.

- Iammarino, S. – Santangelo, G. (2000): Foreign direct investment and regional attractiveness in the eu integration process: Some evidence for the italian regions. *European Urban and Regional Studies*, 7(1):5–18, DOI: <http://dx.doi.org/doi10.1177/096977640000700101>.
- 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>.
- Jacob, B. A. (2002): Where the boys aren't: non-cognitive skills, returns to school and the gender gap in higher education. *Economics of Education Review*, 21(6):589 – 598, DOI: [http://dx.doi.org/doi10.1016/S0272-7757\(01\)00051-6](http://dx.doi.org/doi10.1016/S0272-7757(01)00051-6).
- Jaeger, D. A. – Page, M. E. (1996): Degrees matter: New evidence on sheepskin effects in the returns to education. *The Review of Economics and Statistics*, 78(4):733–740, DOI: <http://dx.doi.org/doi10.2307/2109960>.
- 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>.
- Jolliffe, D. – Campos, N. F. (2005): Does market liberalisation reduce gender discrimination? econometric evidence from hungary, 1986–1998. *Labour Economics*, 12(1):1 – 22, DOI: <http://dx.doi.org/doi10.2139/ssrn.542702>.
- Kertesi, G. (2000): Ingázás a falusi magyarországon. egy megoldatlan probléma. *Közgazdasági Szemle*, 47(10):775–798.

- Kertesi, G. – Köllő, J. (1996): A bér alakulását meghatározó tényezők. In Halpern, L., szerk., *Bérköltség és versenyképesség*, pp. 74–76. MTA Közgazdaságtudományi Intézet, január.
- Kertesi, G. – Köllő, J. (2001): A gazdasági átalakulás két szakasza és az emberi tőke átértékelődése. a bérszerkezet átalakulása magyarországon 1986-1999 - iii. rész. *Közgazdasági Szemle*, 48(11):897–919.
- Kertesi, G. – Köllő, J. (2003): Ágazati bérkülönbségek magyarországon, ii. rész. járadékokon való osztozkodás koncentrált ágazatokban, szakszervezeti aktivitás jelenlétében. *Közgazdasági Szemle*, 50(12):1049–1074.
- Köllő, J. (2000a): Nemek szerinti kereseti különbségek. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2000*, pp. 74–76. MTA Közgazdaságtudományi Kutatóközpont.
- Köllő, J. (2000b): Regionális kereseti és bérköltség-különbségek. In Fazekas, K., szerk., *Munkaerőpiaci tükör 2000*, pp. 90–95. MTA Közgazdaságtudományi Kutatóközpont.
- Köllő, J. (2000c): 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.
- Koenker, R. – Bassett, G. (1978): Regression quantiles. *Econometrica*, 46(1):33–50, DOI: <http://dx.doi.org/doi10.2307/1913643>.
- Koenker, R. – Bassett, G. (1982): Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50(1):43–61, DOI: <http://dx.doi.org/doi10.2307/1912528>.
- Kreif, N. – DiazOrdaz, K. – Moreno-Serra, R. – Mirelman, A. – Hidayat, T. – Suhrcke, M. (2022): Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in indonesia. *Health Services and Outcomes Research Methodology*, 22:192–227, DOI: <http://dx.doi.org/doi10.1007/s10742-021-00259-3>.

- Kőrösi, G. (2006): Vállalatok közti bérkülönbségek dinamikája. In Fazekas, K. – Kézdi, G., szerk., *Munkaerőpiaci tükrök 2006*, pp. 48–59. MTA Közgazdaságtudományi Kutatóközpont.
- Krueger, A. B. – Summers, L. H. (1986a): Efficiency wage and the wage structure. NBER Working Paper 1952, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w1952>.
- Krueger, A. B. – Summers, L. H. (1986b): Reflections on the inter-industry wage structure. Working Paper 1968, National Bureau of Economic Research, DOI: <http://dx.doi.org/doi10.3386/w1968>.
- Központi Statisztikai Hivatal (2010): Módszertani útmutató a feor-08 foglalkozásainak besorolásához. https://www.ksh.hu/docs/szolgaltatasok/hun/feor08/pdf/feor08_modszertan.pdf. Hozzáférés: 2020. december 18.
- 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.
- Lewis, H. G. (1974): Comments on selectivity biases in wage comparisons. *Journal of Political Economy*, 82(6):1145–1155, DOI: <http://dx.doi.org/doi10.1086/260268>.
- Leythienne, D. – Ronkowski, P. (2018): A decomposition of the unadjusted gender pay gap using structure of earnings survey data. Statistical working paper, Eurostat.
- Loh, W.-Y. (2014): Fifty years of classification and regression on trees. *International Statistical Review*, 82(3):329–348, DOI: <http://dx.doi.org/doi10.1111/insr.12016>.
- Lovász, A. (2009): A verseny hatása a női-férfi bérkülönbségre Magyarországon 1986 és 2003 között. In Fazekas, K. – Lovász, A. – Telegdy, Á., szerk., *Munkaerőpiaci tükrök 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.

- Lovász, A. – Rigó, M. (2009): A nők és férfiak relatív termelékenységének és bérének becslése Magyarországon. In Fazekas, K. – Lovász, A. – Telegdy, Á., szerk., *Munkaerőpiaci tükrök 2009*, pp. 159–170. MTA Közgazdaságtudományi Intézet.
- Lovász, A. – Telegdy, Á. (2009): Munkapiaci diszkrimináció – típusok, mérési problémák, empirikus megoldások. In Fazekas, K. – Lovász, A. – Telegdy, Á., szerk., *Munkaerőpiaci tükrök 2009*, pp. 46–66. MTA Közgazdaságtudományi Intézet.
- Manning, A. (2003): The real thin theory: monopsony in modern labour markets. *Labour Economics*, 10(2):105 – 131, DOI: [http://dx.doi.org/doi10.1016/S0927-5371\(03\)00018-6](http://dx.doi.org/doi10.1016/S0927-5371(03)00018-6).
- Mata, J. – Machado, J. A. F. (2005): Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4):445–465, DOI: <http://dx.doi.org/doi10.1002/jae.788>.
- Miller, P. – Mulvey, C. – Martin, N. (1995): What do twins studies reveal about the economic returns to education? a comparison of Australian and U.S. findings. *The American Economic Review*, 85(3):586–599.
- Miller, S. (2020): Causal forest estimation of heterogeneous and time-varying environmental policy effects. *Journal of Environmental Economics and Management*, 103:102337, DOI: <http://dx.doi.org/doi10.1016/j.jeem.2020.102337>.
- 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>.
- Mincer, J. (1974): *Schooling, Experience, and Earnings*. National Bureau of Economic Research, Inc.
- Neuman, S. – Oaxaca, R. (2004): Wage decompositions with selectivity-corrected wage equations: A methodological note. *The Journal of Economic Inequality*, 2(1):3–10, DOI: <http://dx.doi.org/doi10.1023/b:joei.0000028395.38694.4b>.

- Newell, A. – Reilly, B. (2001): The gender pay gap in the transition from communism: some empirical evidence. *Economic Systems*, 25:287–304, DOI: [http://dx.doi.org/doi10.1016/s0939-3625\(01\)00028-0](http://dx.doi.org/doi10.1016/s0939-3625(01)00028-0).
- Nicodemus, K. K. – Malley, J. – Strobl, C. – Ziegler, A. (2010): The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, 11:110, DOI: <http://dx.doi.org/doi10.1186/1471-2105-11-110>.
- 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>.
- OECD (2018): *OECD Employment Outlook 2018*. OECD Publishing, Párizs, DOI: http://dx.doi.org/doi10.1787/empl_outlook-2018-en.
- O'Neill, E. – Weeks, M. (2018): Causal tree estimation of heterogeneous household response to time-of-use electricity pricing schemes. DOI: <http://dx.doi.org/doi10.48550/ARXIV.1810.09179>.
- Reilly, B. (2001): A nemek közötti bérkülönbségek elemzésének statisztikai módszerei. *Statisztikai Szemle*, 79(1):5–17.
- Reimers, C. W. (1983): Labor market discrimination against hispanic and black men. *The Review of Economics and Statistics*, 65(4):570–579, DOI: <http://dx.doi.org/doi10.2307/1935925>.
- Rigó, M. (2012): A szakszervezeti bérrés becslése magyarországon. In Fazekas, K. – Benczúr, P. – Telegdy, Á., szerk., *Munkaerőpiaci tükrök 2012*, pp. 200–214. MTA Közgazdaságtudományi Kutatóközpont.
- Rodríguez, J. – Muro, J. (2015): On the size of sheepskin effects: A meta-analysis. *Economics: The Open-Access, Open-Assessment E-Journal*, 9:1–18, DOI: <http://dx.doi.org/doi10.5018/economics-ejournal.ja.2015-37>.
- Shmueli, G. (2010): To explain or to predict? *Statistical Science*, 25(3):289–310, DOI: <http://dx.doi.org/doi10.1214/10-sts330>.
- Sinning, M. – Hahn, M. – Bauer, T. (2008): The blinder–oaxaca decomposition for nonlinear regression models. *Stata Journal: Promot-*

- ting communications on statistics and Stata*, 8(4):480–492, DOI: <http://dx.doi.org/doi10.1177/1536867x0800800402>.
- Strasser, H. – Weber, C. (1999): On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics*, 2:220–250.
- 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>.
- 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(25), DOI: <http://dx.doi.org/doi10.1186/1471-2105-8-25>.
- 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*, 14:323–48, DOI: <http://dx.doi.org/doi10.1037/a0016973>.
- Szabó, P. A. (2006): Regionális kereseti és bérkülönbségek. In Fazekas, K. – Kézdi, G., szerk., *Munkaerőpiaci tükör 2006*, pp. 70–79. MTA Közgazdaságtudományi Kutatóközpont.
- Szabó-Morvai, Á. (2018): Háztartási munkamegosztás. In Fazekas, K. – Szabó-Morvai, Á., szerk., *Munkaerőpiaci tükör 2017*, pp. 168–171. MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaságtudományi Intézet.
- Taber, D. R. – R., R. W. – Bleich, S. N. – C., W. Y. (2016): Deconstructing race and gender differences in adolescent obesity: Oaxaca-blinder decomposition. *Obesity*, 24(3):719–726, DOI: <http://dx.doi.org/doi10.1002/oby.21369>.
- 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. (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. KRTK-KTI műhelytanulmány 44, MTA Közgazdaság- és Regionális Tudományi Kutatóközpont.
- Telegdy, Á. (2013): A közszféra és a vállalatok bérei közötti áttérjedési hatás magyarországon. *Közgazdasági Szemle*, 60(5):555–578.
- Tibshirani, J. – Athey, S. – Friedberg, R. – Hadad, V. – Hirshberg, D. – Miner, L. – Sverdrup, E. – Wager, S. – Wright, M. (2022a): Package ‘grf’: Generalized random forests. <https://github.com/grf-labs/grf>. Hozzáférés: 2022. július 4.
- Tibshirani, J. – Athey, S. – Sverdrup, E. – Wager, S. (2022b): The grf algorithm. <https://grf-labs.github.io/grf/REFERENCE.html>. Hozzáférés: 2022. július 31.
- Varga, J. (2018): A külföldi beruházások hatásai a munkavállalók béreire. In Fazekas, K. – Szabó-Morvai, Á., szerk., *Munkaerőpiaci tükrök 2017*, pp. 45–53. MTA Közgazdaságtudományi 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>.
- Wager, S. – Athey, S. (2018): Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, DOI: <http://dx.doi.org/doi10.1080/01621459.2017.1319839>.
- 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>.

Wolf, E. (2002): Lower wage rates for fewer hours? a simultaneous wage-hours model for germany. *Labour Economics*, 9(5):643 – 663, DOI: [http://dx.doi.org/doi10.1016/s0927-5371\(02\)00055-6](http://dx.doi.org/doi10.1016/s0927-5371(02)00055-6).

Függelék

F.1. Függelék: Bértarifa adatbázisban szereplő változók leírása

A Bértarifa adatokban a következő aggregált kategóriák szerepelnek az egyes inputváltozóknál.

Az iskolai végzettség szerint a következő 9 kategória különböztethető meg:

1. Általános iskola 0-7 osztály
2. Általános iskola 8 osztály
3. Szakiskola
4. Szakmunkásképző iskola
5. Szakközépiskola
6. Gimnázium
7. Technikum
8. Főiskola, Alapfokozat
9. Egyetem, Mesterfokozat

A külföldi részarány 4 kategóriára oszlik:

1. 100% külföldi részarány
2. 50% feletti külföldi részarány
3. 50% alatti külföldi részarány
4. 0% külföldi részarány

Az állami és önkormányzati részarány 4 kategóriára oszlik:

1. 100% állami és önkormányzati részarány
2. 50% feletti állami és önkormányzati részarány
3. 50% alatti állami és önkormányzati részarány
4. 0% állami és önkormányzati részarány

A régió 7 kategóriára bomlik szét a NUTS2 felosztás szerint:

1. Közép-Magyarország
2. Közép-Dunántúl
3. Nyugat-Dunántúl
4. Dél-Dunántúl
5. Észak-Magyarország
6. Észak-Alföld
7. Dél-Alföld

F.1.1. táblázat. Foglalkozások – FEOR – lista

FEOR	Megnevezés
1	GAZDASÁGI, IGAZGATÁSI, ÉRDEK-KÉPVISELETI VEZETŐK, TÖRVÉNYHOZÓK
11	Törvényhozók, igazgatási, érdek-képviselési vezetők
12	Gazdasági, költségvetési szervezetek vezetői
13	Termelési és szolgáltatást nyújtó egységek vezetői
14	Gazdasági tevékenységet segítő egységek vezetői
2	FELSŐFOKÚ KÉPZETTSÉG ÖNÁLLÓ ALKALMAZÁSÁT IGÉNYLŐ FOGLALKOZÁSOK
21	Műszaki, informatikai és természettudományi foglalkozások
22	Egészségügyi foglalkozások (felsőfokú képzettséghez kapcsolódó)
23	Szociális szolgáltatási foglalkozások (felsőfokú képzettséghez kapcsolódó)
24	Oktatók, pedagógusok
25	Gazdálkodási jellegű foglalkozások
26	Jogi és társadalomtudományi foglalkozások
27	Kulturális, sport-, művészeti és vallási foglalkozások (felsőfokú képzettséghez kapcsolódó)
29	Egyéb magasan képzett ügyintézők
3	EGYÉB FELSŐFOKÚ VAGY KÖZÉPFOKÚ KÉPZETTSÉGET IGÉNYLŐ FOGLALKOZÁSOK
31	Technikusok és hasonló műszaki foglalkozások
32	Szakmai irányítók, felügyelők
33	Egészségügyi foglalkozások
34	Oktatási asszisztensek
35	Szociális gondozási és munkaerő-piaci szolgáltatási foglalkozások
36	Üzleti jellegű szolgáltatások ügyintézői, hatósági ügyintézők
37	Művészeti, kulturális, sport- és vallási foglalkozások
39	Egyéb ügyintézők

FEOR	Megnevezés
4	IRODAI ÉS ÜGYVITELI (ÜGYFÉLKAPCSOLATI) FOGLALKOZÁSOK
41	Irodai, ügyviteli foglalkozások
42	Ügyfélkapcsolati foglalkozások
5	KERESKEDELMI ÉS SZOLGÁLTATÁSI FOGLALKOZÁSOK
51	Kereskedelmi és vendéglátó-ipari foglalkozások
52	Szolgáltatási foglalkozások
6	MEZŐGAZDASÁGI ÉS ERDŐGAZDÁLKODÁSI FOGLALKOZÁSOK
61	Mezőgazdasági foglalkozások
62	Erdőgazdálkodási, vadgazdálkodási és halászati foglalkozások
7	IPARI ÉS ÉPÍTŐIPARI FOGLALKOZÁSOK
71	Élelmiszer-ipari foglalkozások
72	Könnyűipari foglalkozások
73	Fém- és villamosipari foglalkozások
74	Kézműipari foglalkozások
75	Építőipari foglalkozások
79	Egyéb ipari és építőipari foglalkozások
8	GÉPKEZELŐK, ÖSSZESZERELŐK, JÁRMŰVEZETŐK
81	Feldolgozóipari gépek kezelői
82	Összeszerelők
83	Helyhez kötött gépek kezelői
84	Járművezetők és mobil gépek kezelői
9	SZAKKÉPZETTSÉGET NEM IGÉNYLŐ (EGYSZERŰ) FOGLALKOZÁSOK
91	Takarítók és hasonló jellegű egyszerű foglalkozások
92	Egyszerű szolgáltatási, szállítási és hasonló foglalkozások
93	Egyszerű ipari, építőipari, mezőgazdasági foglalkozások
0	FEGYVERES SZERVEK FOGLALKOZÁSAI
01	Fegyveres szervek felsőfokú képesítést igénylő foglalkozásai
02	Fegyveres szervek középfokú képesítést igénylő foglalkozásai
03	Fegyveres szervek középfokú képesítést nem igénylő foglalkozásai

Forrás: Központi Statisztikai Hivatal

F.1.2. táblázat. Nemzetgazdasági ágak – TEÁOR – listája

TEÁOR	Megnevezés
A	MEZŐGAZDASÁG, ERDŐGAZDÁLKODÁS, HALÁSZAT
B	BÁNYÁSZAT, KŐFEJTÉS
C	FELDOLGOZÓIPAR
D	VILLAMOSENERGIA-, GÁZ-, GŐZELLÁTÁS, LÉGKONDI- CIONÁLÁS
E	VÍZELLÁTÁS; SZENNYVÍZ GYŰJTÉSE, KEZELÉSE, HUL- LADÉKGAZDÁLKODÁS, SZENNYEZŐDÉSMENTESÍTÉS
F	ÉPÍTŐIPAR
G	KERESKEDELEM, GÁPJÁRMŰJAVÍTÁS
H	SZÁLLÍTÁS, RAKTÁROZÁS
I	SZÁLLÁSHELY-SZOLGÁLTATÁS, VENDÉGLÁTÁS
J	INFORMÁCIÓ, KOMMUNIKÁCIÓ
K	PÉNZÜGYI, BIZTOSÍTÁSI TEVÉKENYSÉG
L	INGATLANÜGYLETEK
M	SZAKMAI, TUDOMÁNYOS, MŰSZAKI TEVÉKENYSÉG
N	ADMINISZTRATÍV ÉS SZOLGÁLTATÁST TÁMOGATÓ TE- VÉKENYSÉG
O	KÖZIGAZGATÁS, VÉDELEM; KÖTELEZŐ TÁRSADALOM- BIZTOSÍTÁS
P	OKTATÁS
Q	HUMÁN-EGÉSZSÉGÜGYI, SZOCIÁLIS ELLÁTÁS
R	MŰVÉSZET, SZÓRAKOZTATÁS, SZABADIDŐ
S	EGYÉB SZOLGÁLTATÁS
T	HÁZTARTÁS MUNKAADÓI TEVÉKENYSÉGE; TERMÉK ELŐÁLLÍTÁSA, SZOLGÁLTATÁS VÉGZÉSE SAJÁT FO- GYASZTÁSRA
U	TERÜLETEN KÍVÜLI SZERVEZET

Forrás: Központi Statisztikai Hivatal

F.2. Függelék: Foglalkozás besorolás és a hozzájuk kapcsolódó végzettségi szintek

F.2.1. táblázat. Foglalkozásokhoz – FEOR-hoz kötött – elvárt iskolai végzettségi szintek

FEOR kód Foglalkozás megnevezése	Iskolai végzettségi szint
1 – Gazdasági, igazgatási, érdekképviseleti vezetők, törvényhozók	Egyetemi, főiskolai (vagy azzal egyenértékű) végzettség, illetve egyes esetekben felsőfokú vagy érettségire épülő szakképzésben szerzett végzettség, gimnáziumban, szakközépiskolában, technikumban szerzett képesítés
2 – Felsőfokú végzettség önálló alkalmazását igénylő foglalkozások	Egyetemi, főiskolai (vagy azzal egyenértékű) végzettség
3 – Egyéb, felsőfokú, vagy középfokú képzettséget igénylő foglalkozások	Felsőfokú vagy érettségire épülő szakképzésben szerzett végzettség, gimnáziumban, szakközépiskolában, technikumban szerzett képesítés.
4 – Irodai és ügyviteli (ügyfélkapcsolati) foglalkozások	Érettségire épülő szakképzésben szerzett végzettség, gimnáziumban, szakközépiskolában szerzett képesítés, szakiskolai (szakmunkásképző iskolai) szakképesítés alapfokú (általános iskolai) végzettség
5 – Kereskedelmi és szolgáltatási foglalkozások	
6 – Mezőgazdasági és erdőgazdálkodási foglalkozások	
7 – Ipari és építőipari foglalkozások	
8 – Gépkezelők, összeszerelők, járművezetők	Nem igényel szakmai végzettséget
9 – Szakképzettséget nem igénylő (egyszerű) foglalkozások	
0 – Fegyveres szervek foglalkozásai	Az egyes csoportok különböző készség szintekbe tartoznak.

Forrás: Központi Statisztikai Hivatal (2010), p. 9

F.3. Függelék: Adatbázist leíró statisztikák

F.3.1. táblázat. Nemek szerinti megoszlás

	Nők	Férfiak
Megfigyelések száma	43470	69531
Megoszlás (%)	38.47	61.53

Forrás: Bértarifa, saját számítás

F.3.2. táblázat. Iskolai végzettség szerinti megoszlás

	1	2	3	4	5	6	7	8	9
Megfigyelések száma	107	12894	4376	28853	22465	12586	4591	15717	11412
Megoszlás (%)	0.09	11.41	3.87	25.53	19.88	11.14	4.06	13.91	10.10

Megjegyzés: Az iskolai végzettség kategóriái: 1: Általános iskola 0-7 osztály,
2: Általános iskola 8 osztály, 3: Szakiskola, 4: Szakmunkásképző iskola,
5: Szakközépiskola, 6: Gimnázium, 7: Technikum, 8: Főiskola, 9: Egyetem.

Forrás: Bértarifa, saját számítás

F.3.3. táblázat. Foglalkozás szerinti megoszlás

	1	2	3	4	5	6	7	8	9
Megfigyelések száma	10330	12696	20517	9686	9480	991	15481	20756	13064
Megoszlás (%)	9.14	11.24	18.16	8.57	8.39	0.88	13.70	18.37	11.56

Megjegyzés: Az egyjegyű FEOR kódok megnevezései megtalálhatóak az F.1. Függelékben.

Forrás: Bértarifa, saját számítás

F.3.4. táblázat. Régió szerinti megoszlás

	1	2	3	4	5	6	7
Megfigyelések száma	46410	12396	11788	8075	8688	13063	12581
Megoszlás (%)	41.07	10.97	10.43	7.15	7.69	11.56	11.13

Megjegyzés: Régió kategóriák: 1: Közép-Magyarország, 2: Közép-Dunántúl, 3: Nyugat-Dunántúl, 4: Dél-Dunántúl, 5: Észak-Magyarország, 6: Észak-Alföld, 7: Dél-Alföld.

Forrás: Bértarifa, saját számítás

F.3.5. táblázat. Településtípus szerinti megoszlás

	1	2	3
Megfigyelések száma	10260	64780	37961
Megoszlás (%)	9.08	57.33	33.59

Megjegyzés: Településtípus kategóriák: 1: Budapest, 2: város, 3: egyéb.

Forrás: Bértarifa, saját számítás

F.3.6. táblázat. Ágazat szerinti megoszlás

	A	B	C	D	E	F	G	H	I
Megfigyelések száma	4384	357	39096	2078	2553	5710	20106	9991	2715
Megoszlás (%)	3.88	0.32	34.6	1.84	2.26	5.05	17.79	8.84	2.4

Megjegyzés: Az ágazatok megnevezései megtalálhatóak az F.1. Függelékben.

Forrás: Bértarifa, saját számítás

Ágazat szerinti megoszlás (folyt.)

	J	K	L	M	N	P	Q	R	S
Megfigyelések száma	5820	6007	1167	4268	6073	186	742	933	815
Megoszlás (%)	5.15	5.32	1.03	3.78	5.37	0.16	0.66	0.83	0.72

Megjegyzés: Az ágazatok megnevezései megtalálhatóak az F.1. Függelékben.

Forrás: Bértarifa, saját számítás

F.3.7. táblázat. Kollektív szerződések szerinti megoszlás

Szerződés típusa		Nincs	Van
Vállalati	Megfigyelések száma	92506	20495
	Megoszlás (%)	81.86	18.14
Ágazati	Megfigyelések száma	111334	1667
	Megoszlás (%)	98.52	1.48
Több munkáltatós	Megfigyelések száma	92506	20495
	Megoszlás (%)	81.86	18.14

Forrás: Bértarifa, saját számítás

F.3.8. táblázat. Tulajdoni részarány szerinti megoszlás

Tulajdon		100%	>50%	<50%	0
Külföldi	Megfigyelések száma	24542	5752	2115	80592
	Megoszlás (%)	21.72	5.09	1.87	71.32
Állami	Megfigyelések száma	9542	1069	2940	99450
	Megoszlás (%)	8.44	0.95	2.60	88.01

Forrás: Bértarifa, saját számítás

F.3.9. táblázat. Folytonos változók leíró statisztikái

	Átlag	Szórás
Életkor	41.49	11.11
Szolgálati idő	87.10	97.10
Vállalati létszám	1352.65	3400.78

Forrás: Bértarifa, saját számítás

F.4. Függelék: Közepes fa

F.4.1. táblázat. Közepes fa

Levél szám	Megfigyelések száma	Vágási pont	Átlag
1)	113001	root	300414
2)	23026	isco_1 in 1, 2	603294
4)	5240	educ in 2, 3, 4, 5, 6, 7	366574
8)	1151	foreignown in 1, 2, 3	539428
16)	349	gender in 0	449978
32)*	238	size <= 1458	404446
33)*	111	size > 1458	547562
17)	802	gender in 1	578383
34)*	726	region in 1, 2, 3, 5, 7	595729
35)*	76	region in 4, 6	411861
9)	4089	foreignown in 4	318626
18)	2054	sector_1 in 1-6, 8, 10-13, 19	383904
36)*	1052	tenure <= 84	342416
37)*	1002	tenure > 84	427144
19)	2035	sector_1 in 7, 9, 14-18	255859
38)*	792	region in 1, 2	326806
39)*	1243	region in 3, 4, 5, 6, 7	213913
5)	17786	educ in 8, 9	671817
10)	6089	isco_1 in 1	923651
20)	2253	foreignown in 1, 2, 3	1136251
40)*	1077	educ in 8	972587
41)*	1176	educ in 9	1288376
21)	3836	foreignown in 4	783952
42)*	1466	sector_1 in 1, 2, 6, 7, 9, 14, 17-19	590957
43)*	2370	sector_1 in 3-5, 8, 10-13, 16	902336
11)	11697	isco_1 in 2	555756
22)	932	sector_1 in 1, 2, 5, 6, 16-18	390329
44)*	578	size <= 206	336185
45)*	354	size > 206	476724
23)	10765	sector_1 in 3, 4, 7, 8, 10-14, 19	569419
46)*	2468	age <= 30	419767
47)*	8297	age > 30	615312

Megjegyzés: Az első oszlopban a * jelöli a fa végsőleveleit. Az átlagokat, akárcsak a 8. ábrán forintban tüntettem fel.

Forrás: Bértarifa, saját számítás

Közepes fa (folyt.)

Levél szám	Megfigyelések száma	Vágási pont	Átlag
3)	89975	isco_1 in 3, 4, 5, 6, 7, 8, 9	234168
6)	30203	isco_1 in 3, 4	304639
12)	21672	educ in 2, 3, 4, 5, 6, 7	271846
24)	6319	foreignown in 1, 2, 3	327579
48)*	3492	gender in 0	297719
49)*	2827	gender in 1	365109
25)	15353	foreignown in 4	248737
50)*	6560	sector_1 in 1, 2, 5-7, 9, 12, 14, 17, 19	218971
51)*	8793	sector_1 in 3, 4, 8, 10, 11, 13, 16, 18	271479
13)	8531	educ in 8, 9	391377
26)	5848	region in 1, 2	426963
52)*	3020	tenure <= 41	362600
53)*	2828	tenure > 41	497706
27)	2683	region in 3, 4, 5, 6, 7	315124
54)*	890	foreignown in 1, 2, 3	379856
55)*	1793	foreignown in 4	282952
7)	59772	isco_1 in 5, 6, 7, 8, 9	200419
14)	23535	isco_1 in 5, 6, 9	169396
28)	23126	educ in 1, 2, 3, 4, 5, 6, 7	167658
56)*	22159	sector_1 in 1-7, 9-19	165817
57)*	967	sector_1 in 8	210589
29)	409	educ in 8, 9	279760
58)*	311	isco_1 in 5, 6	312104
59)*	98	isco_1 in 9	184552
15)	36237	isco_1 in 7, 8	221009
30)	23168	tenure <= 95	204376
60)*	6375	foreignown in 1, 2, 3	235110
61)*	16793	foreignown in 4	192697
31)	13069	tenure > 95	250998
62)*	10409	sector_1 in 1-3, 5-7, 9, 11-19	240695
63)*	2660	sector_1 in 4, 8, 10	291151

Megjegyzés: Az első oszlopban a * jelöli a fa végsőleveleit. Az átlagokat, akárcsak a 8. ábrán, forintban tüntettem fel.

Forrás: Bértarifa, saját számítás

F.5. Függelék: Szurrogátumok a változóknál

Az egyes változókhoz kapcsolódó legfontosabb szurrogátumokat tartalmazza az F.5.1. táblázat. Az F.5.1. táblázat azt mutatja, hogy a sorokban szereplő változók alapján képzett vágásoknál az oszlopban szereplő változók hányadik helyen szerepelnek szurrogátumként. Az első helyen szereplő mutatja, hogy a CTree alapján eszerint a változó szerint kapnánk a leginkább hasonló felosztást, ha az eredeti változó nem szerepelne a változók között. Az F.5.2. táblázat összevontan tartalmazza, hogy a sorokban szereplő vágóváltozóknak az oszlopban szereplő változók az esetek hány százalékában szerepeltek szurrogátumként. Tulajdonképpen ez a táblázat az F.5.1. táblázat összevonása.

F.5.1. táblázat. A legfontosabb szurrogátumok az optimális fában

Változók	Foglalkozás						Iskolai végzettség						Ágazat						Külföldi részarány						Régió					
Sorszám	1	2	3	4	5		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5		1	2	3	4	5	
Foglalkozás							9	13	9	9	3		26	12	7	1	1		4	3	3	5			5	6	10	5	8	
Iskolai végzettség	11	7	8	5	6								27	16	14	5	6		4	3	6	3			15	10	11	9	6	
Ágazat	13	12	11	8	11		15	18	23	24	19								15	7	10	4	6		18	31	28	33	25	
Külföldi részarány	2	1	1	1	5		3	2	2	3	3		11	13	7	4	4								3	8	8	7	5	
Régió	7	8	11	4	8		9	16	20	28	23		45	50	13	16	18		4	9	5	4	7							
Szolgálati idő	3	7	8	6	8		10	18	17	15	15		9	31	10	12	10		1	3	4	6	2		9	7	12	12	13	
Vállalati létszám	5	13	7	14	9		14	15	18	26	15		33	27	23	14	12		4	5	6	9	2		33	24	15	14	16	
Nem	7	1	2	4	2		12	12	4		5		22	11	9	5	3		1	2	3		5			5	9	10	4	
Településtípus	2	4	3	7	6		3	8	19	12	7		24	27	6	8	6		2	6	4	6	3		44	13	12	5	5	
Vállalati szerződés				1	5			1	8	4	6		12	13	6	4	3		2	3	2	1	1		5	8	3	5	8	
Életkor		7	1	1			7	7	5	5	4		4	5	7	8	6		1	2	3	2	1		2	6	3	5	4	
Állami részarány	1		2	1				2	4	9	8		25	9	2	2			3	3		3			3	10	8	4	4	
Ágazati szerződés		2	2		1		1	2	2	3	8		8	6	4	5	3								6	8	5	4	3	
Több munkáltatós sz.		1	1		1		1	1	1	1			4	1	1		1		2	1					1	4	1		2	

Forrás: Bértarifa, saját számítás

A legfontosabb szurogátumok az optimális fában (folyt.)

Változók	Szolgálati idő					Vállalati létszám					Nem					Településtípus					Vállalati szerződés				
Sorszám	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Foglalkozás	4	5	9	6	11	10	8	4	11	4	8	2	4	2	1	3	5	5	10	7	2		4	3	6
Iskolai végzettség	1	11	13	7	13	3	10	9	10	8	9	3	5	1	3	6	14	4	13	8	1	4	1	1	4
Ágazat	5	16	17	27	20	33	28	31	25	22	1	6	2	5	10	25	27	17	21	18	15	16	12	7	15
Kiírföldi részarány	2	1	1	5	4	6	5	3	5	4				2		6	3	11	5	6	1	6	7	7	5
Régió	7	12	22	24	26	20	25	26	15	21		1	4	4	5	68	27	14	14	8	5	6	17	10	4
Szolgálati idő						10	14	24	11	12		4	6	2	6	6	8	10	11	13	3	10	11	12	6
Vállalati létszám	8	7	24	19	21						1	5	3	4	7	17	18	14	15	11	12	8	9	3	8
Nem		2	6	3	5	5	7	7	6	7						6	5	2	5	6	1	3	1	1	3
Településtípus	3	3	9	11	15	13	5	14	11	12	1	3	1	5	5						1	8	6	5	6
Vállalati szerződés		6	3	8	7	15	7	7	7	7	1				2	1	2	8	3	5					
Életkor	31	7		9	2	5	8	8	7	7	2	3	2	3	4		1	10	2	6		3	5	3	3
Állami részarány	2	1	8	7	5	10	6	3	3	12				1	2	4	5	6	1	8		2	5	4	2
Ágazati szerződés		2	4	6	9	18	3	6	3	2				2		3	1	5	4	4	2	9	5	2	
Több munkáltatós sz.			2	1	3	1		2	2	1							1		4	1	1	1	1	2	

Forrás: Bértarifa, saját számítás

A legfontosabb szurrogátumok az optimális fában (folyt.)

Változók	Életkor					Állami részarány					Ágazati szerződés					Több munkáltatós sz.				
Sorszám	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Foglalkozás	4	6	9	10	7			5	3	2	8		4	4	6	4		1		3
Iskolai végzettség	14	9	11	17	8	3	3	7	8	9		2	2	5	6	12	3	2	3	4
Ágazat	11	16	13	18	19	36	18	22	15	19	20	13	17	17	11	3	2	7	6	13
Külföldi részarány	1	2		4	4	15	7	3	3	4		1	2	6	4	5		1	2	2
Régió	5	5	15	24	23	6	11	20	17	16	9	15	11	16	15	4	4	11	11	9
Szolgálati idő	79	19	8	6	10	4	8	8	17	11	2	6	11	13	10	1	2	2	3	9
Vállalati létszám	10	9	17	21	22	15	10	15	6	15	11	18	9	14	10		4	3	3	6
Nem	3	2	3	4	6	1	2	9	7	7		3	2	10	3		3	1	3	2
Településtípus		2	5	9	11	2	11	7	14	14	4	7	10	4	7		2	3	2	2
Vállalati szerződés		2	1	2	1	13	8	6	8	4	7	6	8	9	2			4	4	5
Életkor						1	2	4	4	7	1	4	3	2	5	1		2	2	1
Állami részarány		3	3	8	6						2	8	8	5	3		1	1	2	
Ágazati szerződés	1		2	3	3	2	7	3	3	2						1	2	4		2
Több munkáltatós sz.		1		1	1		2	1		2	2		2	1						

Forrás: Bértarifa, saját számítás

F.5.2. táblázat. A szurrogátumok használatának százalékos aránya az összes szurrogátumhoz viszonyítva

Változók	Foglalkozás	Iskolai végzettség	Ágazat	Külföldi részarány	Régió	Szolgálati idő	Vállalati létszám
Foglalkozás		12,29	13,43	4,29	9,71	10	10,57
Iskolai végzettség	7,96		14,62	3,44	10,97	9,68	8,6
Ágazat	5,25	9,45		4,01	12,88	8,11	13,26
Külföldi részarány	3,92	5,1	15,29		12,16	5,1	9,02
Régió	4,06	10,25	15,15	3,09		9,71	11,42
Szolgálati idő	4,88	11,43	10,98	2,44	8,08		10,82
Vállalati létszám	5,96	10,93	13,54	3,23	12,67	9,81	
Nem	5,52	11,38	17,24	3,79	9,66	5,52	11,03
Településtípus	4,44	9,9	14,34	4,24	15,96	8,28	11,11
Vállalati szerződés	2,14	6,79	13,57	3,21	10,36	8,57	15,36
Életkor	3,38	10,53	11,28	3,38	7,52	18,42	13,16
Állami részarány	1,6	9,2	15,2	3,6	11,6	9,2	13,6
Ágazati szerződés	2,38	7,62	12,38	5,71	12,38	10	15,24
Több munkáltatós sz.	5	5	11,67	5	13,33	10	10

Forrás: Bértarifa, saját számítás

A szurrogátumok használatának százalékos aránya az összes szurrogátumhoz viszonyítva (folyt.)

Változók	Nem	Településtípus	Vállalati szerződés	Életkor	Állami részarány	Ágazati szerződés	Több munkáltatós sz.
Foglalkozás	4,86	8,57	4,29	10,29	5,14	5,14	1,43
Iskolai végzettség	4,52	9,68	2,37	12,69	6,45	5,81	3,23
Ágazat	2,29	10,31	6,2	7,35	10,5	7,44	2,96
Külföldi részarány	0,78	12,16	10,2	4,31	12,55	7,06	2,35
Régió	1,49	13,98	4,48	7,68	7,47	7,04	4,16
Szolgálati idő	2,74	7,32	6,4	18,6	7,32	6,4	2,59
Vállalati létszám	2,48	9,32	4,97	9,81	7,58	7,7	1,99
Nem		8,28	3,1	6,21	8,97	6,21	3,1
Településtípus	3,03		5,25	5,45	9,7	6,46	1,82
Vállalati szerződés	1,07	6,79		2,14	13,93	11,43	4,64
Életkor	5,26	7,14	5,26		6,77	5,64	2,26
Állami részarány	1,2	9,6	5,2	8		10,4	1,6
Ágazati szerződés	0,95	8,1	8,57	4,29	8,1		4,29
Több munkáltatós sz.	0	10	8,33	5	8,33	8,33	

Forrás: Bértarifa, saját számítás

F.6. Függelék: A 100 véletlen erdő leíró statisztikái

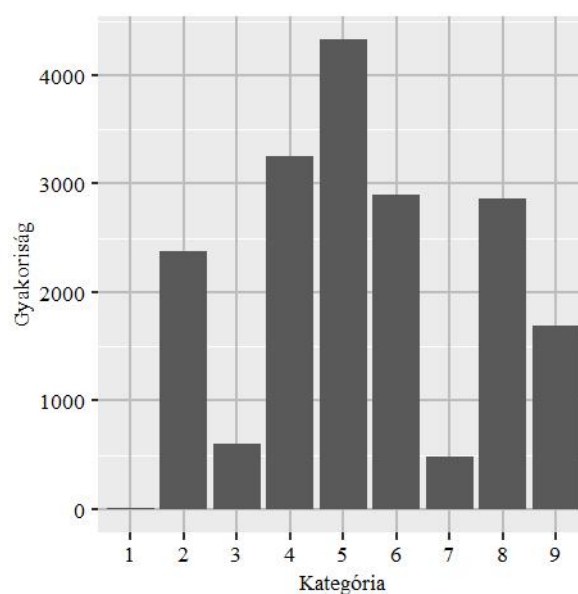
100 véletlen erdő leíró statisztikái

Változó	Átlag	Szórás	Minimum	Maximum
Foglalkozás	36,89	1,08	34,46	39,44
Iskolai végzettség	19,46	0,92	17,11	21,86
Ágazat	8,6	0,16	8,21	9,05
Külföldi részarány	9,37	0,14	9,03	9,67
Régió	6,08	0,19	5,72	6,61
Szolgálati idő	3,8	0,08	3,62	3,95
Vállalati létszám	3,58	0,12	3,32	3,85
Nem	3,71	0,07	3,54	3,87
Településtípus	2,21	0,13	1,97	2,59
Vállalati szerződés	2,47	0,07	2,32	2,67
Életkor	2,09	0,06	1,96	2,28
Állami részarány	1,03	0,06	0,92	1,17
Ágazati szerződés	0,65	0,03	0,58	0,73
Több munkáltató sz.	0,06	0,01	0,06	0,08

Forrás: Bértarifa, saját számítás

F.7. Függelék: Heterogén bérstruktúráhatás elemzéséhez tartozó változónkénti megoszlások

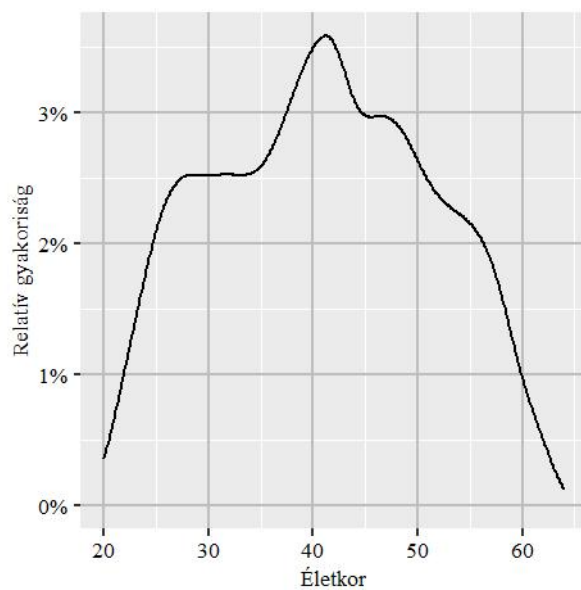
F.7.1. ábra. Megfigyelések megoszlása iskolai végzettség szerint



Megjegyzés: Az iskolai végzettségi kategóriák ugyanazok, mint a 16. ábrán.

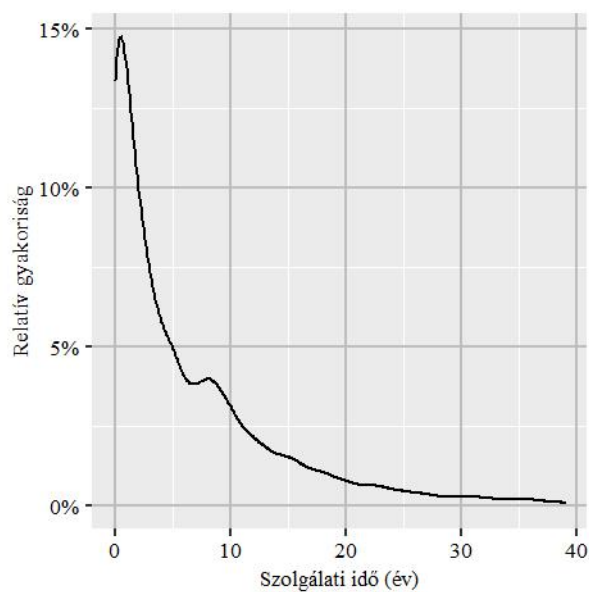
Forrás: Bértarifa, saját számítás

F.7.2. ábra. Megfigyelések megoszlása az életkor szerint



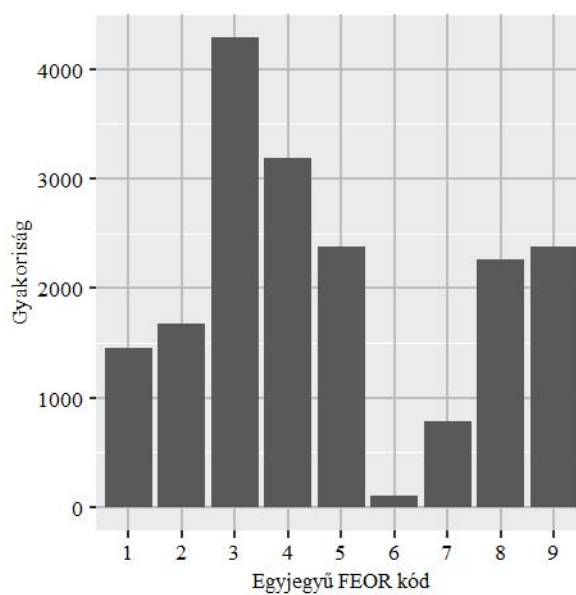
Forrás: Bértarifa, saját számítás

F.7.3. ábra. Megfigyelések megoszlása a szolgálati idő (évek) szerint



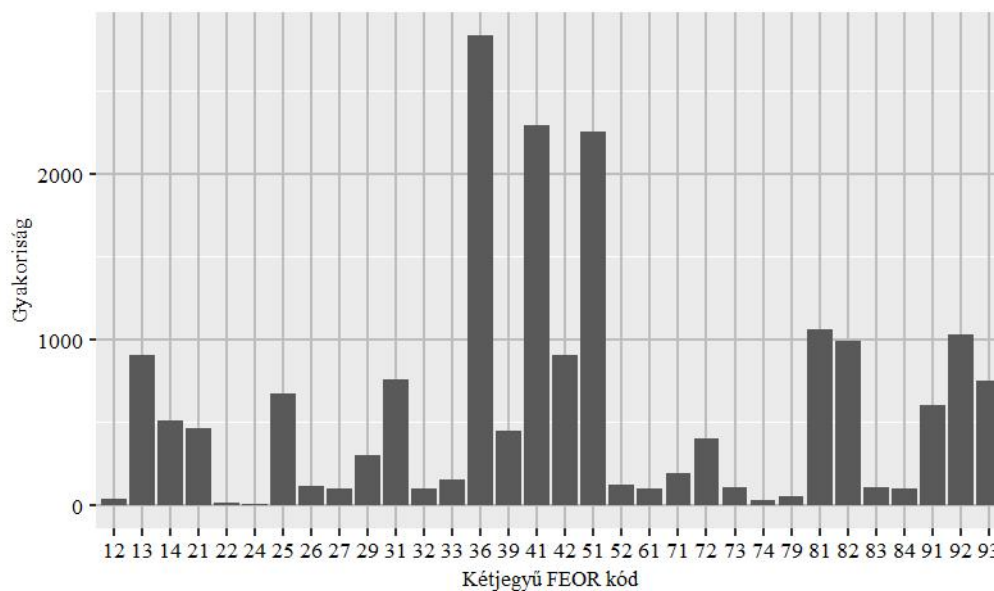
Forrás: Bértarifa, saját számítás

F.7.4. ábra. Megfigyelések megoszlása a foglalkozások – egyjegyű FEOR-kódok – szerint



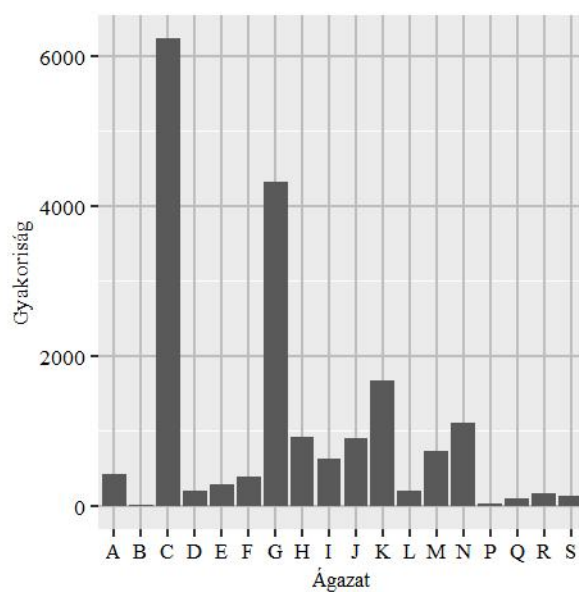
Forrás: Bértarifa, saját számítás

F.7.5. ábra. Megfigyelések megoszlása a foglalkozások – kétjegyű FEOR-kódok – szerint



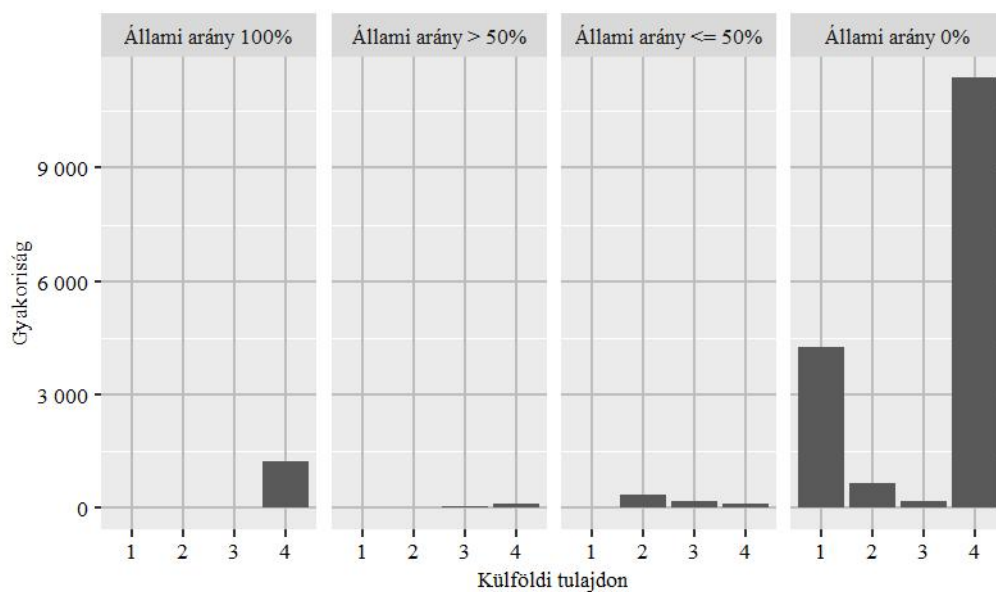
Forrás: Bértarifa, saját számítás

F.7.6. ábra. Megfigyelések megoszlása a nemzetgazdasági ágak – egyjegyű TEÁOR-kódok – szerint



Forrás: Bértarifa, saját számítás

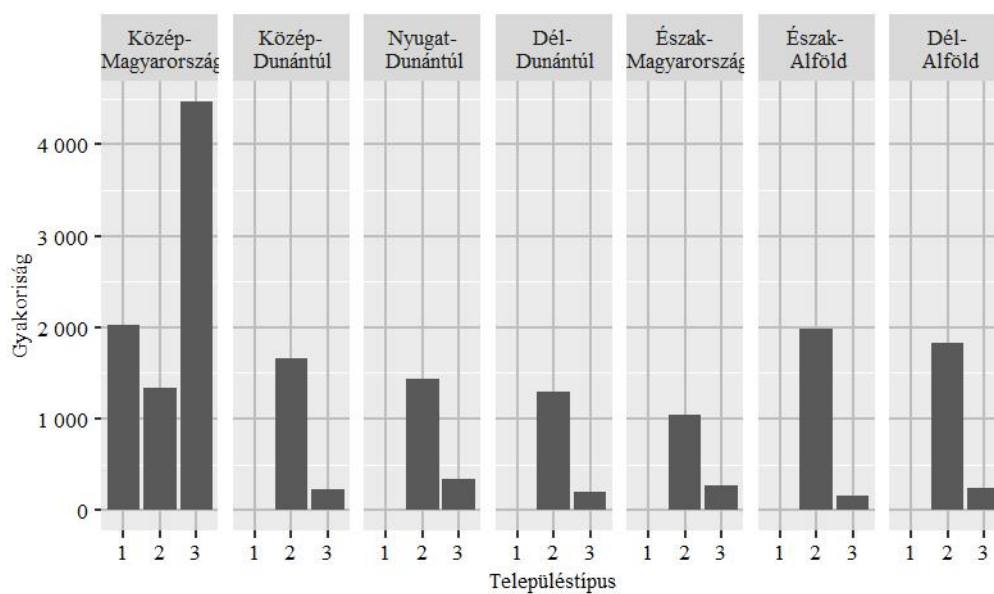
F.7.7. ábra. Megfigyelések megoszlása a tulajdontípusok szerint



Megjegyzés: A külföldi tulajdoni kategóriák ugyanazok, mint a 22. ábrán.

Forrás: Bértarifa, saját számítás

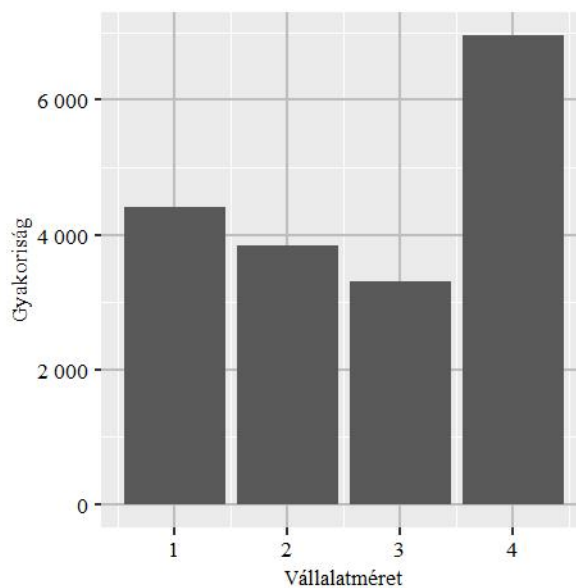
F.7.8. ábra. Megfigyelések megoszlása a régió és a településtípusok szerint



Megjegyzés: A településtípus kategóriák ugyanazok, mint a 23. ábrán.

Forrás: Bértarifa, saját számítás

F.7.9. ábra. Megfigyelések megoszlása a vállalatméret (létszám) szerint



Megjegyzés: A vállalati létszám kategóriák ugyanazok, mint a 24. ábrán.

Forrás: Bértarifa, saját számítás

F.8. Függelék: OLS-becslésben szereplő változók

A F.8.1. táblázat tartalmazza az összes olyan változót, melyeket felhasználtak a bérfüggvények becslésénél. A regressziókban a multikollinearitás és a kis létszámú csoportok miatt nem minden lehetséges keresztszorzat szerepel.

F.8.1. táblázat. Változók az OLS becslésekben 2008 és 2016 között

Változó	Típus
Iskolai végzettség	Faktor
Életkor	Folytonos
Szolgálati idő	Folytonos
Vállalati létszám	Folytonos
Állami részarány	Faktor
Külföldi részarány	Faktor
Régió	Faktor
Kollektív szerződés	Faktor
Ágazat	Faktor
Településtípus	Faktor
Foglalkozás	Faktor
Iskolai végzettség x Életkor	Folytonos
Iskolai végzettség x Szolgálati idő	Folytonos
Iskolai végzettség x Vállalati létszám	Folytonos
Életkor x Szolgálati idő	Folytonos
Életkor x Vállalati létszám	Folytonos
Életkor x Ágazat	Folytonos
Életkor x Településtípus	Folytonos
Életkor x Állami részarány	Folytonos
Életkor x Külföldi részarány	Folytonos
Életkor x Régió	Folytonos
Életkor x Kollektív szerződés	Folytonos
Életkor x Foglalkozás	Folytonos

Forrás: Bértarifa

Változó	Típus
Szolgálati idő x Vállalati létszám	Folytonos
Szolgálati idő x Ágazat	Folytonos
Szolgálati idő x Településtípus	Folytonos
Szolgálati idő x Állami részarány	Folytonos
Szolgálati idő x Külföldi részarány	Folytonos
Szolgálati idő x Régió	Folytonos
Szolgálati idő x Kollektív szerződés	Folytonos
Szolgálati idő x Foglalkozás	Folytonos
Vállalati létszám x Ágazat	Folytonos
Vállalati létszám x Településtípus	Folytonos
Vállalati létszám x Állami részarány	Folytonos
Vállalati létszám x Külföldi részarány	Folytonos
Vállalati létszám x Régió	Folytonos
Vállalati létszám x Kollektív szerződés	Folytonos
Vállalati létszám x Foglalkozás	Folytonos
Településtípus x Állami részarány	Faktor
Településtípus x Külföldi részarány	Faktor
Településtípus x Régió	Faktor
Településtípus x Kollektív szerződés	Faktor
Állami részarány x Külföldi részarány	Faktor
Állami részarány x Régió	Faktor
Állami részarány x Kollektív szerződés	Faktor
Külföldi részarány x Régió	Faktor
Külföldi részarány x Kollektív szerződés	Faktor
Régió x Kollektív szerződés	Faktor
Életkor négyzete	Folytonos
Szolgálati idő négyzete	Folytonos
Vállalati létszám négyzete	Folytonos

Forrás: Bértarifa

F.9. Függelék: Átlagos heterogén bérstruktúra-hatás és az egyes változók szerinti relatív gyakoriságai

F.9.1. táblázat. A heterogén bérstruktúrahatás (BS hatás) nagysága és az egyes változók relatív gyakoriságai (százalék) a tanulómintában

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	0,158	0,1347	0,1589	0,1657	0,1684	0,1611	0,1438	0,1623	0,1616
Életkor	1	6	6	5	5	5	5	6	6	5
	2	30	29	28	28	27	27	27	25	25
	3	27	28	29	30	32	32	31	32	31
	4	29	28	25	25	25	25	25	26	26
	5	8	10	12	11	11	11	11	12	12
Szolgálati idő	1	25	25	25	25	25	24	24	25	25
	2	26	25	25	26	26	27	26	26	26
	3	25	24	24	24	26	25	27	26	25
	4	24	26	25	25	24	24	24	24	24
Iskolai végzettség	1	16	14	14	13	13	13	13	12	12
	2	21	19	20	21	23	23	21	22	21
	3	45	46	45	45	42	40	42	43	42
	4	18	21	21	21	22	23	24	23	25
Régió	1	42	45	46	47	43	44	47	50	43
	2	30	27	27	27	29	27	26	25	28
	3	29	28	27	26	28	28	27	26	29
Ágazat	1	19	21	22	19	20	19	20	21	24
	2	35	30	30	33	34	36	33	34	32
	3	7	7	7	6	6	6	6	4	5
	4	39	42	42	41	40	39	42	41	39
Vállalatméret	1	31	29	26	26	25	26	26	21	24
	2	19	17	15	16	17	17	18	16	20
	3	19	18	19	18	20	20	19	17	18
	4	31	36	39	39	39	37	37	46	38
Külföldi részarány	1	25	20	26	27	28	29	29	23	23
	2	7	7	6	7	7	6	5	7	5
	3	2	2	2	2	1	4	3	3	2
	4	66	71	65	64	64	62	63	67	70

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

F.9.2. táblázat. A legkisebb bérstruktúrahatással (BS hatás) rendelkező csoportban (A-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tanulómintán

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	-0,1229	-0,1353	-0,1398	-0,0626	-0,0907	-0,0547	-0,0772	-0,0371	-0,0481
Életkor	1	7	5	7	7	3	4	4	9	8
	2	30	32	29	26	32	29	29	34	27
	3	26	25	26	29	26	28	32	26	25
	4	28	26	25	25	25	25	24	21	26
	5	9	11	13	13	14	13	11	10	13
Szolgálati idő	1	30	25	33	26	30	24	21	36	30
	2	26	28	21	25	27	23	27	28	32
	3	22	22	20	24	22	25	26	16	16
	4	22	25	26	26	21	28	26	20	22
Iskolai végzettség	1	3	2	1	2	3	2	3	8	4
	2	3	6	4	6	7	4	6	6	9
	3	76	68	77	87	73	85	72	66	79
	4	18	25	17	6	16	10	18	19	8
Régió	1	64	61	62	60	63	59	61	70	61
	2	18	15	18	18	18	15	17	16	17
	3	18	24	20	22	18	26	22	14	22
Ágazat	1	16	19	15	21	17	14	16	19	22
	2	8	11	9	10	11	15	17	19	17
	3	4	6	5	6	4	6	9	6	7
	4	72	64	70	62	68	65	58	56	54
Vállalatméret	1	53	58	52	51	38	42	41	41	45
	2	27	16	23	21	19	20	22	19	26
	3	10	10	8	7	12	11	10	7	10
	4	10	16	16	21	31	27	26	33	19
Külföldi részarány	1	6	8	8	9	19	19	13	19	8
	2	2	2	1	1	6	7	4	4	4
	3	3	1	3	5	2	8	5	3	5
	4	89	89	88	85	72	66	78	74	83

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

F.9.3. táblázat. A legnagyobb bérstruktúrahatással (BS hatás) rendelkező csoportban (M-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tanulómintán

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	0,4316	0,406	0,4355	0,4023	0,4449	0,3953	0,3842	0,3919	0,3764
Életkor	1	2	2	2	1	1	1	2	0	1
	2	24	19	21	18	17	17	19	13	14
	3	34	34	35	36	39	38	39	42	40
	4	33	33	30	32	30	31	29	33	30
	5	7	12	12	13	13	14	11	12	14
Szolgálati idő	1	12	12	13	12	11	10	14	7	13
	2	20	19	20	19	15	20	19	14	16
	3	34	34	35	29	32	33	29	35	29
	4	34	35	31	39	42	38	38	43	42
Iskolai végzettség	1	16	23	15	20	12	9	7	8	12
	2	42	34	35	35	29	43	27	20	30
	3	33	29	33	24	28	19	30	36	24
	4	10	14	18	21	32	28	35	36	35
Régió	1	22	32	27	28	35	27	39	44	31
	2	46	38	45	43	40	45	31	30	37
	3	32	30	27	29	26	28	30	26	32
Ágazat	1	21	23	25	13	13	15	21	21	26
	2	64	54	53	56	55	58	45	47	42
	3	3	4	4	5	6	6	4	3	4
	4	12	19	19	26	26	21	30	29	28
Vállalatméret	1	10	8	12	13	13	12	13	13	17
	2	13	17	12	13	14	14	13	12	14
	3	30	28	28	30	31	30	30	26	23
	4	47	47	48	44	43	44	43	49	46
Külföldi részarány	1	56	45	56	52	51	41	51	39	38
	2	15	12	11	11	12	6	3	9	8
	3	2	3	3	2	1	3	1	2	3
	4	28	40	30	36	36	51	44	50	52

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

F.9.4. táblázat. A legnagyobb és a legkisebb bérstruktúrahatással rendelkező csoportok (M- és A-csoportok) különbsége (százalékpont) a tesztmintán

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	0,5452	0,5285	0,5639	0,4552	0,5071	0,4362	0,4559	0,4283	0,4181
Életkor	1	-6	-6	-4	-4	-5	-5	-3	-10	-7
	2	-10	-16	-8	-11	-16	-13	-10	-22	-11
	3	11	11	6	5	14	10	8	19	14
	4	6	10	7	9	7	10	6	11	5
	5	-2	2	-1	2	-1	-2	-1	2	0
Szolgálati idő	1	-18	-16	-23	-15	-22	-17	-9	-32	-17
	2	-4	-15	-4	-8	-11	-6	-13	-14	-15
	3	7	16	15	9	9	10	3	19	10
	4	15	15	11	15	24	13	18	25	21
Iskolai végzettség	1	13	20	15	18	8	7	4	-2	5
	2	42	33	33	24	25	41	20	15	23
	3	-42	-37	-44	-58	-50	-67	-40	-30	-55
	4	-11	-15	-3	16	18	20	16	17	27
Régió	1	-45	-27	-35	-30	-32	-35	-21	-26	-29
	2	30	27	27	22	21	31	15	17	22
	3	15	0	7	8	11	4	6	10	7
Ágazat	1	5	0	6	-8	-1	0	3	1	6
	2	56	48	46	47	44	46	30	29	24
	3	-1	-1	-2	-1	1	1	-4	-1	-1
	4	-60	-47	-51	-39	-45	-48	-30	-30	-28
Vállalatméret	1	-41	-53	-42	-40	-30	-32	-29	-27	-30
	2	-15	-4	-12	-7	-7	-7	-9	-7	-13
	3	21	21	21	22	22	21	20	18	13
	4	35	36	33	24	16	17	20	17	31
Külföldi részarány	1	52	39	49	42	36	23	37	20	32
	2	11	12	10	10	6	-2	0	6	4
	3	-1	2	1	-4	-2	-5	-4	-1	-2
	4	-62	-53	-59	-48	-40	-15	-34	-25	-34

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

F.9.5. táblázat. A heterogén bérstruktúrah hatás (BS hatás) nagysága és az egyes változók relatív gyakoriságai (százalék) a tesztmintában

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	0,1576	0,1358	0,1562	0,1644	0,1662	0,1599	0,1455	0,1617	0,1612
Életkor	1	6	5	5	5	5	5	6	6	5
	2	30	29	28	28	27	27	26	26	25
	3	27	28	29	31	31	31	32	32	31
	4	29	27	26	24	26	26	25	26	27
	5	8	10	11	12	12	11	11	11	12
Szolgálati idő	1	25	25	25	24	26	24	24	24	26
	2	25	25	26	26	26	26	25	26	25
	3	25	25	24	26	26	26	26	26	26
	4	24	25	25	24	23	24	24	24	24
Iskolai végzettség	1	16	14	14	13	13	13	13	12	12
	2	21	19	20	21	23	23	22	22	21
	3	44	47	45	45	41	41	41	42	42
	4	18	20	21	21	22	24	24	24	25
Régió	1	42	45	46	47	43	45	47	49	44
	2	29	27	27	26	29	27	26	25	28
	3	29	28	27	27	28	28	27	26	29
Ágazat	1	19	20	22	20	20	19	20	21	23
	2	35	30	30	33	34	37	33	34	33
	3	7	7	6	6	6	6	5	4	5
	4	39	43	42	41	41	39	42	41	39
Vállalatméret	1	31	28	27	26	24	25	26	21	24
	2	19	17	16	16	17	17	18	16	21
	3	19	18	19	19	20	20	19	17	17
	4	30	36	39	39	39	38	37	46	38
Külföldi részarány	1	25	19	25	26	28	28	30	24	23
	2	7	8	7	7	7	6	5	7	6
	3	2	2	2	2	1	4	2	3	2
	4	65	72	66	65	63	62	62	67	69

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

F.9.6. táblázat. A legkisebb bérstruktúrahatással (BS hatás) rendelkező csoportban (A-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tesztmintán

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	-0,1157	-0,1227	-0,1372	-0,0592	-0,0731	-0,0479	-0,0684	-0,0343	-0,0419
Életkor	1	8	7	6	5	5	6	5	10	8
	2	33	33	29	28	32	28	27	34	27
	3	24	25	28	31	26	29	31	25	26
	4	27	25	24	23	24	24	24	21	25
	5	9	10	13	13	13	13	13	10	13
Szolgálati idő	1	30	27	34	28	32	25	22	38	31
	2	25	30	22	25	25	25	28	27	31
	3	25	21	21	22	23	25	26	17	18
	4	20	22	24	24	20	25	24	19	21
Iskolai végzettség	1	3	1	1	2	3	1	3	9	6
	2	2	4	4	7	7	4	7	6	7
	3	75	67	76	86	76	87	71	65	79
	4	19	27	19	5	13	7	19	20	8
Régió	1	67	59	60	59	65	61	60	69	61
	2	16	13	18	18	17	15	17	15	17
	3	17	28	22	23	18	24	22	16	22
Ágazat	1	15	20	16	21	15	14	17	20	20
	2	9	9	9	10	10	14	16	17	19
	3	4	5	6	6	6	5	8	5	6
	4	72	66	70	64	69	67	60	58	55
Vállalatméret	1	53	61	54	53	40	44	42	42	46
	2	28	19	25	19	20	20	23	19	27
	3	10	8	7	6	10	9	9	6	9
	4	9	12	14	22	29	27	25	32	17
Külföldi részarány	1	6	7	6	8	16	18	13	17	6
	2	2	1	1	1	7	8	4	4	4
	3	3	1	2	5	3	7	5	4	5
	4	88	91	90	86	74	66	78	75	85

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

F.9.7. táblázat. A legnagyobb bérstruktúrahatással (BS hatás) rendelkező csoportban (M-csoport) az átlagos bérstruktúrahatás és az egyes változók relatív gyakoriságai (százalék) a tesztmintán

Változók	Kategória	2008	2009	2010	2011	2012	2013	2014	2015	2016
BS hatás	Átlag	0,4295	0,4058	0,4267	0,396	0,434	0,3883	0,3875	0,394	0,3762
Életkor	1	2	1	2	1	0	1	2	0	1
	2	23	17	21	17	16	15	17	12	16
	3	35	36	34	36	40	39	39	44	40
	4	33	35	31	32	31	34	30	32	30
	5	7	12	12	15	12	11	12	12	13
Szolgálati idő	1	12	11	11	13	10	8	13	6	14
	2	21	15	18	17	14	19	15	13	16
	3	32	37	36	31	32	35	29	36	28
	4	35	37	35	39	44	38	42	44	42
Iskolai végzettség	1	16	21	16	20	11	8	7	7	11
	2	44	37	37	31	32	45	27	21	30
	3	33	30	32	28	26	20	31	35	24
	4	8	12	16	21	31	27	35	37	35
Régió	1	22	32	25	29	33	26	39	43	32
	2	46	40	45	40	38	46	32	32	39
	3	32	28	29	31	29	28	28	26	29
Ágazat	1	20	20	22	13	14	14	20	21	26
	2	65	57	55	57	54	60	46	46	43
	3	3	4	4	5	7	6	4	4	5
	4	12	19	19	25	24	19	30	28	27
Vállalatméret	1	12	8	12	13	10	12	13	15	16
	2	13	15	13	12	13	13	14	12	14
	3	31	29	28	28	32	30	29	24	22
	4	44	48	47	46	45	44	45	49	48
Külföldi részarány	1	58	46	55	50	52	41	50	37	38
	2	13	13	11	11	13	6	4	10	8
	3	2	3	3	1	1	2	1	3	3
	4	26	38	31	38	34	51	44	50	51

Megjegyzés: Ugyanaz, mint a 21. táblázatnál.

Forrás: Bértarifa, saját számítás

Saját publikációk a témában

Takács, O. – Vincze, J. (2018a): Bérelőrejelzések - prediktorok és tanulságok. *Közgazdasági Szemle*, 55(6):592–618,
DOI: <http://dx.doi.org/10.18414/kszh.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 bérkülönbségek Magyarországon: a véletlen erdő- és az OLS-becslésen alapuló Blinder–Oaxaca-dekompozíció eredményeinek összehasonlítása. *Statistikai Szemle*, 99(1):5–45.
DOI: <http://dx.doi.org/10.20311/stat2021.1.hu0005>