Statisztikai Szemle

Közzététel: 2021. január 13.

A tanulmány címe:

Nemek közötti bérkülönbségek Magyarországon: a véletlenerdő- és az OLS-becslésen alapuló Blinder–Oaxaca-dekompozíció eredményeinek összehasonlítása

Szerző:

TAKÁCS OLGA, a Budapesti Corvinus Egyetem PhD-hallgatója
E-mail: olga.takacs@stud.uni-corvinus.hu

DOI: https://doi.org/10.20311/stat2021.1.hu0005

Takács Olga

Nemek közötti bérkülönbségek Magyarországon: a véletlenerdő- és az OLS-becslésen alapuló Blinder–Oaxaca-dekompozíció eredményeinek összehasonlítása*

Gender wage gap in Hungary: comparison of Blinder–Oaxaca decompositions based on random forest and OLS estimations

TAKÁCS OLGA, a Budapesti Corvinus Egyetem PhD-hallgatója
E-mail: olga.takacs@stud.uni-corvinus.hu

A Blinder–Oaxaca-dekompozíció a nemek közötti átlagos bérkülönbséget egy megfigyelt jellemzőkkel magyarázott (összetételhatásra) és egy nem magyarázott részre (bérstruktúrahatásra) bontja fel. A szerző a férfiak és a nők bérfüggvényének felbontásához a legkisebb négyzetek módszerét (ordinary least squares, OLS) és a véletlenerdő-algoritmust alkalmazza; célja az így kapott eredmények összehasonlítása magyar adatokon. Az összetételhatás nagyobb a véletlen erdőnél, ami kisebb bérstruktúrahatást eredményez, azonban a becsült bérstruktúrahatások közel azonosak, és az egyéni szinten mértek közötti korreláció alacsony. Vagyis egyéni szinten jelentős különbségek lehetnek a két módszerrel becsült bérstruktúrahatásban. A leíró statisztikai elemzések eredményei a becslésekhez használt változók tekintetében különbségeket mutatnak, amelyek nem vezethetők vissza egyetlen változóra. A szerző következtetései megerősítik *Weichselbaumer–Winter-Ebmer* [2005] eredményeit, akik rámutattak arra, hogy a dekompozíciónál alkalmazott becslési eljárások átlagosan ugyanakkora összetétel- és bérstruktúrahatást becsülnek, azonban a módszertan jelentősen befolyásolja az egyes változók fontosságát.

TÁRGYSZÓ: nemek közötti bérkülönbség, véletlenerdő-algoritmus, Magyarország

The Blinder–Oaxaca decomposition splits the average gender wage gap into two parts: one that is explained by covariates (composition effect) and another that is unexplained (wage structure effect). In this study, male and female wage functions are decomposed by OLS and random forest regressions, and the results obtained on a Hungarian dataset are compared. It is found that the composition effect is larger in random forest regression than in OLS regression, resulting in smaller

* A szerző ezúton mond köszönetet *Vincze János*nak, *Csóka Imolá*nak és *Makai Márton*nak a cikkel kapcsolatos észrevételeikért és hasznos tanácsaikért.

wage structure effects; however, these latter estimated by the two methods are almost identical. Individual wage structure effects have low correlation, i.e. there can be significant differences among them. Descriptive statistics of covariates show differences between the two methods, which cannot be traced back to one single covariate. The author's conclusions confirm the results of *Weichselbaumer–Winter-Ebmer* [2005] who argued that different methodologies of the Blinder–Oaxaca decomposition estimate the same composition and wage structure effects on average but the methodology significantly influences the importance of each variance.

KEYWORD: gender wage gap, random forest, Hungary

A világ nagy részén az utóbbi évtizedekben a nemek közötti bérkülönbségek folyamatos csökkenését dokumentálták, amelyek azonban nem tűntek el teljesen (*Weichselbaumer–Winter-Ebmer* [2007]). Az *OECD* [2018] szerint 39 százalékkal kerestek kevesebbet a nők 2015-ben, mint a férfiak; ugyanakkor az országok között jelentős eltérések tapasztalhatók. A legnagyobb különbséget Japánban, Koreában, Mexikóban és Chilében mutatták ki, Európán belül a mediterrán országokban, a német anyanyelvű országokban, Hollandiában és Csehországban, a legkisebbet – 30 százalék alattit – pedig Kelet-Európában, a skandináv országokban és Portugáliában (*OECD* [2018]).

Az átlagos értékek általános képet adnak arról, hogy mekkorák a nemek közötti bérkülönbségek a világban és az egyes országokban, azonban nem mondanak semmit arról, hogy a különbség megléte mögött ténylegesen mi áll, a nőkkel szembeni diszkriminációval vagy a két nem eltérő munkaerőpiaci jellemzőivel magyarázható-e az eltérés. Ez a dilemma politikai szempontból sem elhanyagolható. Számos országban célzott intézkedéseket hoztak e különbség felszámolására. Például az Európai Unióban született egy stratégiai keretmegállapodás, melyben több ország – közöttük Magyarország is – egyéni szintű intézkedéseket vezetett be a nemek közötti bérkülönbség megszüntetése érdekében (*European Commission* [2016]).

A szakirodalomban a diszkrimináció és a munkaerőpiaci jellemzőkkel magyarázható rész elkülönítésére a Blinder–Oaxaca-felbontást (*Blinder* [1973], *Oaxaca* [1973], *Jann* [2008]) használják. A dekompozíció az átlagos bérkülönbséget magyarázott és nem magyarázott részre bontja fel: előbbi a férfiak és a nők között megfigyelt egyéni jellemzőkből adódó különbséget tartalmazza; utóbbi pedig azt a bérkülönbséget mutatja, amely nem magyarázható e jellemzőkkel. *Blinder* [1973] és *Oaxaca* [1973] szerint a nem magyarázott rész megfeleltethető a diszkrimináció mértékének, azonban feltételezhető, hogy egyéb módszertani okokra – mint a kiha-

gyott változók hatására vagy a munkatapasztalat hibás mérésére – visszavezethető eltéréseket is tartalmaz (*Reilly* [2001]). Mindezek miatt óvatosnak kell lenni a magyarázott és nem magyarázott részek értelmezésénél.

A nemi bérkülönbségek felbontáshoz szükség van a nők és a férfiak alcsoportjára vonatkozó modellekre. Ezek *Blinder* [1973] és *Oaxaca* [1973] esetében OLS-becsléssel készültek. Későbbi kutatásokban a Blinder–Oaxaca-felbontást tovább általánosították és pontosították. *DiNardo–Fortin–Lemieux* [1996] kimutatták, hogy a magyarázó változók eloszlásbeli változásának hatása van a Blinder–Oaxaca-féle felbontással kapott magyarázott és nem magyarázott részekre, emiatt érdemes átsúlyozni az inputváltozókat. *Machado–Mata* [2005] kvantilis regressziót, valamint a munkaerőpiaci részvétel miatt bekövetkező torzítások kezelésére *Heckman*-féle [1979] szelekciót alkalmaztak a Blinder–Oaxaca alapjául szolgáló modellekhez (*Reimers* [1983], *Neuman–Oaxaca* [2004]).[1]

Ezek az általánosítások a különböző kérdések megválaszolására még alkalmasabb eszközzé tették a Blinder–Oaxaca-dekompozíciót azáltal, hogy jobban figyelembe vették a munkaerőpiac jellemzőit. Azonban ezek a regressziós becslések megkövetelték az adatgeneráló-folyamat előzetes ismeretét. A bérfüggvények esetében *Mincer* [1974] kimutatta a bérek és a munkatapasztalat közötti nemlineáris kapcsolatot. Amennyiben a tapasztalat négyzetét is felhasználjuk a dekompozíciós bérbecslésekben, akkor ennek megfelelően javul regressziós előrejelzésük, és feltételezhetően a magyarázott és a nem magyarázott rész nagysága pontosabban meghatározható. Azonban az adatokban lehetnek további rejtett összefüggések, melyeket a hagyományos regressziónál továbbra sem vesznek figyelembe. Magyarország esetében például *Earle–Telegdy* [2012] kimutatták, a bérek alakulására szignifikánsan hat, hogy kinek a tulajdonában van a vállalat. A külföldi tulajdonú vállalatoknál növekedtek a bérek, azonban a bérnövekedés szintje eltérően alakult a különböző dolgozói csoportoknál. A szerzők arra jutottak, hogy amennyiben egy hazai vállalat külföldi tulajdonba került, akkor leginkább a magasan képzett fiatal munkaerő „járt jól". Emellett *Fazekas* [2005] bemutatta, hogy a rendszerváltás után a külföldi vállalatok földrajzi elhelyezkedését erősen befolyásolta a nyugati határhoz való közelség, míg a magyarokét nem. Az ipari hagyományok és azon települések elhelyezkedése, amelyekre a magasabb iskolázottság jellemző, a külföldi és a magyar tulajdonosoknak is számított. Mindkét kutatási eredmény arra utal, hogy Magyarországon a bérfüggvények meghatározásánál vizsgált magyarázó változók feltehetően nem függetlenek, és léteznek közöttük nemlineáris kapcsolatok.

Jelen tanulmányban a Blinder–Oaxaca-dekompozíció bérfüggvénybecsléséhez az OLS mellett egy gépi tanulási algoritmust is alkalmazok, a véletlen erdőt (random

[1] Az itt bemutatott általánosításokat használják a bérkülönbségek vizsgálatánál. A további általánosításokról összefoglaló képet ad *Fortin et al.* [2011].

forest), amely a klasszifikációs és regressziós fákra (classification and regression tree – CART) épül. A gépi tanulás szakirodalmában használt nemparaméteres eljárások előnye a hagyományos módszerekkel szemben, hogy képesek „rátanulni" a rejtett összefüggésekre, így mindenféle előzetes feltételezés nélkül tudják kinyerni az adatban rejlő információkat. Ezért az eljárások jobban teljesítenek, ha vannak nemlineáris összefüggések az inputváltozók között, és ahogyan *Earle–Telegdy* [2012] és *Fazekas* [2005] rámutattak, a magyar adatokban lehetnek is. *Takács–Vincze* [2019a] több évre vonatkozóan is igazolták, hogy a véletlen erdővel készített bérfüggvényeknek jobb az előrejelző képességük, mint az OLS-sel készülteknek. A Blinder–Oaxaca-dekompozícióra kiterjesztett véletlen erdő leírását *Takács–Vincze* [2019b] tanulmánya részletesen tartalmazza.

Célom *Takács–Vincze* [2019a], [2019b] kutatásaira építve a véletlen erdővel és az OLS-sel becsült bérfüggvények alapján készült Blinder–Oaxaca-dekompozíciók összehasonlítása. Az összevetés több szinten történik, ugyanis elvégzem az átlagos magyarázott és nem magyarázott részre is, majd megvizsgálom, hogy a két módszer szerint mennyiben tér el az egyéni szinten számszerűsített nem magyarázott rész nagysága. Ezzel az összehasonlítással pontosabb képet kaphatunk arról, hogy a Blinder–Oaxaca-dekompozícióhoz használt módszertan mennyire befolyásolja az eredményeket.

A bérfüggvények előállításához alkalmazott regressziós fákat és véletlenerdő-eljárást az 1. fejezetben mutatom be. Továbbá ez a fejezet tartalmazza a Blinder–Oaxaca-féle felbontás módszertanát és annak adaptációját a véletlen erdőre. A 2. fejezetben a becslések adatbázisát és a változókat tárgyalom. A 3. fejezetben három lépésben összevetem az OLS-sel és a véletlen erdővel készült eredményeket. Elsőként a becsült bérfüggvények előrejelzéseit hasonlítom össze, majd a Blinder–Oaxaca-dekompozíció eredményeként kapott magyarázott és nem magyarázott rész alakulását vizsgálom meg, végül az egyéni szinten számított nem magyarázott részt elemzem, főként leíró statisztikai módszerekkel. A 4. fejezetben összefoglalom a tanulságokat.

1. Módszertan

A döntési fák felépítésére számos eljárás létezik, melyek közül én a CART-ot, azon belül is a folytonos változókra használt regressziós fa építésének módszertanát mutatom be[2] az 1.1. alfejezetben. Amiatt korlátozom a leírást ennek az egyféle algo-

[2] A klasszifikációs és regressziós fák ugyanolyan fa formában prezentálható eredményt adnak. A különbség közöttük, hogy a klasszifikációs fák kategóriaváltozókra, a regressziós fák pedig folytonos változókra alkalmazhatók.

ritmusnak a bemutatására, mert a véletlen erdő is ezt használja. Ezt követően az 1.2. alfejezetben ismertetem, hogy az OLS- és a véletlenerdő-becslés miként alkalmazható a Blinder–Oaxaca-felbontásban. Az 1.3. alfejezetben az egyéni szinten számszerűsített nem magyarázott részt tárgyalom, összehasonlítom az OLS-sel és a véletlen erdővel készült Blinder–Oaxaca-felbontás eredményeit.

1.1. A klasszifikációs és regressziós fáktól a véletlen erdőig

A fa építésének kiindulópontja egy *N* darab *i*-vel jelölt megfigyelésekből álló adathalmaz, amelyet a fa tetején található gyökér jelenít meg. Ez az adathalmaz tartalmazza a *p* darab inputváltozót és – regressziós fák esetében – a folytonos függőváltozót minden megfigyeléshez.

A fa építésénél első lépésben az eljárás a megfigyeléseket két egymást nem átfedő részhalmazra osztja, és az egyes csoportokhoz hozzárendeli az eredményváltozó átlagát mint a függővváltozóra vonatkozó előrejelzést. Ebben a lépésben az a cél, hogy két olyan – R_j-vel jelölt – csoportot határozzunk meg, amelyek eltérésnégyzetösszege (residual sum of squares, RSS) minimális:

$$RSS = \sum_{j=1}^{J=2} \sum_{i \in R_j} \left(y_i - \widehat{y_{R_j}} \right)^2 ,$$

ahol y_i az egyedi megfigyelésekhez tartozó eredményváltozó, míg y_{R_j} az egyedi megfigyelést tartalmazó R_j csoportban az eredményváltozó átlaga. Ez lesz az első vágási pont, melynek megtalálásához az algoritmus végigmegy a *p* darab inputváltozón, és mindegyiknél megnézi annak összes lehetséges értéke szerint a felosztást. Az algoritmus minden felosztásnál számszerűsíti az RSS csökkenését és kiválasztja a legkisebb RSS-t adó vágást. Az eljárás ezzel meghatározza az első vágási pontot, és a teret két nem átfedő részhalmazra bontja. Grafikusan ez úgy jeleníthető meg, hogy a gyökérből leágazik két levél, amelyek a részhalmazokat testesítik meg. Az egyes levelekhez tartozó megfigyelések pedig az eredetihez képest homogénebb csoportokat alkotnak.

A következő lépésben az algoritmus már nem az egész teret vizsgálja, hanem a két levelet külön-külön: megnézni, hogy miként tudja további két részre bontani az adott levelet úgy, hogy az összes levélre vonatkozó RSS a lehető legkisebb legyen. Ezt követően az eljárás további két levélre bontja az adott levelet, amely így csomóponttá válik, ezzel a teret három részre osztja, három végpontot alkotva. A CART a fa építése során bináris vágásokat hajt végre: mindig két új levelet hoz létre.

Az újabb vágásokkal egyre több diszjunkt részhalmaz keletkezik, vagyis nő a levelek száma, miközben az RSS folyamatosan csökken. Az algoritmus egészen addig növeszti a fát, amíg egy előre meghatározott kilépési kritériumot elér: például, ha a levelek nagysága elér egy minimális értéket, vagy ha az RSS csökkenése kisebb, mint egy előre meghatározott érték.

A faépítés előtt nem kell meghatározni az adatgeneráló folyamatot, mert a leveleken levő végső csoportokat és azok jellemzőit egy „mohónak" tekinthető eljárással határozza meg az algoritmus. A mohóság abból adódik, hogy az eljárás mindig a legkisebb RSS-t eredményező vágást hajtja végre. Az RSS a gyökérnél maximális a fa növekedésével egyre csökken, így az eredményváltozó átlaga egyre jobb becslésnek bizonyul a csoport tagjaira nézve. Tehát a nagyobb fák általában jobb előrejelzési képességgel rendelkeznek. Azonban nagy fák esetében előfordulhat, hogy a CART „túlilleszt" az adatokon: „rátanul" az adat egyéni mintázataira, csökkentve ezzel az általánosítási képességet. Ennek elkerülésére a fákat általában egy tanulómintán építik, és egy tesztmintán külön ellenőrzik a teljesítőképességét. A nagy fák másik problémája, hogy nehezebben értelmezhetők, pedig a döntési fák előnye éppen a könnyű érthetőségben rejlik. A kisebb fák – bár egyszerűbben interpretálhatók – általánosítási képességei rosszabbak lehetnek azáltal, hogy néhány lényeges vágás kimaradhat. Ennek kiküszöbölésére az a gyakorlat, hogy nagy fát kell építeni, amely ezután visszametszhető.

A CART-ok hátránya, hogy egyéb paraméteres eljárásokkal összehasonlítva magasabb a varianciájuk, érzékenyek az adatokban történt kis változásokra, és emiatt előrejelzési képességük is gyenge. Azonban több CART eredményeinek aggregálásával az előrejelző-képesség és az eredmények robusztussága növelhető. Ilyen aggregálási eljárás a bagging, a boosting és a véletlen erdő (*James et al.* [2017]).

A bagging egy-egy fát épít a tanulómintából a bootstrap-eljárással készített almintákon, majd ezek eredményeit összegzi. A boosting egy adatbázisból készít szekvenciális fákat, vagyis az adott fa eredménye függ az előzőtől (*Breiman* [1998], [2001]). A véletlen erdő a bagginghez hasonlóan a tanulómintából képzett bootstrap-mintákon készíti el a lehető legnagyobb metszés nélküli fákat úgy, hogy minden vágásnál csak a változók egy részét használja fel. A következő lépésben az elkészült közel korrelálatlan fák eredményeit összegzi, ami javítja az eljárás előrejelző-képességét (*Loh* [2009]).

A gépi tanulási algoritmusok és így a véletlen erdő is számos ún. hiperparaméter előzetes meghatározását teszi szükségessé: a fák számát, az egyes fáknál a levelek méretét és a potenciális vágóváltozók számát. Ezekre a hiperparaméterekre léteznek ajánlások, illetve különböző optimalizáló algoritmusok. A véletlen erdő készítéséhez *Breiman* [2001] alapján az R program randomForestSRC-csomagját használtam, és ennek tune.rfsrc parancsa segítségével a különböző paraméterbeállításokat vizsgáltam a férfiak és a nők bérére vonatkozóan. Összesen 500 fa

eredményét aggregáltam, és ugyanennyi fa esetében végeztem el a levélméret, valamint a potenciális vágóváltozók számának meghatározását.[3] A levél optimális mérete a férfiaknál 8, a nőknél 5. A felhasznált inputváltozók közül az algoritmus a nőknél 12-t, míg a férfiaknál 13-t tekint potenciális vágóváltozónak.[4] Az optimalizáció eredményét, a validációs hibák nagyságát a levélméret és a potenciális vágóváltozók, illetve fák számának függvényében az 1. Függelék tartalmazza.

1.2. Blinder–Oaxaca-felbontás

A Blinder–Oaxaca-felbontás szerint a nyers bérkülönbség megfeleltethető a férfiak és a nők átlagbérkülönbségének, amelyre *Blinder* [1973] alapján a következő összefüggés áll fenn:

$$av\left(y_M\right) - av\left(y_F\right) = \bar{X}'_M\beta_M - \bar{X}'_F\,\beta_F, \qquad /1/$$

ahol $av\left(y_M\right)$ és $av\left(y_F\right)$ az átlagos béreket, $\bar{X}'_M$ és $\bar{X}'_F$ a magyarázó változók átlagos értékeit jelölik a két nem esetében, β_M és β_F e két alcsoport OLS-sel becsült együtthatói. Az /1/ összefüggés tovább bontható, és a következő formában írható fel:

$$\bar{X}'_M\beta_M - \bar{X}'_F\,\beta_F = \left(\bar{X}_M - \bar{X}_F\right)'\beta_M + \bar{X}'_F\left(\beta_M - \beta_F\right), \qquad /2/$$

ahol a jobb oldalon az első tag a magyarázó változók által magyarázott rész, a második tag pedig a nem magyarázott rész. Ez utóbbi mutatja, hogy átlagosan mennyit kapnának a nők, ha munkájuk a férfiak bérfüggvényével lenne beárazva. Amennyiben a nem magyarázott rész nulla, akkor átlagosan a férfiak és a nők munkáját ugyanúgy árazzák a munkaerőpiacon.

A véletlen erdő esetében az /1/ és a /2/ egyenletekben található β együtthatók nem becsülhetők, így a következő összefüggést használtam a nyers különbségek meghatározására:

$$av(y_M) - av(y_F) = av(P^M(M)) - av(P^F(F)) + torzítás, \qquad /3/$$

[3] Az erdő mérete is olyan hiperparaméter, melyet érdemes lehet optimalizálni, azonban jelen esetben úgy vélem, hogy a lehetséges hiperparaméterek közül kettő optimalizálása is elégséges.

[4] A 12 és 13 lehetséges vágóváltozó a nők és a férfiak esetében meglehetősen magas, tekintve, hogy a hüvelykujjszabály az inputváltozók harmadát ajánlja. A potenciális vágóváltozók ilyen nagy száma arra utal, hogy a változók között van néhány lényeges, és a cél az, hogy ezek minden lépésben bekerüljenek a potenciális vágóváltozók közé.

ahol P^M és P^F jelölik a férfiak és a nők almintáinak véletlen erdővel készített modelljeit, $av\left(P^M\left(M\right)\right)$ a férfiak almintájára a férfi modellel készített bérelőrejelzések átlagát, $av\left(P^F\left(F\right)\right)$ pedig a nők alcsoportjára a női modellel készített bérelőrejelzések átlagát. A magyarázott és a nem magyarázott rész a következőképpen áll elő:

$$\begin{aligned} & av\left(P^M\left(M\right)\right) - av\left(P^F\left(F\right)\right) = \\ & = \left[av\left(P^M\left(M\right)\right) - av\left(P^M\left(F\right)\right)\right] + \left[av\left(P^M\left(F\right)\right) - av\left(P^F\left(F\right)\right)\right]. \end{aligned} \qquad /4/$$

Ahogyan a /2/ egyenletnél, úgy itt is, jobb oldalon az első tag a megfigyelt változók által magyarázott, a második tag pedig a nem magyarázott részként értelmezhető (*Takács–Vincze* [2019b]).

A Blinder–Oaxaca-dekompozícióval kapcsolatban *Fortin–Lemieux–Firpo* [2011] kiemelik azt a tényt, hogy a magyarázott rész a két csoport jellemzőinek különböző eloszlására vezethető vissza. A nem magyarázott rész létezését pedig a becslőfüggvények közötti különbség okozza. Emiatt a szerzők a magyarázott részre összetételhatásként, a nem magyarázott részre pedig bérstruktúrahatásként utalnak. A továbbiakban én is ezeket a megnevezéseket használom.

1.3. Az egyéni szintű bérstruktúrahatás vizsgálata

A Blinder–Oaxaca-eljárás felbontja a nyers bérkülönbséget összetétel- és bérstruktúrahatásra, emellett megmutatja, mekkora az egyes magyarázó változók hatása a különbségre. OLS-sel becsült modellek esetében a β_M és a β_F együtthatók, illetve ezek különbségei mutatják e hatásokat. A véletlen erdőnél nem állnak rendelkezésre hasonló mérőszámok. Azonban OLS-sel és véletlen erdővel is egyéni szinten meg lehet határozni a bérstruktúrahatás nagyságát, mely eredmények így már összehasonlíthatók. Tehát az inputváltozók bérstruktúrára gyakorolt hatásának elemzéséhez egyéni szinten készítek becslést a bérstruktúrahatás nagyságára, és ezeket vizsgálom tovább.

Ahogyan a /2/ egyenletből is látszik, az OLS-együtthatók különbségei csak a nők mintáján számított átlagokkal súlyozódnak. A véletlen erdő esetében a bérelőrejelzést is csak a női almintán számszerűsítjük /4/, emiatt csak rájuk értelmezhető a (férfiaknál nulla) bérstruktúrahatás, amelyet a továbbiakban csak őket tekintve vizsgálok. Egyéni szinten OLS-sel, valamint véletlen erdővel is meghatározom, hogy a férfi és a női bérfüggvény alapján mekkora lenne a nők bére; a két bérfüggvény

eredménye közötti különbség lesz az egyéni szinten számított bérstruktúrahatás. Amennyiben ennek értéke pozitív, úgy a nők alulárazottak a férfiakhoz képest, ha negatív, akkor pedig jobban kellene keresniük egy azonos adottságú férfinál.

Az egyéni szintű bérstruktúrahatás két módszer szerinti összehasonlításához első lépésben az átlagokat és az eloszlást veszem górcső alá. Ezt követően összevetem, hogy a bérfüggvénybecslésekhez használt inputváltozók szerint hogyan alakulnak az átlagos bérstruktúrahatások. Ez az összehasonlítás korlátozottan ugyan, de lehetővé teszi annak vizsgálatát, hogy a két eljárás szerint melyik változó milyen mértékben befolyásolja a bérstruktúrahatás alakulását.

2. Adatbázis

A bérkülönbségek becsléséhez a Nemzeti Munkaügyi Hivatal 2016-ra vonatkozó bértarifaadatait használtam. Ez az adatbázis a megfigyelt munkavállalók béréről, személyes és vállalati jellemzőikről tartalmaz információkat. Eredményváltozóként az alapilletmény mellett a havi szintre vetített bónuszokat és jutalékokat is magába foglaló havi bruttó átlagkeresetek logaritmusát használtam. Az OLS- és a véletlenerdő-becslés inputváltozóit az 1. táblázat tartalmazza.

1. táblázat

A Blinder–Oaxaca-felbontáshoz használt inputváltozók
(Input variables used for Blinder–Oaxaca decomposition)

Inputváltozó	Megfigyelési egység
Életkor	Év
Szolgálati idő	Adott munkáltatónál eltöltött hónapok
Iskolai végzettség	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
Foglalkozás	1. és 2. szintű FEOR-kód

(A táblázat folytatása a következő oldalon)

(Folytatás)

Inputváltozó	Megfigyelési egység
Külföldi tulajdon aránya	1: 100% külföldi tulajdon 2: 50% feletti külföldi tulajdon 3: 50% alatti külföldi tulajdon 4: 0% külföldi tulajdon
Állami és önkormányzati tulajdon aránya	1: 100% állami és önkormányzati tulajdon 2: 50% feletti állami és önkormányzati tulajdon 3: 50% alatti állami és önkormányzati tulajdon 4: 0% állami és önkormányzati tulajdon
Vállalatméret	Foglalkoztatottak száma
Településtípus	1: főváros; 2: város; 3: egyéb
Régió	NUTS2 régiók – 7 kategória
Ágazat	Nemzetgazdasági ág (1. szintű TEÁOR-kód)*
Vállalati szintű kollektív szerződés	0: nem, 1: igen
Ágazati szintű kollektív szerződés	0: nem, 1: igen
Több munkáltatóra kiterjedő, de nem ágazati szintű kollektív szerződés	0: nem, 1: igen

* Az ágazatok közül a kevés számú megfigyelés miatt a közigazgatást, védelmet és kötelező társadalombiztosítást tartalmazó O ágazatot kizártam.

Megjegyzés. NUTS2 (Nomenclature des unités territoriales statistiques): Statisztikai célú területi egységek nómenklatúrája; FEOR: Foglalkozások egységes osztályozási rendszere; TEÁOR: Gazdasági tevékenységek egységes ágazati osztályozási rendszere.

Az adatbázisban rendelkezésre állt az életkorra és a potenciális munkatapasztalatra[5] vonatkozó információ is. Mivel a kettő erősen korrelál, számításaimhoz az életkort választottam; az OLS-becslés tartalmazza ennek négyzetes értékeit is.

A havi bérek torzító hatását elkerülve a vizsgálataimat kizárólag a teljes munkaidőben foglalkoztatottakra végeztem el, illetve kizártam az elemzésből a nem versenyszférában dolgozókat is, mivel ott a bérezési rendszert más tényezők határozzák meg. Továbbá *Lovász* [2013] példáját követve nem vettem be a mintába a 20 fő alatti vállalatokat, mert ezek béradatai nem megbízhatók (*Elek et al*. [2009]).

[5] A potenciális munkatapasztalat az életkor és a legmagasabb iskolai végzettség befejezéséhez köthető életkor különbségeként adódik.

A 2016. évi bértarifa-adatbázis a szűréseket követően összesen 110 003 megfigyelést tartalmaz, ebből 50 000 véletlenszerűen kiválasztott képezi a tanulómintát, a többi a tesztmintában szerepel. A 2016. évi nyers bérkülönbség a bruttó havi átlagkeresetek logaritmusa alapján az egész mintában 0,1534. A nők és férfiak átlagos bérének logaritmusait, illetve a nyers bérkülönbségeket a tanuló- és a tesztmintán a 2. táblázat mutatja.

2. táblázat

A nők és a férfiak bérének leíró statisztikái a tanuló- és a tesztmintán
(Descriptive statistics of wages for men and women, by training and test samples)

Leíró statisztika	Tanulóminta			Tesztminta		
	Nők	Férfiak	*Összes*	Nők	Férfiak	*Összes*
Elemszám	19 824	30 176	*50 000*	23 623	36 380	*60 003*
Átlag	12,3343	12,5059	*12,4379*	12,3476	12,4979	*12,4387*
Szórás	0,5490	0,6325	*0,6066*	0,5561	0,6298	*0,6063*
Minimum	10,6685	11,4631	*10,6685*	11,4284	10,7195	*10,7195*
Maximum	16,1108	16,1131	*16,1131*	16,1181	16,3155	*16,3155*
Nyers különbség	0,1716			0,1503		

A minta férfi-női összetétele bár nem 50-50 százalékban oszlik meg, kellően nagyszámú megfigyelést tartalmaz mindkét csoport. A 2. táblázat adataiból kitűnik, hogy mindkét mintában a nők átlagbére alacsonyabb a férfiakénál, valamint a nyers bérkülönbség az átlagbérek különbségét ragadja meg, amely pozitív. Ezeket az átlagos bérkülönbségeket vizsgálom tovább a Blinder–Oaxaca-dekompozíció segítségével.

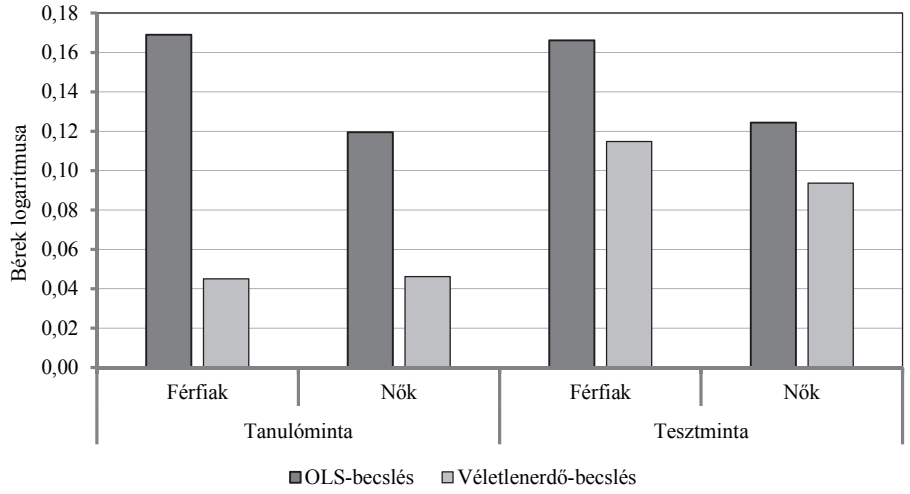
3. Eredmények

Az eredményeket három részletben tárgyalom: a 3.1. alfejezetben az OLS- és a véletlenerdő-becsléssel kapott bérfüggvények statisztikai tulajdonságait vizsgálom; a 3.2. alfejezetben a Blinder–Oaxaca-dekompozíció két módszerének eredményeit vetem össze; végül a 3.3. és a 3.4. alfejezetben a két eljárás egyéni szinten számított bérstruktúrahatásait hasonlítom össze leíró statisztikai módszerekkel.

3.1. A férfiak és a nők bérfüggvényeinek összehasonlítása

Az OLS- és a véletlenerdő-eljárás teljesítőképességének összehasonlításához MSE-t (mean squared error – átlagos négyzetes hiba) használtam; az eredményt az 1. ábra mutatja.

1. ábra. Átlagos négyzetes eltérés a véletlenerdő- és az OLS-becslés szerint a tanuló- és a tesztmintán, 2016
(Mean squared error for random forest and OLS estimations, by training and test samples, 2016)

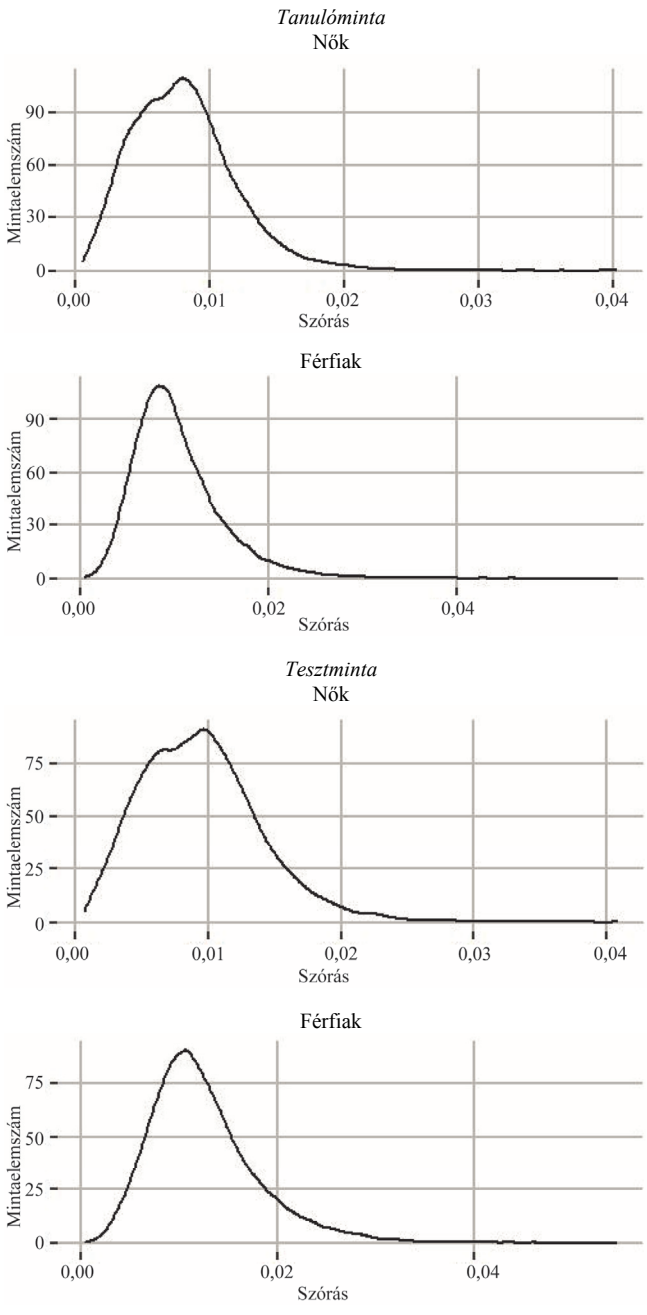


Az 1. ábrán látható, hogy a véletlenerdő-becslés MSE-értéke jelentősen kisebb, mint az OLS-é. Ez egyrészt megfelel az általános várakozásoknak egy nemparametrikus módszernél, másrészt igazolja, hogy lehetnek nemlineáris kapcsolatok az inputváltozók között. A nőket tekintve az MSE-érték kisebb mind a tanuló-, mind a tesztmintán, ennek oka az lehet, hogy a nők bére kevésbé szóródik a két mintában. (Lásd a 2. táblázatot.)

A véletlen erdőnél a tesztmintán megnő a hiba a tanulómintához képes, de még így is elmarad az OLS-becslés hibájától. Ez alátámasztja, hogy a véletlen erdő jobb előrejelzési képességgel rendelkezik ebben az esetben, mint a hagyományos OLS. Azonban a hiba növekedése arra utal, hogy a tanulóminta egyes sajátosságaira jobban „rátanult" a véletlen erdő, így az általánosítási képessége romlott a tesztmintán.

Lényeges megjegyezni, hogy a véletlenerdő-becslés – szemben az OLS-sel – ugyanazon a megfigyeléshalmazon nem adja mindig ugyanazt az eredményt. Ennek magyarázata, hogy az erdőben levő fák létrehozásánál a véletlen fontos szerepet játszik mind az alminták meghatározásánál, mind a vágásoknál a korlátozott számú változó kiválasztásánál. Amiatt, hogy az erdők változhatnak, megvizsgáltam a kapott eredmények robusztusságát 100 véletlen erdő létrehozásával, melyekhez az 1. fejezetben leírt paraméterbeállításokat használtam. Az adott személyhez tartozó becslés mindig a saját neméhez tartozó becslőfüggvénnyel készült. A 100 erdővel 100 bérbecslést adtam minden egyes megfigyelésre, majd megvizsgáltam, hogy az ezekhez tartozó szórások mekkorák. A tanuló- és a tesztmintára vonatkozó eredményeket a 2. ábra mutatja, ahol a vízszintes tengelyen az egyes megfigyelésekhez tartozó becslések szórása, a függőlegesen pedig az ahhoz tartozó mintaelemszám található.

2. ábra. A szórások megoszlása a 100 véletlen erdő szerint a tanuló- és a tesztmintán
(Distribution of standard deviations of 100 random forest estimations, by training and test samples)

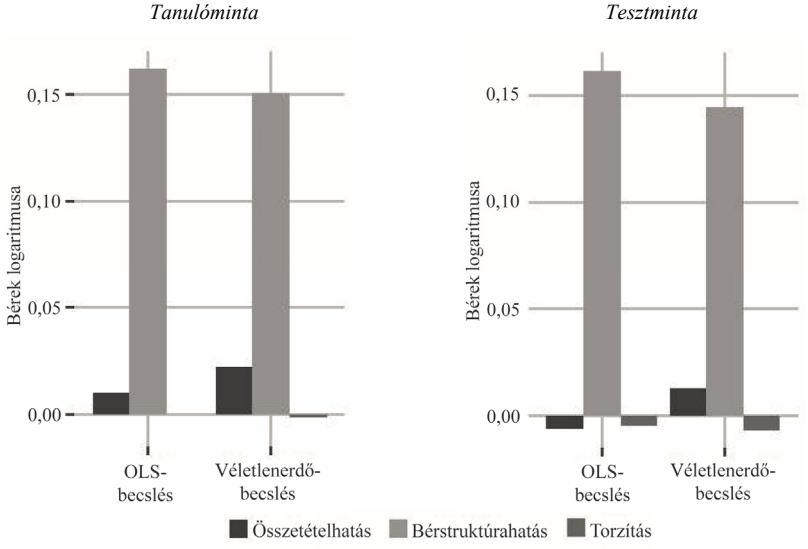


A 2. ábra azt mutatja, hogy a véletlenerdő-eljárásnál vannak kiugró eltérések, azonban az esetek többségében alacsony a szórásuk, ami arra mutat, hogy az így készült becslések robusztusak. A következőkben csak egy véletlen erdővel készült becslés eredményét tárgyalom.

3.2. Blinder–Oaxaca-felbontás

A 3. ábra a nyers bérkülönbséget és az 1.2. alfejezetben ismertetett Blinder–Oaxaca-felbontást mutatja a véletlen erdő és az OLS esetében.

3. ábra. A Blinder–Oaxaca-felbontás alakulása a véletlenerdő- és az OLS-becslés szerint a tanuló- és a tesztmintán
(Blinder–Oaxaca decomposition based on random forest and OLS estimations, by training and test samples)



Összességében az egyes becslési eljárásoknál hasonló eredményeket látunk a tanuló- és a tesztmintán. Az OLS-nél a tanulómintán a torzítás mértéke a definíció szerint nulla, míg a véletlen erdőnél ez a tag nem nulla, azonban – ahogyan a 3. ábrán látható – elhanyagolhatóan kicsi. A tesztmintákon a torzítás már nagyobb, de még így sem jelentős. Ez igazolja, hogy a véletlen erdő is alkalmas a Blinder–Oaxaca-dekompozíció létrehozásához.

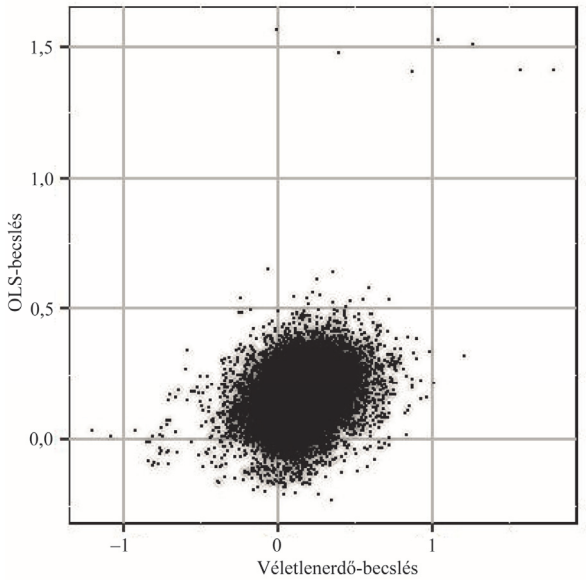
A 3. ábra jól szemlélteti, hogy az összetételhatás abszolút nagysága mind a tanuló-, mind a tesztmintán nagyobb a véletlenerdő-eljárásnál. Ez okozza, hogy a bérstruktúrahatás viszont kisebb a véletlen erdőnél, de így is közel azonos a két mód-

szer esetében. Emellett a bérstruktúrahatás jelentősen meghaladja az összetételhatás nagyságát mindkét metódusnál, továbbá minden esetben pozitív, ami arra utal, hogy ha a férfiak bérfüggvényével áraznák a nőket, akkor a nőknek többet kellene keresniük.

3.3. Egyéni szintű bérstruktúrahatások

A 4. ábra az egyéni szintű bérstruktúrahatás (a bérek logaritmusában kifejezett) szóródását mutatja a tanulómintán az OLS-sel és a véletlen erdővel készült becslések esetében.

4. ábra. Az egyéni szintű bérstruktúrahatás szóródása a véletlenerdő- és az OLS-becslés szerint a tanulómintán
(Dispersion of wage structure effects in the training sample, by random forest and OLS estimations)



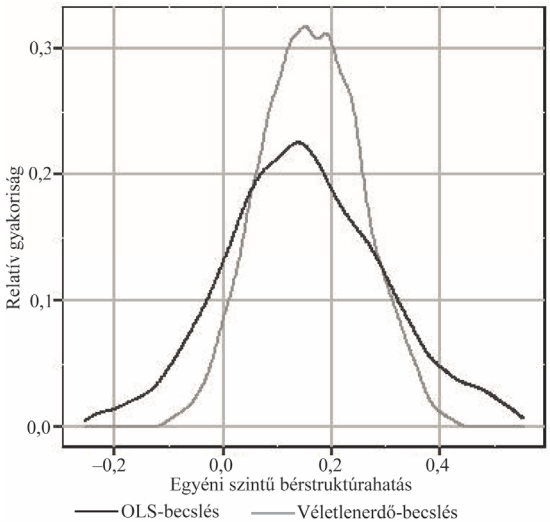
A 4. ábrán látható, hogy a legtöbb esetben a bérstruktúrahatás a véletlen erdőnél –1 és 1 között, az OLS-regressziónál ennél szűkebb intervallumban szóródik. Emellett mindkét módszernél adódtak kiugró értékek, melyeket eltávolítottam az adatbázisból.[6] Így a tanulóminta 19 824 megfigyeléséből 19 124 maradt, amelyek további elemzéseim alapjául szolgáltak.

Az eltávolított értékek nélkül a két módszer szerinti egyéni bérstruktúrahatás eloszlása normális, és az átlagok hasonlók (lásd az 5. ábrát), de a véletlen erdő eseté-

[6] Azokat az értékeket távolítottam el, amelyek kívül estek a Q_{25} –1,5 * *IQR* és a Q_{75} + 1,5 * *IQR* tartományon, ahol az *IQR* a középső 50 százalék terjedelme.

ben az eloszlás laposabb: itt az egyéni bérstruktúrahatás jobban szóródik. Az eloszlások összehasonlítására elvégzett Kolmogorov–Szmirnov-teszt[7] szerint a két eloszlás szignifikánsan különbözik egymástól. A 3. táblázat a két módszer szerinti bérstruktúrahatások leíró statisztikáit tartalmazza a nők esetében a tanuló- és a kiugró értékek nélküli mintán. Az OLS- és a véletlenerdő-eljárással készített becslések átlagát a 4. táblázat mutatja.

5. ábra. Az egyéni szintű bérstruktúrahatás eloszlása a véletlenerdő- és az OLS-becslés szerint a tanulómintán
(Distribution of individual wage structure effects in the training sample, by random forest and OLS estimations)



3. táblázat

Leíró statisztikák a véletlenerdő- és az OLS-becslés szerint a tanuló- és a kiugró értékek nélküli mintán
(Descriptive statistics of the training and outlier-free samples, by random forest and OLS estimations)

Leíró statisztika	Tanulóminta		Kiugró értékek nélküli minta	
	OLS-becslés	Véletlenerdő-becslés	OLS-becslés	Véletlenerdő-becslés
Átlag	0,1617	0,1504	0,1616	0,1496
Szórás	0,1014	0,1705	0,0932	0,1470
Minimum	–0,2295	–1,2031	–0,1071	–0,2563
Maximum	1,5647	1,7879	0,4296	0,5534
Korreláció	0,3081		0,2767	

[7] A kétoldalú Kolmogorov–Szmirnov-teszt értéke 0,13057, ahol a *p*-érték kisebb, mint $2{,}2*10^{-16}$. Vagyis az eloszlások egyezőségére vonatkozó H_0 hipotézist elvetem.

4. táblázat

A bérek logaritmusával kifejezett előrejelzések átlagai a véletlenerdő- és az OLS-becslés szerint a kiugró értékek nélküli mintán
(Averages of wage estimations [in log wages] in the outlier-free sample, by random forest and OLS estimations)

Bérfüggvény	OLS-becslés	Véletlenerdő-becslés	Különbség
Férfi	12,4870	12,4740	0,0130
Női	12,3253	12,3244	0,0009
Különbség	0,1616	0,1496	

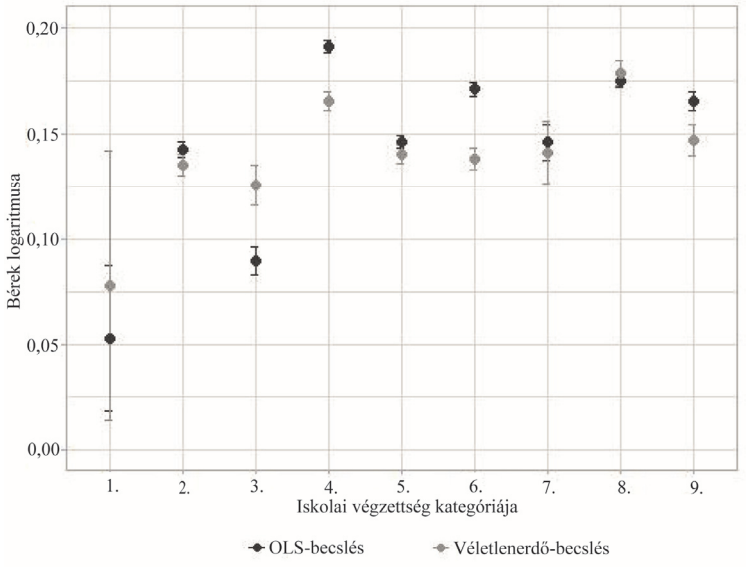
A 4. táblázat összefoglalóan mutatja, hogy a férfiak modelljével készített előrejelzés mindkét módszer esetében magasabb, ami nem meglepő, hiszen a férfiak bérei átlagosan magasabbak, mint a nőké. Az OLS-sel készült előrejelzések átlagai magasabbak, mint a véletlen erdőé. A 3. táblázat alapján pedig az látszik, hogy bár az egyéni bérstruktúrahatások átlaga közel azonos, a korreláció viszonylag alacsony. Vagyis a véletlenerdő- és az OLS-becslés szerinti egyéni bérstruktúrahatások között nagyok a különbségek. Annak magyarázatára, hogy mi okozza a két módszer szerinti eltérést, a következőkben megvizsgálom, hogy az egyes változók alapján mekkorák az általos bérstruktúrahatások.

3.4. Egyéni szintű bérstruktúrahatások változónként

A következőkben a változónként számított egyéni szintű bérstruktúrahatásokat hasonlítom össze. A cél annak vizsgálata, hogy egy változón belül az egyes kategóriákhoz tartozóan mekkorák a különbségek a véletlenerdő- és az OLS-becsléssel számított átlagos bérstruktúrahatások között. A 3. Függelékben a változók szerinti megoszlások találhatók, a 4. Függelékben pedig a leíró statisztikákat alátámasztó – a kétféle módszerrel kapott eredmények különbségére számított – páros ANOVA-tesztek eredményeit mutatom be.

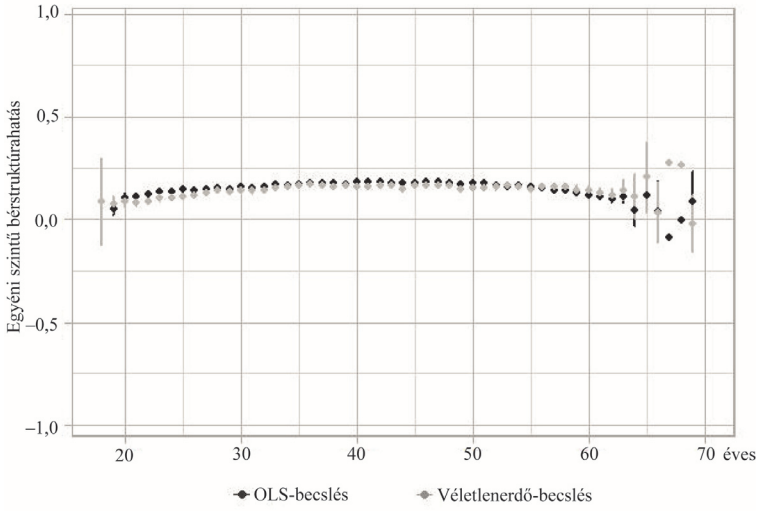
Az iskolai végzettségnél a két módszer szerinti becsült átlagok erősen együttmozognak. (Lásd a 6. ábrát.) A bérstruktúrahatás az általános iskolát nem befejezetteknél (1. kategória) a legkisebb, azonban elég kevés megfigyelés esik ebbe a kategóriába, és az adatok is erősen szóródnak. A következő legalacsonyabb bérstruktúrahatással a 3., szakiskolai végzettség kategória rendelkezik, azonban itt legnagyobb a különbség a két eljárás között. Az első érettségit adó – szakközépiskolát jelölő – 5. kategória felett az átlagok ugyanolyan sávban szóródnak. Figyelemre méltó, hogy a véletlen erdő az OLS-sel összehasonlítva nagyobb bérstruktúrahatást becsül a főiskolát (8. kategória) végzetteknél, míg az egyetem (9. kategória) esetében ez pont fordítva van.

6. ábra. Átlagos bérstruktúrahatások és azok
95 százalékos konfidenciaintervallumai iskolai végzettség szerint
(Average wage structure effects and their 95% confidence intervals by level of education)



Megjegyzés. Az iskolai végzettség kategóriáit lásd az 1. táblázatban.

7. ábra. Átlagos bérstruktúrahatások és azok
95 százalékos konfidenciaintervallumai különböző életkorok szerint
(Average wage structure effects and their 95% confidence intervals by age)

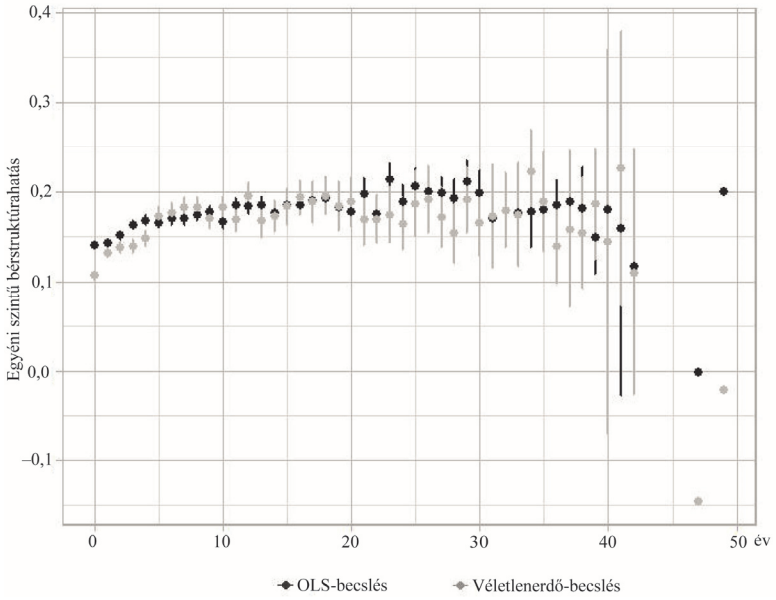


Megjegyzés. Az életkor és az egyéni szintű bérstruktúrahatások csak 18 és 70 éves kor között vannak feltüntetve.

A 20 év alattiak és a 61 év felettiek tekintetében erősen szóródnak az átlagok, amit az magyaráz, hogy ezekhez a korcsoportokhoz általában kevesebb mint 100 megfigyelés tartozik, így ezek az eredmények nem tekinthetők statisztikailag erősnek. A 20 és 61 év közöttiekről elmondható, hogy az életkor növekedésével a bérstruktúrahatás kezdetben növekszik, majd 50 éves kor felett csökkenni kezd. A 7. ábrán az is látható, hogy 20 és 41 éves kor között az OLS legalább akkora bérstruktúrahatást eredményez, mint a véletlen erdő, majd ezt követően ez utóbbi nagyobb hatást becsül. Azonban a becslési eredmények így is meglehetősen közel állnak egymáshoz. Az itt kapott négyzetesen emelkedő bérstruktúrahatás egybevág *Gábor* [2008] eredményeivel, aki kimutatta, hogy a férfiak bérfüggvénye a nőkéhez képest magasabbról indul, és kezdeti erősebb emelkedése miatt konkávabb.

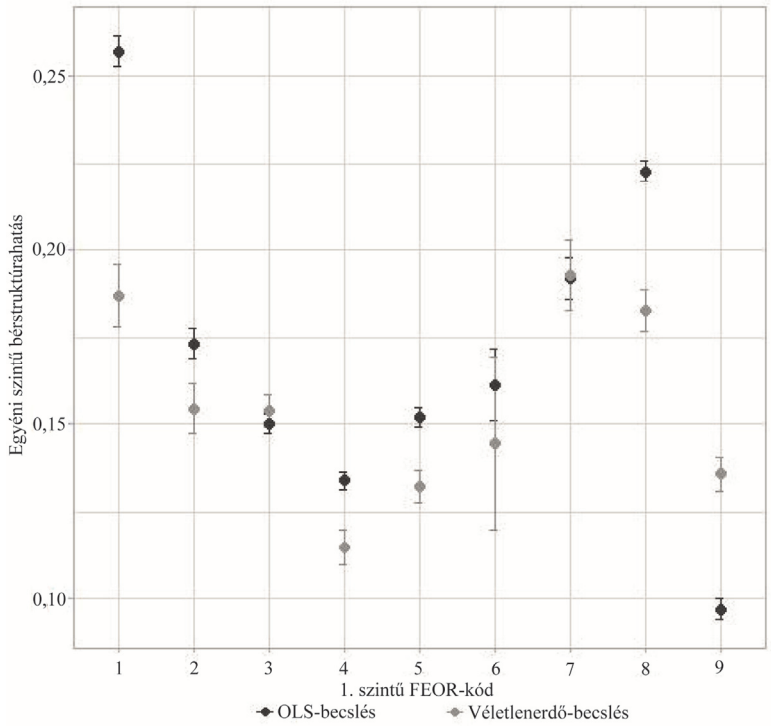
A Blinder–Oaxaca-dekompozícióban a szolgálati időt hónapokban határoztam meg, azonban a 8. ábrán már – a könnyebb értelmezhetőség miatt – években tüntettem fel. Ez a változó rámutat arra, hogy a vállalathoz való belépéskor sem nulla a bérstruktúrahatás, vagyis a nők feltételezhetően, már kezdetben is kevesebb fizetést kapnak, mint az ugyanolyan adottságokkal rendelkező férfi kollégáik. A vállalatnál eltöltött évek során pedig a bérstruktúrahatás egyre inkább emelkedik, majd 10 év után tulajdonképpen stagnál. Az eredményekből arra is következtethetünk, hogy az egyéni bérstruktúrahatások a szolgálati idő emelkedésével egyre nagyobb szóródást mutatnak, ahogyan az adott szolgálati időhöz tartozóan egyre csökkennek a megfigyelésszámok, továbbá a szolgálati idő előrehaladtával egyre inkább eltér a két becslés eredménye.

8. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai szolgálati évek szerint
(Average wage structure effects and their 95% confidence intervals by year of service)



A 9. ábra az 1. szintű, a 10. ábra a 2. szintű FEOR-kódokhoz tartozó átlagos bérstruktúrahatásokat tartalmazza. A Blinder–Oaxaca-dekompozícióban 2. szintű foglalkozási kódok szerepelnek.

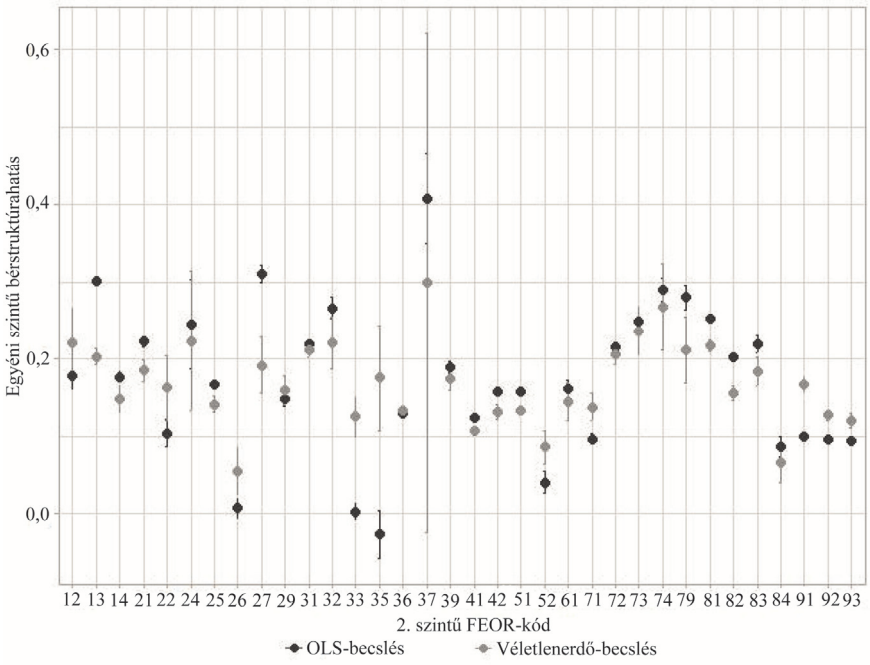
9. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai az 1. szintű FEOR-kód szerint
(Average wage structure effects and their 95% confidence intervals by 1-digit FEOR [Hungarian Standard Classification of Occupation] code)



Megjegyzés. Itt és a 10. ábránál a FEOR struktúráját lásd a https://www.ksh.hu/feor_menu honlapon.

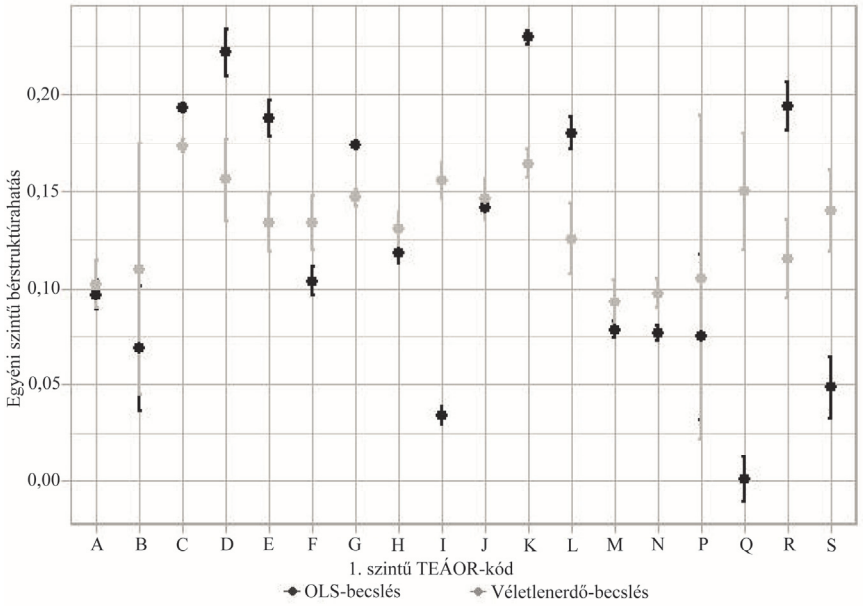
A 9. ábra alapján az OLS-sel és a véletlen erdővel számított átlagos bérstruktúrahatások erősen együttmozognak. Ez alapján a gazdasági, igazgatási, érdekképviseleti vezetőket és törvényhozókat magába foglaló 1-es és a szakképzettséget nem igénylő foglalkozásokat tömörítő 9-es kategóriában legnagyobb az átlagok eltérése a két módszertan között. A legkisebb bérstruktúrahatás mindkét eljárás szerint az irodai és ügyviteli (ügyfélkapcsolati) foglalkozásokat tömörítő 4-es kategóriában van. A kétjegyű foglalkozáskódokat tartalmazó 10. ábra arra hívja fel a figyelmet, hogy az egyjegyű kategóriákon belül a foglalkozások között nagyok lehetnek az eltérések a két metódus alapján.

10. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai a 2. szintű FEOR-kód szerint
(Average wage structure effects and their 95% confidence intervals by 2-digit FEOR code)



A foglalkozások megkülönböztetése mellett az ágazatoknál látszanak a legnagyobb módszertani különbségek. Az A és a B ágazatnál a véletlen erdő közel azonos bérstruktúrahatást, míg az OLS a B ágazatnál jelentősen alacsonyabbat becsül. A C–J ágazatoknál a véletlen erdő eredményei hasonló értékeket mutatnak, az OLS-éi pedig jobban szóródnak. Az N és P ágazatok azok, amelyek egy részét a költségvetési szektorhoz sorolják. A saját mintámban csak a versenyszféra vállalatai vannak, azonban a költségvetési szférában alkalmazott bérszint hathat a magánszektorra (*Telegdy* [2013]), és végső soron a nemek közötti béregyenlőtlenségre is. Emiatt a többi ágazattal összehasonlítva itt alacsonyabbak a bérstruktúrahatások. Az egészségügyet magába foglaló Q ágazatban a két módszer szerinti bérstruktúrahatások erősen eltérnek, így erről a szintén jelentős részben a közszférához tartozó ágazatról nehéz egyértelmű megállapítást tenni. A különbség oka az lehet, hogy ehhez az ágazathoz kevés számú megfigyelés tartozik.

11. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai az 1. szintű TEÁOR-kód szerint
(Average wage structure effects and their 95% confidence intervals by the first level of NACE Rev. 2)



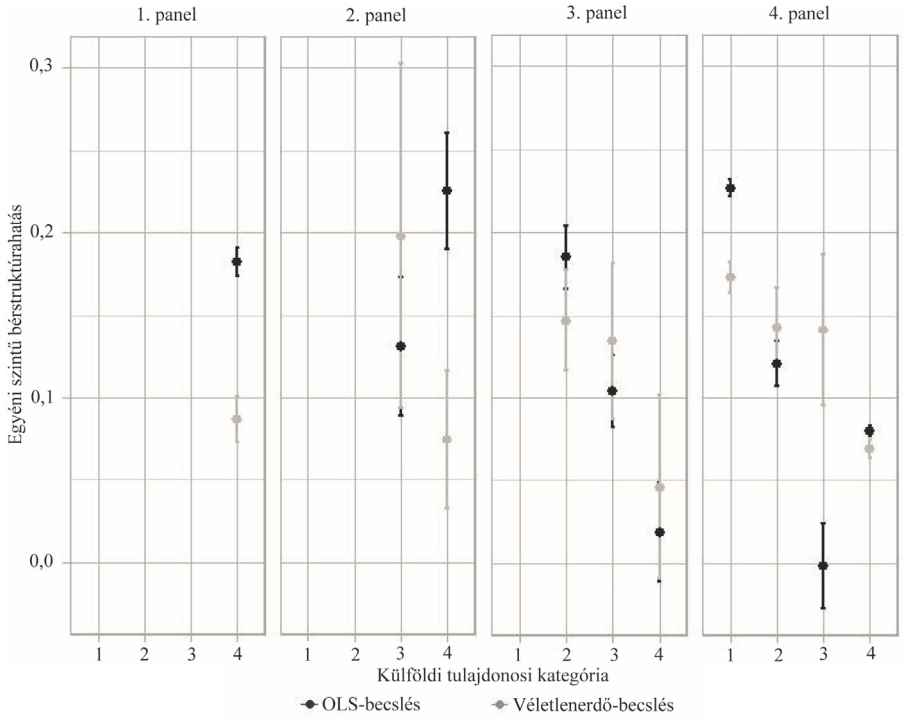
Megjegyzés. A TEÁOR struktúráját lásd a https://www.ksh.hu/docs/osztalyozasok/teaor/teaor08_struktura_2018_08_01.pdf honlapon.

A 12. ábrán a kétféle módszer szerint számított átlagos bérstruktúrahatást az állami és a külföldi tulajdonú vállalatok arányában vetem össze. Az 1. panelen az állami tulajdonú vállalatok 100 százalékos, a 2. panelen többségi, a 3. panelen kevesebb mint 50 százalékos, a 4. panelen 0 százalékos aránnyal szerepelnek. Az egyes paneleken a külföldi tulajdon szerinti kategóriák láthatók a vízszintes tengelyen.[8]

Az 1. és 2. panelen közel azonos a bérstruktúrahatás átlagainak eltérése. A 3. panelen az látszik, hogy a két eljárás egy irányba mutat, tehát minél kisebb a külföldi tulajdonosi hányad, annál inkább csökken a bérstruktúrahatás. A 4. panelen a bérstruktúrahatás, ha van bármekkora külföldi tulajdon, akkor közel azonos a véletlen erdőnél. Ezzel szemben az OLS itt is a külföldi tulajdon mérséklődésével együtt a bérstruktúrahatás csökkenését mutatja. Ez arra vezethető vissza, hogy az OLS-becslésben nem szerepeltek a két változó keresztszorzatai, a véletlenerdő-módszer azonban képes volt „rátanulni" erre a mintára. A 0 százalék külföldi, illetve állami tulajdonú vállalatoknál található a legtöbb megfigyelés, és ezeknél a két módszertan szerinti átlagos bérstruktúrahatások között kicsi a különbség.

[8] Fontos megjegyezni, hogy az OLS-sel készült bérfüggvényekben a két változó keresztszorzata nem szerepelt, így ez a fajta összevetés ugyan informatív, de nem igazán korrekt.

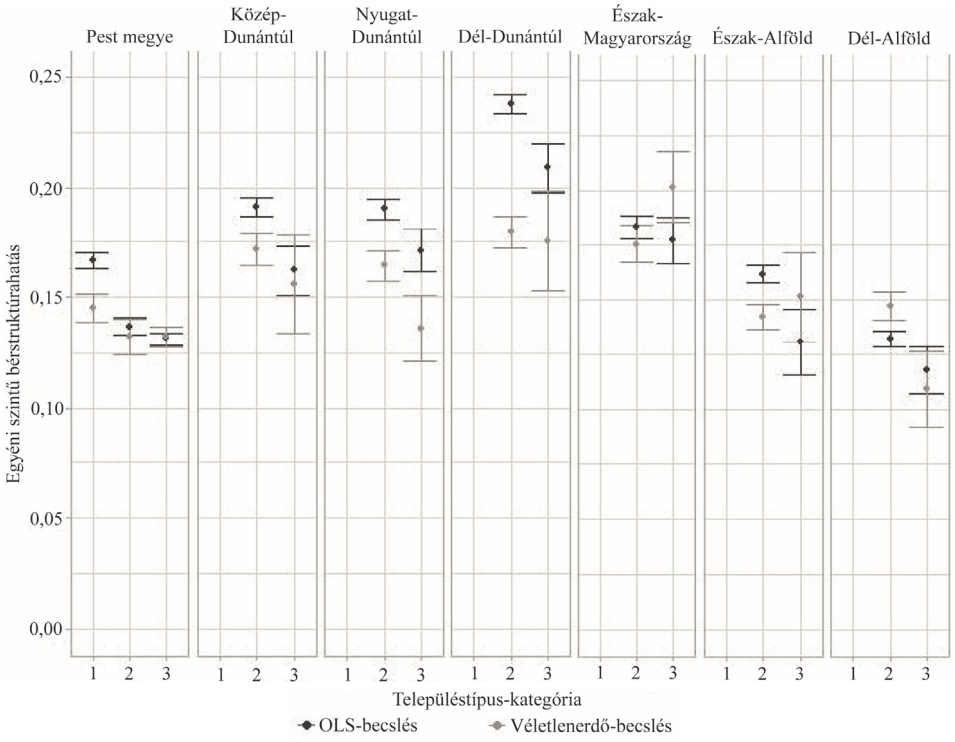
12. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai állami és külföldi tulajdonú vállalatok szerint
(Average wage structure effects and their 95% confidence intervals by state- and foreign-owned companies)



Megjegyzés. A tulajdonosi kategóriákat lásd az 1. táblázatban.

A tulajdonhoz hasonló módon vontam össze a régióra és a településtípusra vonatkozó adatokat a 13. ábrán. A régiók és a településtípusok összevetéséből az látszik, hogy Pest megyében alacsony a városra és az egyéb településtípusokra vonatkozó bérstruktúrahatás, és a két módszer szerinti átlagok közel állnak egymáshoz. A többi régiót összehasonlítva, az Alföldön a legalacsonyabb a bérstruktúrahatás mind a város, mind az egyéb településtípusok tekintetében. A budapesti bérstruktúrahatás magasabb, mint a többi Pest megyei településtípusban, azonban országos összehasonlításban alacsonyabb, mint az egész Dunántúlon és az Észak-Magyarországon mért, városokra vonatkozó átlagos bérstruktúrahatás.

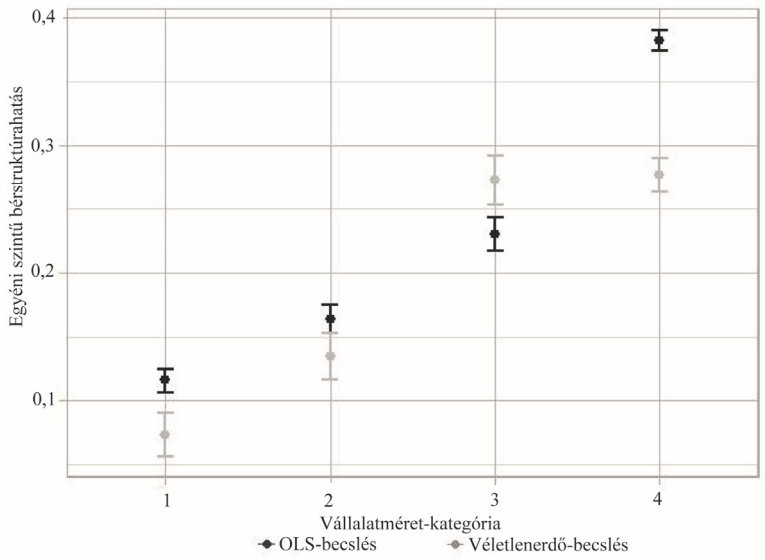
13. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai a magyar régiókban különböző településtípusok szerint
(Average wage structure effects and their 95% confidence intervals in the Hungarian regions, by type of settlement)



Megjegyzés. Településkategória: 1: főváros, 2: város, 3: egyéb.

A vállalatméretet a folytonos létszám változó jeleníti meg a bérfüggvény-becslésekben. A 14. ábrán viszont összevontabb kategóriákat tüntettem fel. Az OLS-sel számított bérstruktúrahatás erősebb, ha a vállalatméret nagyobb. Az 1–3 vállalat-kategóriákat tekintve ez a véletlen erdő esetében is igaz, azonban a 4-es kategóriában a bérstruktúrahatás nagyságrendileg megegyezik a 3-as kategóriában számítottal. Az eltérést az okozza, hogy az OLS-ben a létszám és a bér között lineáris kapcsolatot tételeztem fel.

14. ábra. Átlagos bérstruktúrahatások és azok 95 százalékos konfidenciaintervallumai vállalatméret szerint
(Average wage structure effects and their 95% confidence intervals by company size)



Megjegyzés. Vállalatméret-kategória: 1: 20–49 fő, 2: 50–149 fő, 3: 150–499 fő, 4: 500 fő és afeletti.

4. Összefoglalás

Tanulmányomban magyar adatokon hasonlítottam össze az OLS- és a véletlenerdő-becsléssel felírt nemek szerinti bérfüggvények Blinder–Oaxaca-dekompozíciós eredményeit. Célom az volt, hogy pontosabb képet kaphassak a bérdiszkriminációról, illetve annak mértékéről a módszertan függvényében.

Az adatokban rejlő nemlinearitások miatt a véletlenerdő-algoritmus bérelőrejelzései pontosabbak, mint az OLS-sel készültek. Az összehasonlítás azonban korántsem igazságos, ugyanis az OLS-regresszió felírásánál a leggyakrabban használt függvényformát alkalmaztam (*Blau–Kahn* [2017], *Leythienne–Ronkowski* [2018]), amelyben az életkor négyzetén kívül egyéb nemlinearitást nem vettem figyelembe. A véletlenerdő-eljárás képes „rátanulni” az adatokban rejlő összefüggésekre. További kutatási kérdés, hogy az eredmények közötti különbség mennyire lenne jelentős abban az esetben, ha az OLS-becslés kereszthatásokat is tartalmazna.

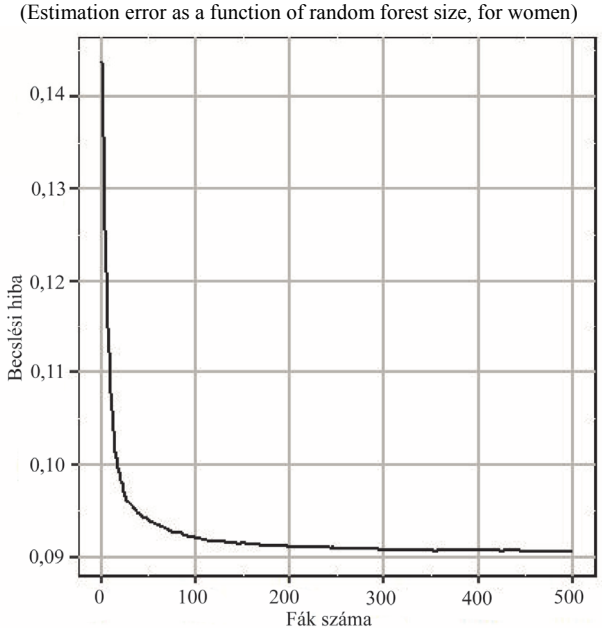
A véletlen erdővel készített Blinder–Oaxaca-dekompozíció esetében megjelenő torzítás még elfogadható mértékű. Tehát ez a módszertani váltás nem hat jelentősen a

torzítottságra, és vélhetően pontosabban becsli a férfi és női bérfüggvényeket. A dekompozíción belül az összetételhatás nagysága abszolút értelemben nagyobb a véletlen erdő esetében, ami végső soron kisebb bérstruktúrahatást eredményez. A kétféle módszerrel számított átlagos bérstruktúrahatás-eredmények így is közel esnek egymáshoz, az eloszlásuk azonban különböző, és a közöttük levő korreláció alacsony. Mindez azért következik be, mert az OLS- és a véletlenerdő-becslés különböző mértékben „árazza félre" a nőket a férfiakhoz képest. Abban az esetben, ha az adott változónál – mint például a foglalkozásnál – a különböző aggregáltsági szintű adatokat is összevetettem, akkor a magasabb szinten aggregált csoportnál a két módszer szerinti átlagos bérstruktúrahatások közelítettek egymáshoz. A két módszer szerinti átlagos bérstruktúrahatások közel azonos eredményeket adtak azokban a csoportokban, amelyek sok megfigyelést tartalmaztak. A leíró statisztikai elemzés alapján azonban nem emelhető ki egy változó sem, amely egyértelműen felelős lenne azért, hogy a kétféle eljárás szerinti bérstruktúrahatások eltérők. Ez egybevág *Weichselbaumer–Winter-Ebmer* [2005] eredményeivel, akik a nemek szerinti bérkülönbségek vizsgálatánál többféle módszertant is összevetettek, és arra a következtetésre jutottak, hogy a bérkülönbség-felbontás szempontjából a metódusok nem különösebben meghatározók, amelyek azonban eltérően értékelik a változók szignifikanciáját.

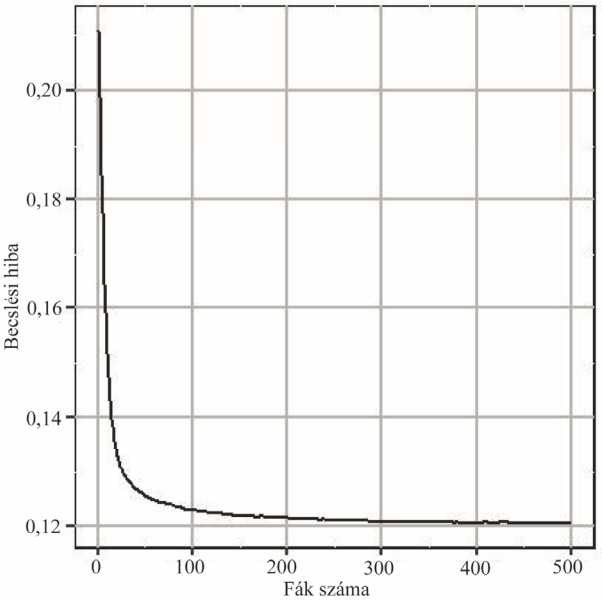
Összességében tehát átlagosan hasonló eredményt ad az OLS- és a véletlenerdő-becslés, azonban az egyéni szintű bérstruktúrahatás mértékénél és a változók fontosságának meghatározásánál már van jelentősége annak, hogy melyiket használjuk. Mivel a véletlen erdő esetében az egyedi változók hatását nehéz számszerűsíteni, így további kutatások alapjául szolgálhat a két eljárás összehasonlítása. Ezáltal lehetővé válik a regresszióban olyan nemlineáris kapcsolatok definiálása, amelyek javítják a bérfüggvények előrejelző-képességét, és az egyedi hatások értelmezésének lehetősége is megmarad.

1. Függelék

F1. ábra. Becslési hibanagyság az erdő méretének függvényében a nők mintájában
(Estimation error as a function of random forest size, for women)



F2. ábra. Becslési hibanagyság az erdő méretének függvényében a férfiak mintájában
(Estimation error as a function of random forest size, for men)

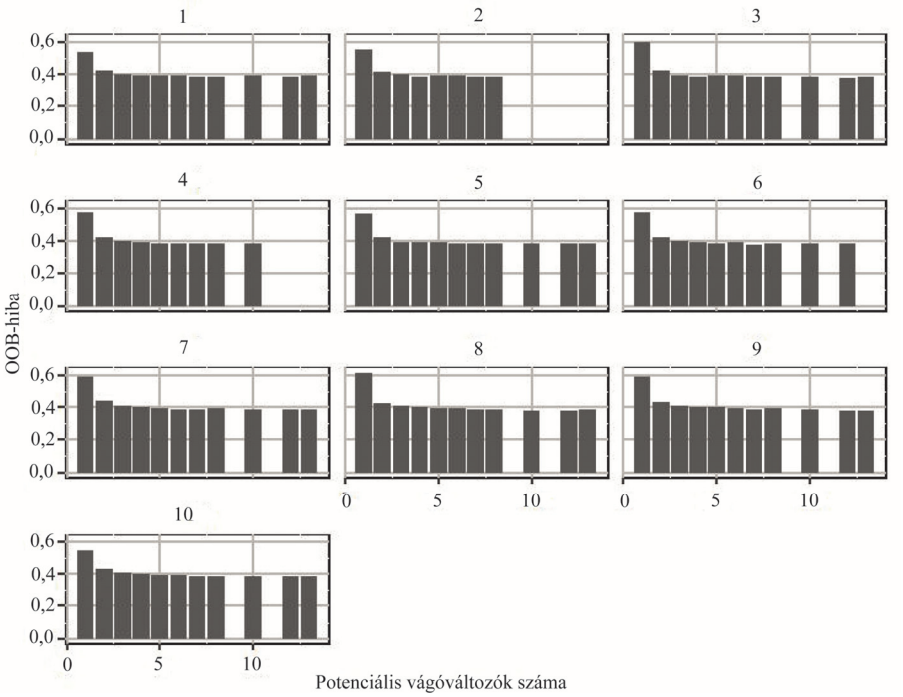


F1. táblázat

A paraméteroptimalizálás eredményei (tune.rfsrc paranccsal)
(Parameter optimisation results [by tune.rfsrc])

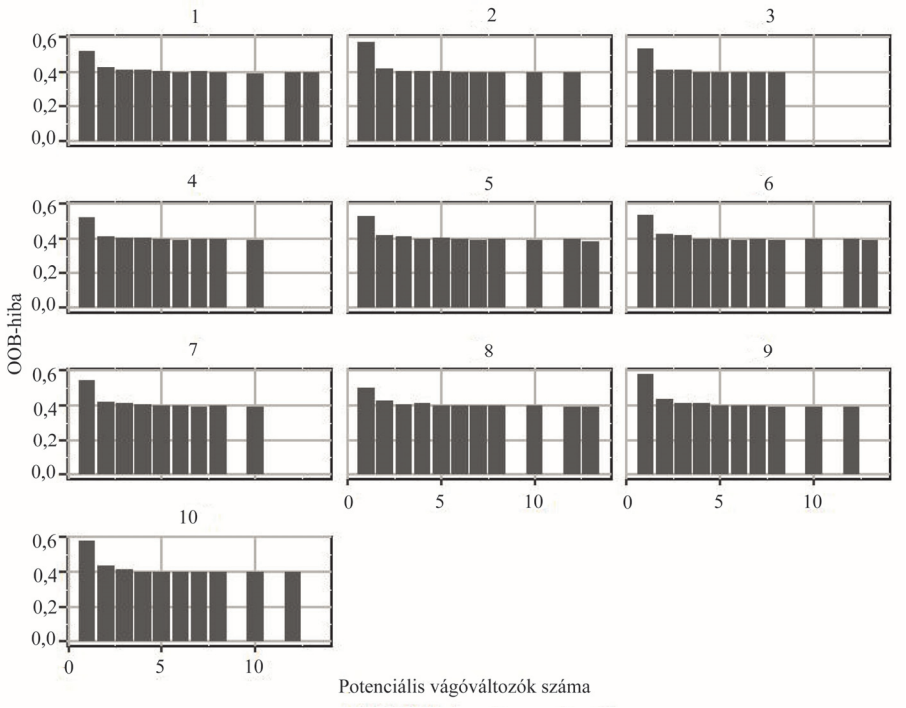
Paraméter	Nők	Férfiak
Mintanagyság	19 824	30 176
Fák száma	500	500
Levélméret	8	5
Átlagos levélméret	219,948	476,964
Potenciális vágóváltozók száma	12	13
Változók száma	13	13
Mintavétel	Visszatevés nélkül	Visszatevés nélkül
Bootstrap mintanagyság	1 671	2 290
Magyarázott variancia (%)	64,39	63,9
Hibanagyság	0,11	0,14

F3. ábra. A becslési (OOB-) hiba nagysága a levélméret és a potenciális vágóváltozók számának függvényében a nők mintájában
(OOB [out-of-bag] error in the subsample of women, by node size and the size of potential split variables)



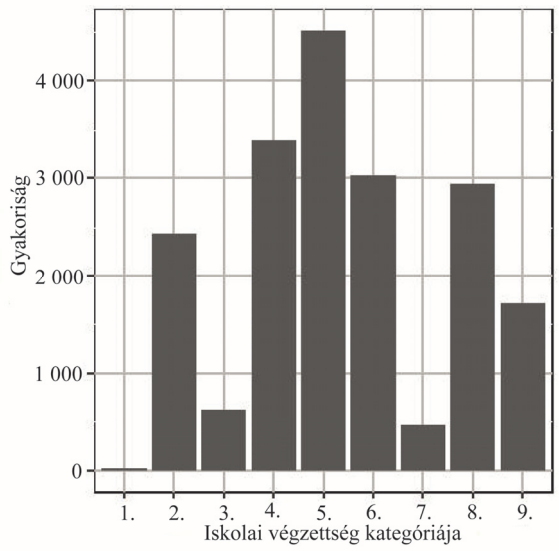
Megjegyzés. OOB (out-of-bag error): előrejelzési hiba. Itt és az F4. ábra esetén az egyes paneleken (1–10-ig) a levélnagyság, a horizontális tengelyen a potenciális vágóváltozók száma van feltüntetve.

F4. ábra. A becslési (OOB-) hiba nagysága a levélméret
és a potenciális vágóváltozók számának függvényében a férfiak mintájában
(OOB error in the subsample of men, by node size and the size of potential split variables)



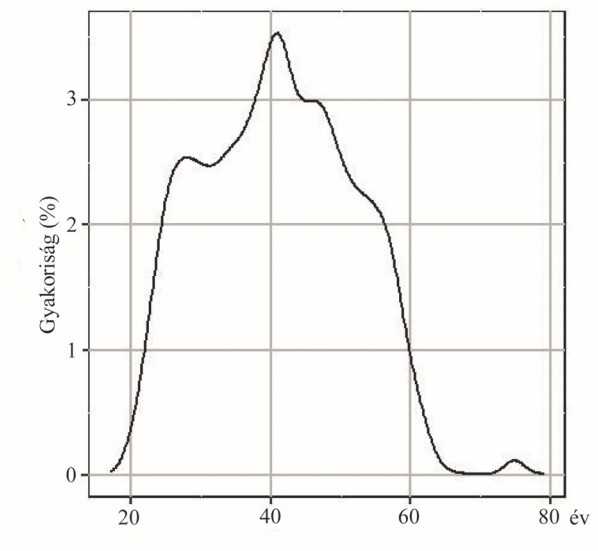
2. Függelék

F5. ábra. Az egyéni szintű bérstruktúrahatás megoszlása iskolai végzettség szerint
(Distribution of individual wage structure effects by level of education)

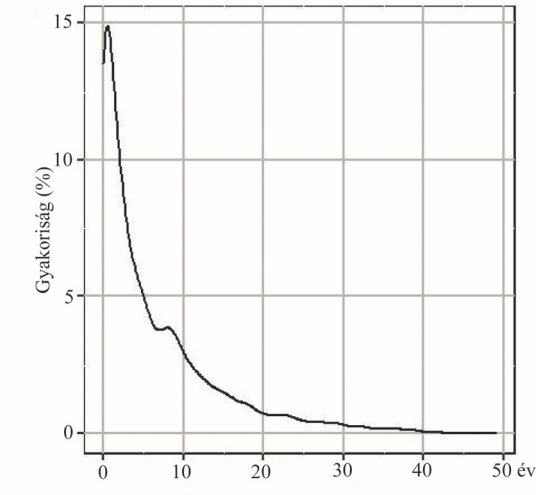


Megjegyzés. Az iskolai végzettség kategóriáit lásd az 1. táblázatban.

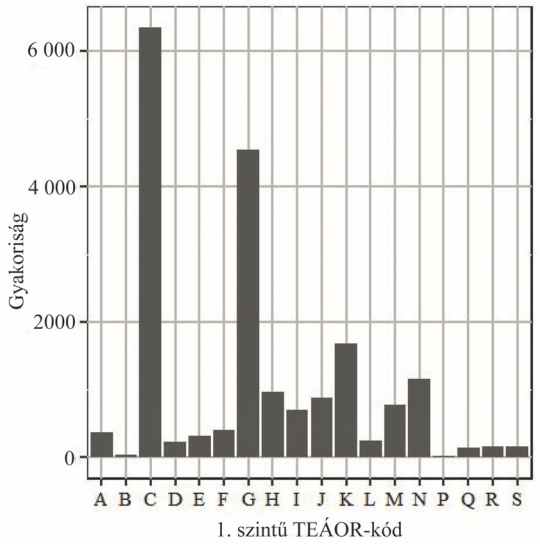
F6. ábra. Az egyéni szintű bérstruktúrahatás megoszlása életkor szerint
(Distribution of individual wage structure effects by age)



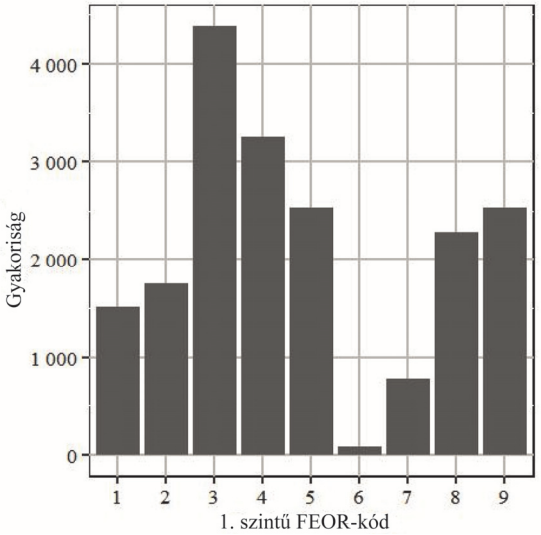
F7. ábra. Az egyéni szintű bérstruktúrahatás megoszlása szolgálati évek szerint
(Distribution of individual wage structure effects by year of service)



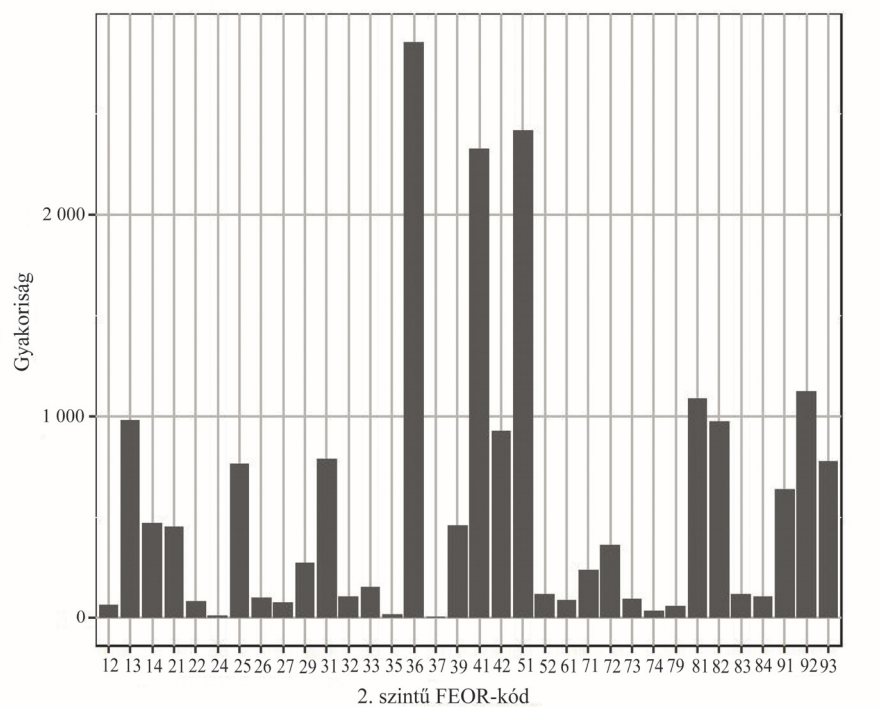
F8. ábra. Az egyéni szintű bérstruktúrahatás megoszlása az 1. szintű TEÁOR-kód szerint
(Distribution of individual wage structure effects by the first level of NACE Rev. 2)



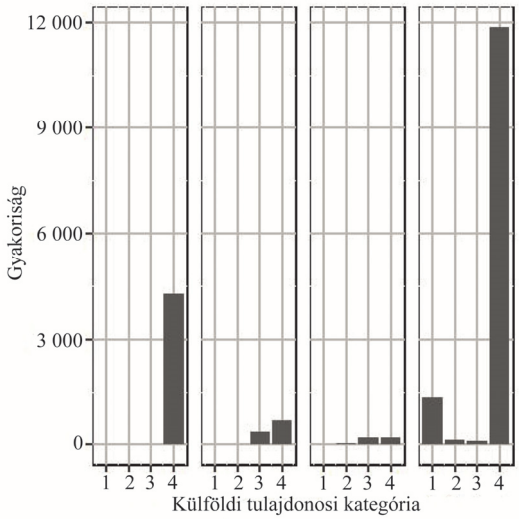
F9. ábra. Az egyéni szintű bérstruktúrahatás megoszlása az 1. szintű FEOR-kód szerint
(Distribution of individual wage structure effects by 1-digit FEOR code)



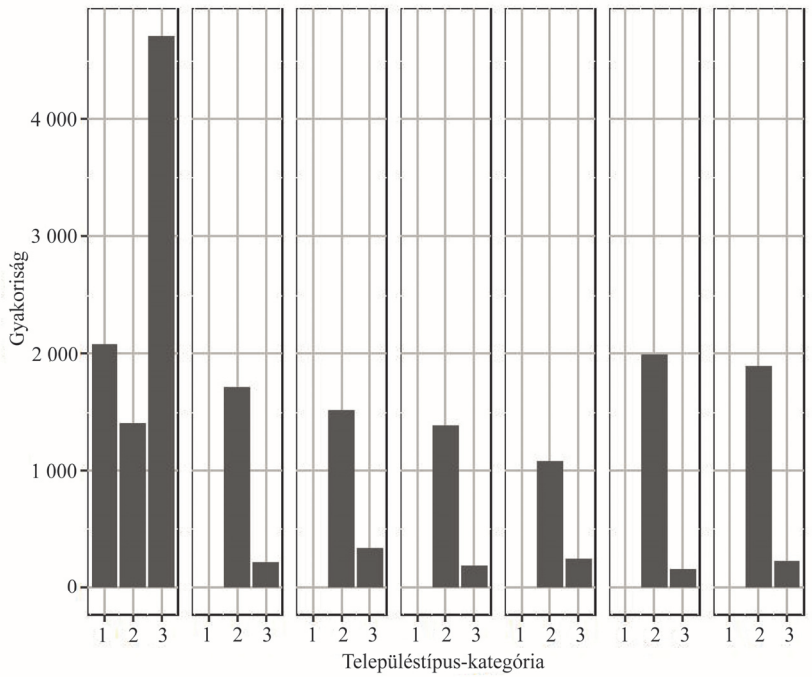
F10. ábra. Az egyéni szintű bérstruktúrahatás megoszlása a 2. szintű FEOR-kód szerinti
(Distribution of individual wage structure effects by 2-digit FEOR code)



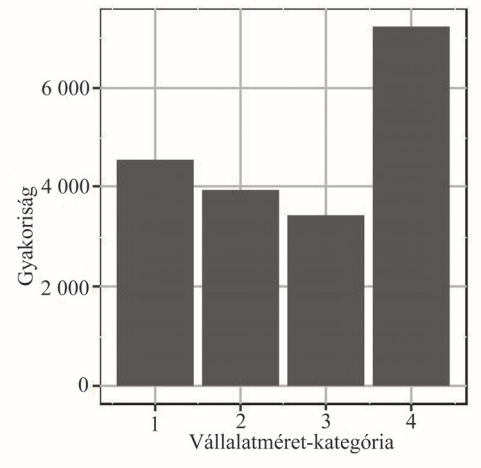
F11. ábra. Az egyéni szintű bérstruktúrahatás megoszlása állami és külföldi tulajdonú vállalatok szerint
(Distribution of individual wage structure effects by state- and foreign-owned companies)



F12. ábra. Az egyéni szintű bérstruktúrahatás megoszlása régió és településtípus szerint
(Distribution of individual wage structure effects by region and type of settlements)



F13. ábra. Az egyéni szintű bérstruktúrahatás megoszlása vállalatméret szerint
(Distribution of individual wage structure effects by company size)



3. Függelék

A kétféle módszerrel számított bérstruktúrahatások változók szerinti különbségét páros ANOVA-teszttel hasonlítottam össze. Ennek során normalitás-[9] és a varianciára vonatkozó homogenitásteszteket[10] végeztem. A varianciák általában nem egyeztek meg a csoportok között, emiatt az eltérő szórással rendelkező csoportok összehasonlítására is alkalmas Welch-próbát alkalmaztam. A következő táblázatokban a normalitásra, a varianciaegyezőségre és a Welch-próbára vonatkozó *p*-értékek szerepelnek.

[9] A normalitástesztekhez használt Shapiro-féle nullhipotézis: adott csoporton belül a megfigyelések eloszlása normális.

[10] A variancia homogenitásának tesztelésére alkalmazott Levene-teszt nullhipotézise: az összehasonlított csoportokhoz tartozó szórások megegyeznek.

F2. táblázat

A normalitásvizsgálathoz, a varianciaegyezőséghez és a Welch-próbához tartozó p-értékek iskolai végzettség szerint
(*p* values of normality testing, variance homogeneity and the Welch test, by level of education)

Iskolai végzettség kategóriája	Teszt	Iskolai végzettség kategóriája								
		1.	2.	3.	4.	5.	6.	7.	8.	9.
1	Welch-próba		0,44	0,79	0,23	0,46	0,17	0,47	0,61	0,30
	Varianciateszt		0,81	0,39	0,46	0,83	0,66	0,76	0,52	0,50
2	Welch-próba			0,00	0,00	0,66	0,00	0,79	0,01	0,03
	Varianciateszt			0,00	0,00	0,85	0,06	0,00	0,00	0,00
3	Welch-próba				0,00	0,00	0,00	0,00	0,00	0,00
	Varianciateszt				0,47	0,00	0,02	0,00	0,00	0,00
4	Welch-próba					0,00	0,05	0,01	0,00	0,13
	Varianciateszt					0,00	0,00	0,00	0,00	0,00
5	Welch-próba						0,00	0,95	0,02	0,01
	Varianciateszt						0,02	0,00	0,00	0,00
6	Welch-próba							0,00	0,00	0,00
	Varianciateszt							0,00	0,00	0,00
7	Welch-próba								0,27	0,12
	Varianciateszt								0,04	0,04
8	Welch-próba									0,00
	Varianciateszt									0,84
Normalitásteszt		0,49	0,19	0,06	0,00	0,00	0,00	0,80	0,17	0,77

Megjegyzés. Az iskolai végzettség kategóriáit lásd az 1. táblázatban.

F3. táblázat

A varianciaegyezőséghez és a Welch-próbához tartozó p-értékek életkor és szolgálati évek szerint
(*p* values of variance homogeneity and the Welch test, by age and year of service)

Teszt	*p*-érték
Életkor	
Welch-próba	0,00
Varianciaegyezőség	0,00
Szolgálati évek	
Welch-próba	0,00
Varianciaegyezőség	0,00

Megjegyzés. Az életkor és a szolgálati idő esetében azokat a csoportokat vizsgáltam, amelyekhez legalább 100 megfigyelés tartozik. Az életkor esetében ez a 21 és 60 év közötti, a szolgálati idő esetében a 0 és 24 év közötti csoportokat jelenti.

F4. táblázat

A normalitásvizsgálathoz, a varianciaegyezőséghez és a Welch-próbához tartozó p-értékek az 1. szintű FEOR-kód szerint
(*p* values of normality testing, variance homogeneity and the Welch test, by 1-digit FEOR code)

1. szintű FEOR-kód	Teszt	1. szintű FEOR-kód								
		1	2	3	4	5	6	7	8	9
1	Welch-próba		0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Varianciateszt		0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	Welch-próba			0,00	0,96	0,83	0,88	0,00	0,00	0,00
	Varianciateszt			0,00	0,00	0,00	0,00	0,00	0,00	0,00
3	Welch-próba				0,00	0,00	0,13	0,64	0,00	0,00
	Varianciateszt				0,00	0,00	0,02	0,00	0,52	0,00
4	Welch-próba					0,83	0,86	0,00	0,00	0,00
	Varianciateszt					0,00	0,11	0,08	0,00	0,00
5	Welch-próba						0,81	0,00	0,00	0,00
	Varianciateszt						0,69	0,00	0,00	0,02
6	Welch-próba							0,21	0,08	0,00
	Varianciateszt							0,35	0,02	0,85
7	Welch-próba								0,00	0,00
	Varianciateszt								0,00	0,02
8	Welch-próba									0,00
	Varianciateszt									0,00
Normalitásteszt		0,01	0,00	0,01	0,00	0,00	0,35	0,01	0,00	0,00

F5. táblázat

A normalitásvizsgálathoz, a varianciaegyezőséghez és a Welch-próbához tartozó p-értékek az 1. szintű TEÁOR-kód szerint
(*p* values of normality testing, variance homogeneity and the Welch test, by the first level of NACE Rev. 2)

Kód	Teszt	1. szintű TEÁOR-kód																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S
A	Welch-próba		0,26	0,00	0,00	0,00	0,02	0,00	0,40	0,00	0,95	0,00	0,00	0,30	0,05	0,53	0,00	0,00	0,00
	Varianciateszt		0,56	0,03	0,00	0,36	0,01	0,14	0,58	0,73	0,00	0,00	0,44	0,00	0,85	0,09	0,00	0,67	0,04
B	Welch-próba			0,05	0,00	0,01	0,72	0,03	0,35	0,01	0,24	0,00	0,01	0,39	0,52	0,83	0,00	0,00	0,12
	Varianciateszt			0,99	0,28	0,82	0,69	0,88	0,68	0,61	0,35	0,33	0,81	0,51	0,61	0,53	0,27	0,74	0,76
C	Welch-próba				0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,21	0,00	0,00	0,00
	Varianciateszt				0,00	0,40	0,10	0,07	0,01	0,01	0,00	0,00	0,40	0,00	0,00	0,26	0,00	0,35	0,37
D	Welch-próba					0,38	0,00	0,00	0,00	0,00	0,00	0,96	0,43	0,00	0,00	0,03	0,00	0,40	0,00
	Varianciateszt					0,00	0,06	0,00	0,00	0,00	0,65	0,57	0,00	0,25	0,00	0,95	0,90	0,00	0,12
E	Welch-próba						0,00	0,00	0,00	0,00	0,00	0,20	0,95	0,00	0,00	0,05	0,00	0,07	0,00
	Varianciateszt						0,08	0,82	0,58	0,45	0,00	0,00	0,94	0,01	0,37	0,20	0,00	0,78	0,23
F	Welch-próba							0,00	0,05	0,00	0,01	0,00	0,00	0,11	0,31	0,99	0,00	0,00	0,00
	Varianciateszt							0,03	0,00	0,00	0,06	0,04	0,10	0,34	0,00	0,52	0,10	0,11	0,90
G	Welch-próba								0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,16	0,00	0,00	0,00
	Varianciateszt								0,18	0,14	0,00	0,00	0,77	0,00	0,04	0,22	0,00	0,63	0,20
H	Welch-próba									0,00	0,29	0,00	0,00	0,74	0,13	0,65	0,00	0,00	0,00
	Varianciateszt									0,79	0,00	0,00	0,68	0,00	0,62	0,13	0,00	0,93	0,06

(A táblázat folytatása a következő oldalon)

(Folytatás)

Kód	Teszt	1. szintű TEÁOR-kód																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	P	Q	R	S
I	Welch-próba										0,00	0,00	0,00	0,00	0,00	0,03	0,08	0,00	0,02
	Varianciateszt										0,00	0,00	0,55	0,00	0,84	0,09	0,00	0,81	0,04
J	Welch-próba											0,00	0,00	0,22	0,02	0,52	0,00	0,00	0,00
	Varianciateszt											0,89	0,00	0,27	0,00	0,93	0,62	0,00	0,15
K	Welch-próba												0,28	0,00	0,00	0,02	0,00	0,24	0,00
	Varianciateszt												0,00	0,24	0,00	0,91	0,55	0,00	0,13
L	Welch-próba													0,00	0,00	0,04	0,00	0,09	0,00
	Varianciateszt													0,01	0,48	0,21	0,01	0,84	0,24
M	Welch-próba														0,35	0,69	0,00	0,00	0,00
	Varianciateszt														0,00	0,74	0,30	0,02	0,43
N	Welch-próba															0,81	0,00	0,00	0,00
	Varianciateszt															0,11	0,00	0,74	0,04
P	Welch-próba																0,01	0,01	0,15
	Varianciateszt																0,90	0,21	0,54
Q	Welch-próba																	0,00	0,00
	Varianciateszt																	0,01	0,15
R	Welch-próba																		0,00
	Varianciateszt																		0,23
Normalitásteszt		0,88	0,34	0,21	0,30	0,84	0,42	0,00	0,11	0,04	0,21	0,00	0,17	0,00	0,00	0,34	0,04	0,02	0,00

F6. táblázat

A normalitásvizsgálathoz, a varianciaegyezőséghez és a Welch-próbához tartozó p-értékek az állami és a külföldi tulajdonú vállalatok aránya szerint
(*p* values of normality testing, variance homogeneity and the Welch test, by state and foreign ownership ratio of companies)

Kategória-pár	Teszt	Állami és külföldi tulajdonú vállalatok szerinti kategóriapár									
		1_4	2_3	2_4	3_2	3_3	3_4	4_1	4_2	4_3	4_4
1_4	Welch-próba		0,00	0,05	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Variancia		0,31	0,41	0,04	0,00	0,34	0,00	0,00	0,18	0,62
2_3	Welch-próba			0,00	0,04	0,51	0,48	0,02	0,38	0,16	0,11
	Variancia			0,67	0,84	0,50	0,21	0,52	0,56	0,80	0,37
2_4	Welch-próba				0,00	0,00	0,00	0,00	0,00	0,00	0,00
	Variancia				0,70	0,12	0,21	0,08	0,12	0,83	0,52
3_2	Welch-próba					0,03	0,04	0,36	0,00	0,00	0,09
	Variancia					0,08	0,06	0,02	0,06	0,89	0,06
3_3	Welch-próba						0,94	0,00	0,79	0,00	0,12
	Variancia						0,01	0,65	0,72	0,13	0,00
3_4	Welch-próba							0,00	0,87	0,00	0,16
	Variancia							0,00	0,00	0,14	0,27
4_1	Welch-próba								0,00	0,00	0,00
	Variancia								0,98	0,07	0,00
4_2	Welch-próba									0,00	0,01
	Variancia									0,11	0,00
4_3	Welch-próba										0,00
	Variancia										0,24
Normalitásteszt		0,10	0,27	0,37	0,80	0,04	0,02	0,08	0,02	0,05	0,00

Megjegyzés. A kategóriapárok első tagja az állami, a második a külföldi tulajdont jelöli.

F7. táblázat

A normalitásvizsgálathoz, a varianciaegyezőséghez és a Welch-próbához tartozó p-értékek régiók és településtípusok szerint
(*p* values of normality testing, variance homogeneity and the Welch test, by region and type of settlements)

Kategóriapár	Teszt	Régiók és településtípusok szerinti kategóriapár														
		1_1	1_2	1_3	2_2	2_3	3_2	3_3	4_2	4_3	5_2	5_3	6_2	6_3	7_2	7_3
1_1	Welch-próba		0,00	0,00	0,66	0,13	0,43	0,10	0,00	0,30	0,01	0,00	0,68	0,00	0,00	0,14
	Varianciateszt		0,00	0,07	0,41	0,03	0,21	0,00	0,00	0,11	0,00	0,00	0,00	0,00	0,00	0,00
1_2	Welch-próba			0,26	0,01	0,86	0,00	0,00	0,00	0,01	0,55	0,00	0,00	0,02	0,00	0,63
	Varianciateszt			0,00	0,02	0,48	0,09	0,30	0,02	0,85	0,70	0,01	0,01	0,01	0,00	0,00
1_3	Welch-próba				0,00	0,48	0,00	0,00	0,00	0,00	0,08	0,00	0,00	0,05	0,00	0,26
	Varianciateszt				0,01	0,00	0,00	0,00	0,00	0,03	0,00	0,00	0,00	0,00	0,00	0,00
2_2	Welch-próba					0,20	0,24	0,06	0,00	0,22	0,04	0,00	0,94	0,00	0,00	0,23
	Varianciateszt					0,06	0,64	0,01	0,00	0,21	0,01	0,00	0,00	0,00	0,00	0,00
2_3	Welch-próba						0,06	0,02	0,00	0,06	0,88	0,01	0,18	0,05	0,03	0,84
	Varianciateszt						0,12	0,88	0,63	0,70	0,60	0,10	0,61	0,07	0,35	0,01
3_2	Welch-próba							0,27	0,00	0,51	0,00	0,00	0,23	0,00	0,00	0,06
	Varianciateszt							0,04	0,00	0,32	0,04	0,00	0,00	0,00	0,00	0,00
3_3	Welch-próba								0,01	0,88	0,00	0,00	0,06	0,00	0,00	0,02
	Varianciateszt								0,72	0,57	0,40	0,09	0,70	0,06	0,37	0,01
4_2	Welch-próba									0,03	0,00	0,00	0,00	0,00	0,00	0,00
	Varianciateszt									0,33	0,06	0,09	0,98	0,06	0,39	0,00
4_3	Welch-próba										0,03	0,00	0,22	0,00	0,00	0,06
	Varianciateszt										1,00	0,04	0,32	0,03	0,16	0,00
5_2	Welch-próba											0,00	0,02	0,01	0,00	0,93
	Varianciateszt											0,01	0,04	0,01	0,01	0,00
5_3	Welch-próba												0,00	0,78	0,29	0,00
	Varianciateszt												0,09	0,63	0,19	0,27
6_2	Welch-próba													0,00	0,00	0,20
	Varianciateszt													0,06	0,36	0,00
6_3	Welch-próba														0,60	0,02
	Varianciateszt														0,11	0,62
7_2	Welch-próba															0,01
	Varianciateszt															0,01
Normalitásteszt		0,47	0,00	0,00	0,13	0,10	0,00	0,01	0,11	0,60	0,44	0,66	0,00	0,03	0,00	0,00

Megjegyzés. A kategóriapárok első tagja a régióra, a második a településtípusra utal. Régiókódok: 1: Pest megye, 2: Közép-Dunántúl, 3: Nyugat-Dunántúl, 4: Dél-Dunántúl, 5: Észak-Magyarország, 6: Észak-Alföld, 7: Dél-Alföld; településkategória: 1: főváros, 2: város, 3: egyéb.

F8. táblázat

A normalitásvizsgálathoz, a varianciaegyezőséghez és a Welch-próbához tartozó p-értékek vállalatméret szerint
(*p* values of normality testing, variance homogeneity and the Welch test, by company size)

Vállalatméret-kategóira	Teszt	Vállalatméret-kategóira			
		20–49 fő	50–149 fő	150–499 fő	500 fő és afeletti
20–49 fő	Welch-próba		0,32	0,00	0,00
	Varianciaegyezőség		0,00	0,52	0,00
50–149 fő	Welch-próba			0,00	0,00
	Varianciaegyezőség			0,02	0,13
150–499 fő	Welch-próba				0,00
	Varianciaegyezőség				0,00
Normalitásteszt		0,00	0,00	0,13	0,02

Irodalom

BLAU, F. – KAHN, L. [2017]: The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*. Vol. 55. No. 3. pp. 789–865. http://dx.doi.org/10.1257/jel.20160995

BLINDER, A. S. [1973]: Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*. Vol. 8. No. 4. pp. 436–455. http://dx.doi.org/10.2307/144855

BREIMAN, L. [1998]: Arcing classifier (with discussion and a rejoinder by the author). *Annals of Statistics*. Vol. 26. No. 3. pp. 801–824. http://dx.doi.org/10.1214/aos/1024691079

BREIMAN, L. [2001]: Random forests. *Machine Learning*. Vol. 45. No. 1. pp. 5–32. http://dx.doi.org/10.1023/A:1010933404324

DINARDO, J. – FORTIN, N. M. – LEMIEUX, T. [1996]: Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*. Vol. 64 No. 5. pp. 1001–1044. http://dx.doi.org/10.2307/2171954

EARLE, J. S. – TELEGDY Á. [2012]: Privatizáció, foglalkoztatás és bérek. In: *Fazekas K. – Benczúr P. – Telegdy Á.* (szerk.): *Munkaerőpiaci Tükör 2012*. MTA Közgazdaság- és Regionális Tudományi Kutatóközpont Közgazdaság-tudományi Intézet, Országos Foglalkoztatási Közhasznú Nonprofit Kft. Budapest. 231–248. old. http://econ.core.hu/file/download/mt_2012_hun/egyben.pdf

ELEK P. – SCHARLE Á. – SZABÓ B. – SZABÓ P. A. [2009]: *A bérekhez kapcsolódó adóeltitkolás Magyarországon.* Közpénzügyi Füzetek. 23. sz. ELTE Empirikus Tanulmányok Intézete. Budapest. http://adko.hu/01_files/adotanulmanyok/2009/beradok-eltitk.pdf

EUROPEAN COMMISSION [2016]: *Strategic Engagement for Gender Equality 2016–2019.* PUBLICATIONS Office of the European Union. Luxembourg. http://dx.doi.org/10.2838/722771

FAZEKAS K. [2005]: A hazai és a külföldi tulajdonú vállalkozások területi koncentrációjának hatása a foglalkoztatás és munkanélküliség területi különbségeire. In: *Faluvégi A. –*

Fazekas K. (szerk.): *A hely és a fej: Munkapiac és regionalitás Magyarországon*. MTA Közgazdaságtudományi Intézet. Budapest. 47–74. old.

FORTIN, N. – LEMIEUX, T. – FIRPO, S. [2011]: Chapter 1 – Decomposition methods in economics. *Handbook of Labor Economics*. Vol. 4. Part A. pp. 1–102. https://doi.org/10.1016/S0169-7218(11)00407-2

GÁBOR R. I. [2008]: A hiányzó láncszem? Életpálya-keresetek és keresetingadozás. *Közgazdasági Szemle*. LV. évf. December. 1057–1074. old.

HECKMAN, J. J. [1979]: Sample selection bias as a specification error. *Econometrica*. Vol. 47. No. 1. pp. 153–161. http://dx.doi.org/10.2307/1912352

JAMES, G. – WITTEN, D. – HASTIE, T. – TIBSHIRANI, R. [2017]: *An Introduction to Statistical Learning with Applications in R*. Springer. New York.

JANN, B. [2008]: The Blinder–Oaxaca decomposition for linear regression models. *The Stata Journal*. Vol. 8. No. 4. pp. 453–479. http://dx.doi.org/10.1177/1536867X0800800401

LEYTHIENNE, D. – RONKOWSKI, P. [2018]: *A Decomposition of the Unadjusted Gender Pay Gap Using Structure of Earnings Survey Data.* Statistical Working Paper. Eurostat. http://dx.doi.org/10.2785/796328

MACHADO, J. A. F. – MATA, J. [2005]: Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*. Vol. 20. Issue 4. pp. 445–465. http://dx.doi.org/10.1002/jae.788

MINCER, J. A. [1974]: *Schooling, Experience, and Earnings*. National Bureau of Economic Research. Cambridge.

LOH, B. W.-Y. [2009]: Improving the precision of classification trees. *The Annals of Applied Statistics*. Vol. 3. No. 4. pp. 1710–1737. http://dx.doi.org/10.1214/09-AOAS260

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*. XL. évf. Július-augusztus. 814–836. old.

NEUMAN, S. – OAXACA, R. L. [2004]: Wage decompositions with selectivity-corrected wage equations: A methodological note. *Journal of Economic Inequality.* Vol. 2. No. 1. pp. 3–10. http://dx.doi.org/10.1023/B:JOEI.0000028395.38694.4b

OAXACA, R. [1973]: Male-female wage differentials in urban labor markets. *International Economic Review.* Vol. 14. No. 3. pp. 693–709.

OECD (ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT) [2018]: *OECD Employment Outlook 2018*. Paris. https://doi.org/10.1787/19991266

REILLY, B. [2001]: A nemek közötti bérkülönbségek elemzésének statisztikai módszerei. *Statisztikai Szemle*. 79. évf. 1. sz. 5–17. old.

REIMERS, C. W. [1983]: Labor market discrimination against hispanic and black men. *Review of Economics and Statistics*. Vol. 65. No. 4. pp. 570–579. http://dx.doi.org/10.2307/1935925

TAKÁCS O. – VINCZE J. [2019a]: *The Gender Pay Gap in Hungary: New Results with a New Methodology*. Discussion Paper. No. 24. Institute of Economics, Centre for Economic and Regional Studies. Budapest.

TAKÁCS O. – VINCZE J. [2019b]: *Blinder–Oaxaca Decomposition with Recursive Tree-based Methods: A Technical Note*. Discussion Paper. No. 23. Institute of Economics, Centre for Economic and Regional Studies. Budapest.

TELEGDY Á. [2013]: A közszféra és a vállalatok bérei közötti átterjedési hatás Magyarországon. *Közgazdasági Szemle*. LX. évf. Május. 555–578. old.

WEICHSELBAUMER, D. – WINTER-EBMER, R. [2005]: A meta-analysis of the international gender wage gap. *Journal of Economic Surveys*. Vol. 19. No. 3. pp. 479–511. http://dx.doi.org/10.1111/j.0950-0804.2005.00256.x

WEICHSELBAUMER, D. – WINTER-EBMER, R. [2007]: The effects of competition and equal treatment laws on gender wage differentials. *Economic Policy*. Vol. 22. No. 50. pp. 237–287. http://dx.doi.org/10.1111/j.1468-0327.2007.00177.x