Közzététel: 2025. augusztus 8.

A tanulmány címe:

Imputálási eljárások hatékonyságának tesztelése egyváltozós paneladatokon: szimulációs elemzés és gyakorlati ajánlások

Szerzők:
BILICZ HANGA LILLA
a Pécsi Tudományegyetem Közgazdaságtudományi Kar Közgazdaság és Ökonometria Intézetének adjunktusa
E-mail: bilicz.hanga@ktk.pte.hu
BILICZ DÁVID
a Pécsi Tudományegyetem Közgazdaságtudományi Kar Közgazdaság és Ökonometria Intézetének tanársegédje
E-mail: bilicz.david@ktk.pte.hu

DOI: https://doi.org/10.20311/stat2025.08.hu0719

Bilicz Hanga Lilla – Bilicz Dávid

Imputálási eljárások hatékonyságának tesztelése egyváltozós paneladatokon: szimulációs elemzés és gyakorlati ajánlások

Evaluating the effectiveness of imputation methods for univariate panel data: a simulation analysis and practical recommendations

Bilicz Hanga Lilla, a Pécsi Tudományegyetem Közgazdaságtudományi Kar Közgazdaság és Ökonometria Intézetének adjunktusa
E-mail: bilicz.hanga@ktk.pte.hu
Bilicz Dávid, a Pécsi Tudományegyetem Közgazdaságtudományi Kar Közgazdaság és Ökonometria Intézetének tanársegédje
E-mail: bilicz.david@ktk.pte.hu

A hiányzó adatok kezelése kiemelt jelentőségű a statisztikai modellezés és az empirikus kutatások pontosságának biztosítása érdekében. Különösen a területi paneladatok esetében kulcsfontosságú a megfelelő imputációs módszerek kiválasztása, mivel a hagyományos törlési eljárások jelentős adatvesztéssel és torzítással járhatnak. Tanulmányunkban egyváltozós paneladatok hiányzó értékeinek imputációját vizsgáljuk öt, különböző karakterisztikákkal rendelkező ismérv esetében, szimulációs kísérlet segítségével tesztelve tizenegy különböző, széles körben alkalmazott eljárást. Eredményeink alapján az adatok panel jellegét figyelembe vevő, peremeloszlásra építő módszerek általánosan jobb teljesítményt nyújtanak, függetlenül az adatok jellegétől. Következésképpen alkalmazásukat különösen ajánljuk a hiányzó területi adatok imputációjára, hiszen nemcsak a panelelemzések, hanem sok esetben a keresztmetszeti adatokra építő kutatások esetén is elérhető panelmegfigyelés a területi ismérvek tekintetében, amely karakterisztika figyelembevétele jelentősen javíthatja az imputálás és következésképpen a statisztikai eljárások megbízhatóságát. Kutatásunk rávilágít továbbá arra is, hogy az aggregált értékek ismerete nem feltétlenül javítja az imputálás pontosságát, valamint bármilyen csoportosításon alapuló technika hatékonysága erősen függ a csoportképzés relevanciájától.

Kulcsszavak: hiányzó adatok, imputációs módszerek, panelelemzés

Handling missing data is crucial for ensuring the accuracy of statistical modeling and empirical research. The selection of appropriate imputation methods is particularly important for spatial panel data, as traditional deletion techniques may lead to significant data loss and bias. In our study, we examine the imputation of missing values in univariate panel data across five variables with different characteristics, testing eleven widely used methods through a simulation experiment. Our results indicate that methods based on marginal distributions, which account for the panel nature of the data, consistently outperform other approaches, regardless of data characteristics. Consequently, we strongly recommend their application for imputing missing spatial data, as panel observations are often available even in cross-sectional studies involving spatial indicators. Considering these characteristics can significantly improve imputation accuracy and, consequently, the reliability of statistical analyses. Furthermore, our

research highlights that the knowledge of aggregated values does not necessarily enhance imputation accuracy, and the effectiveness of any grouping-based technique heavily depends on the relevance of the grouping criteria.

Keywords: missing data, imputation methods, panel data analysis

A hiányzó adatok problémája az empirikus kutatások egyik legnagyobb kihívása: a megfelelő kezelési módszer kiválasztása alapvetően befolyásolja az elemzések pontosságát és megbízhatóságát *(Little–Rubin, 2002; Allison, 2009)*, így az imputációs technikák fejlesztése és összehasonlítása kiemelt fontosságú a statisztikai modellezésben és az adatelemzésben egyaránt. Az adathiány okai igen sokrétűek lehetnek: szerepelhet közöttük többek között a technikai hiba, a válaszmegtagadás vagy az adminisztratív korlátozások, területi adatgyűjtés esetében pedig a leggyakoribb az adatszolgáltatók közötti eltérésekből, illetve az adatvédelmi és hozzáférési korlátozásokból adódó hiányzó adat *(Baltagi, 2021)*. A területi adatok hiányzó értékeinek kezelése különösen fontos *(Anselin, 1988)*, hiszen ezek az adatok gyakran döntéshozatali folyamatok alapját képezik. Emellett számos térbeli elemzés sem folytatható hiányzó adatstruktúrákon. Ennek ellenére kevés az olyan átfogó módszertani tanulmány, amely paneladatokra fókuszál. Fontos azt is megemlíteni, hogy a különböző területi egységekre rendelkezésre álló adatok skálája igen széles lehet, azonban a több változó együttes figyelembevétele mentén történő adathiánykezelés olyan implicit együttmozgásokat is feltételez, amelyek nem biztos, hogy adott területi szinten is megfigyelhetők. Mivel paneladatok esetében sok megfigyelés együttes figyelembevétele már egyetlen változó vizsgálatakor is biztosítható, az egy változóból kiinduló adatbecslés csökkentheti a szubjektum szerepét az adathiány kezelésének módszerválasztásában. Az adatimputáció szempontjából fontos továbbá a területi adatok típusainak pontos ismerete is, mivel az adatok térbeli szerkezete és jellege befolyásolja a hiányzó értékek kezelésének módszertanát. Léteznek expliciten meghatározott szomszédsági viszonyokkal rendelkező adatok, ahol a területi egységek közötti kapcsolatok meghatározzák a térbeli autokorrelációt, és amelyek becslésére gyakran súlymátrixokat is alkalmaznak. Az ilyen adatok esetében a térbeli interpoláció kulcsfontosságú módszertani megközelítés a hiányzó értékek pótlásában *(Jakobi, 2009)*, hiszen ekkor a megfigyelésekhez pontos földrajzi koordináták is tartoznak. Jelen tanulmány ezen esetekre tartalmi korlátok következtében nem tér ki részletesen, és területi adatok alatt területi szinten aggregált változókat ért a továbbiakban.

A hiányzó adatok kezelésére számos módszer létezik, a legegyszerűbb *listwise* (teljes eset) és *pairwise* (páronkénti) törléstől kezdve a fejlettebb, gépi tanuláson

alapuló imputációs algoritmusokig. Hiányzó adatok kezelésekor fontos kiindulópont a hiány elfogadásának vagy mesterséges pótlásának szakmai és etikai mérlegelése, valamint az imputációs eljáráshoz releváns változók körének kijelölése. Egyváltozós mintáknál a hiányzó értékek pótlására nem állnak rendelkezésre más változók, így az imputáció elsősorban a térbeli autokorreláció és a statisztikai becslési technikák kihasználásán alapul. Az egyszerűbb törlési és imputációs módszerek előnye, hogy szinte minden esetben alkalmazhatók, hátrányuk viszont, hogy gyakran jelentősen torzítják az eredményeket vagy túlzottan szűkítik a mintát *(van Buuren, 2012)*. Ezzel szemben a komplexebb statisztikai módszerekre építő eljárások pontosabb becsléseket kínálnak, azonban jellemzően nagyobb adatigényűek, és bonyolultabb szoftveres háttér szükséges a kivitelezésükhöz. Az utóbbi évtizedekben egyre nagyobb hangsúly kerül továbbá a Bayes-alapú módszerekre *(Schafer, 1997)* és a mélytanulási megközelítésekre *(Gondara–Wang, 2018)*, amelyek az adatok belső mintázatait kihasználva képesek csökkenteni az imputációból fakadó bizonytalanságot. Ennek ellenére az optimális módszer kiválasztása továbbra is függ az adatok jellegétől és a kutatás céljától, ezért a különböző technikák összehasonlítása továbbra is nyitott kutatási kérdésként szolgál. Tanulmányunk célja, hogy szimulációs eljárás segítségével értékeljük néhány könnyen és széles körben alkalmazható imputációs módszer pontosságát és megbízhatóságát különböző területi adatszerkezetek esetén, egyetlen változóból, azonban panel adatstruktúrából kiindulva. Célunk továbbá az is, hogy javaslatokat tegyünk optimális imputációs technika kiválasztására a gyakorlati alkalmazásban. Tanulmányunk elsősorban társadalmi adatok elemzésével foglalkozik, így nem tér ki egyéb diszciplínákban (például klinikai, biológiai, pszichológiai, műszaki) keletkező, területi egységekre szintén értelmezhető, speciális adatokra.

1. Az adathiány karakterisztikái

Mielőtt részletesen bemutatjuk az adathiány kezelésére szolgáló, tanulmányunk szempontjából releváns módszereket, ki kell térnünk a megfelelő technikák kiválasztását megelőző elemzési lépésekre is, azaz az adathiány struktúrájának azonosítására, valamint az adathiány mértékének meghatározására.

Hair és szerzőtársai (2010) az adathiány kezelésének négy fő lépését javasolják. Első lépésben szükséges eldönteni, hogy ignorálható-e az adathiány. Ez területi adatok esetén jellemzően nem áll fenn *(Anselin, 1988)*. Második lépésben az

adathiány mértékének felmérésére térnek ki a szerzők, ennek szakirodalmi támpontjairól az 1.2. alfejezetben beszélünk részletesen. Az adathiány mértékét jelen tanulmányunkban 5%-ban határoztuk meg, de robusztusságellenőrzés végett 10%-os adathiány mellett is elvégeztük tesztjeinket. Harmadik lépésben *Hair és szerzőtársai (2010)* nyomán az adathiány-struktúra azonosítása következik, amelyet részletesen az 1.1. alfejezetünkben tárgyalunk, és ebből az MCAR (*missing completely at random*, teljesen véletlenszerűen hiányzó) -adatokkal foglalkozunk jelen tanulmányunkban. Végül, utolsó lépésként a hiányzó adatok helyettesítésének kérdése következik, amely lehetőségek közül tanulmányunkban mind az ismert, mind a becsült értékek helyettesítésével történő adatkezelési technikák egyes módszereit is teszteljük, ezeket a 2.2. alfejezet mutatja be részletesen.

1.1. Adathiány-struktúrák

Elsőként *Rubin (1976)* vetette fel, hogy nem mindegy, a hiányzó adatok mögött milyen folyamat húzódik meg. Ezek alapján a szakirodalomban három különböző megközelítés terjedt el, amelyek a hiányzó adatok keletkezését azok véletlenszerűsége alapján csoportosítják. E három kategória a teljesen véletlenszerűen hiányzó adatok (*missing at completely random*, MCAR), a véletlenszerűen hiányzó adatok (*missing at random*, MAR) és a nem véletlenszerűen hiányzó adatok (angolul általában *missing not at random*, MNAR).

A teljesen véletlenszerűen hiányzó (MCAR-) adatok esetében arról beszélhetünk, hogy a változó hiányzásának valószínűségét semmilyen megfigyelt vagy meg nem figyelt tényező nem befolyásolja. Következésképpen a hiányzó adatok figyelmen kívül hagyása MCAR-struktúrájú adathiány esetén nem vezet torzításhoz, hanem a csökkent mintaméret miatt a becslések standard hibáját növeli *(Dong–Peng, 2013)*.

A továbbiakban a formalizálás céljából legyen egyetlen ismérvünk mátrix formába (**M**) rendezve, amelynek minden sora egy megfigyelési egységet, minden oszlopa pedig egy időegységet képvisel. Az **M** mátrix $a \times b$ méretű, vagyis a darab sora és b darab oszlopa van. Továbbá legyen az **M** mátrix általános eleme m_{ij}, amely az i-edik megfigyelési egység j-edik időperiódusban felvett értékét jelöli. E mátrix valamekkora hányada esetében MCAR típusú adathiánnyal szembesülünk. Az adathiány szemléltetésére vezessük be az **F**, úgynevezett adathiányindikátor-mátrixot (*flag matrix*) *(Békés–Kézdi, 2021; Little–Rubin, 2002; Oravecz, 2008)*, amely szintén $a \times b$ méretű, általános eleme pedig f_{ij}, amely

$$f_{ij} = \begin{cases} 1 \mid m_{ij} \text{ hiányzik} \\ 0 \mid m_{ij} \text{ nem hiányzik.} \end{cases} \quad (1)$$

Az MCAR struktúrájú adathiány esetében ekkor azt mondhatjuk, hogy az adatállományban foglalt minden egyes adatpont esetében a hiány valószínűsége azonos, azaz:

$$P(\mathbf{F}|\mathbf{M}) = P(\mathbf{F}). \tag{2}$$

A véletlenszerűen hiányzó adatok (MAR) esetében ennél már egy jóval lazább feltételezéssel tudunk élni. Ilyen esetben ugyanis az adathiányok már nem teljesen véletlenszerűek, azonban minden, az adathiány mértékét befolyásoló tényező ismert. A MAR-struktúrájú adathiány definíciójának formalizálásához felírhatjuk ismét az **M-**mel azonos dimenziójú adathiányindikátor-mátrixot (**F**). Ha a hiányzási mechanizmus MAR, akkor a hiányzási minta csak a megfigyelt adatokon keresztül függ az összes adattól *(Schafer, 1997)*. Ekkor tehát, ha $P(\mathbf{F}|\mathbf{M}, \xi)$ alakban írjuk fel M eloszlását, elmondható, hogy:

$$P(\mathbf{F}|\mathbf{M}, \xi) = P\left(\mathbf{F}\middle|\mathbf{M}_{mf}, \mathbf{M}_h, \xi\right) = P\left(\mathbf{F}\middle|\mathbf{M}_{mf}, \xi\right), \tag{3}$$

ahol $\mathbf{M}_{mf}$ az $\mathbf{M}$ mátrix megfigyelt része, $\mathbf{M}_h$ pedig az $\mathbf{M}$ mátrix hiányzó része, valamint ξ a hiányzási paraméter. Vagyis a hiányzás valószínűsége nem függ közvetlenül a hiányzó értékektől ($\mathbf{M}_h$), hanem csak a megfigyelt adatoktól ($\mathbf{M}_{mf}$) és egyéb paraméterektől (ξ). Területi adatok esetében illusztratív példa lehet, ha egy ország régióinak gazdasági fejlődését vizsgáljuk, és rendelkezésünkre állnak a vállalkozói aktivitás (például újonnan alapított vállalkozások száma) és a gazdasági teljesítmény (például GDP/fő) adatai. A vállalkozói aktivitás adatai azonban hiányosak lehetnek bizonyos területeken, különösen ott, ahol a gazdasági teljesítmény kezdetben is alacsony volt. További illusztratív példát szolgáltat az 1. táblázat, amelyben láthatjuk, hogy a nők kisebb valószínűséggel adják meg az életkorukat. Ez azonban az életkor változóban nem azonosítható mintázat, csupán a nem változó figyelembevételével látunk mintázatot a hiányzás struktúrájában. Tehát ez a példa is megfelel a MAR definíciójának.

A nem véletlenszerűen hiányzó (MNAR-) adatok körébe tartozik minden olyan adathiány, amelynél az előző két feltételezés (MCAR és MAR) egyaránt sérül. Ez a gyakorlatban azt jelenti, hogy egy adatpont hiánya összefüggésben van vagy az adott érték nagyságával, vagy egyéb, meg nem figyelhető változókkal. Az előbbiekhez hasonlóan MNAR esetében is fel tudjuk írni a hiányzási valószínűséget a már bevezetett adathiányindikátor-mátrix (**F**) segítségével *(Alwateer et al., 2024)* a következőképpen:

$$P(\mathbf{F}|\mathbf{M}) \neq P\left(\mathbf{F}\middle|\mathbf{M}_{mf}\right). \tag{4}$$

Jó példa az MNAR-adathiányra a gyakorlatban is megfigyelhető azon tendencia, miszerint a magasabb jövedelemmel rendelkezők kevésbé hajlandóak jövedelmi adatokat megadni, mint az átlagos körül vagy az alatt keresők. Az 1. táblázatban látható, hogy azon egyének, akik 40 év feletti életkorcsoportba sorolhatók,

nem adták meg életkorukat az illusztratív példában. Mivel így magában a megfigyelt változóban is felfedezhető saját adatstruktúrájához köthető mintázat, ebben az esetben egyértelműen nem véletlenszerű adathiányról (MNAR) beszélhetünk.

1. táblázat

Hiányzási mechanizmusok (MNAR, MAR, MCAR) illusztrálása példán keresztül
Illustration of missing data mechanisms (MCAR, MAR, MNAR) with an example

X = nem (1 = nő)	Y = kor (év)	MNAR	MAR	MCAR
1	27	27	Hiányzik	27
1	31	31	Hiányzik	Hiányzik
1	33	33	Hiányzik	33
1	33	33	Hiányzik	Hiányzik
1	41	Hiányzik	Hiányzik	41
2	26	26	26	26
2	27	27	27	27
2	29	29	29	Hiányzik
2	29	29	29	29
2	32	32	32	32
2	33	33	33	Hiányzik
2	40	Hiányzik	40	40
2	43	Hiányzik	43	43
2	44	Hiányzik	44	Hiányzik
2	47	Hiányzik	47	47
Mintázat Y változó esetén:		Csak Y > 40 esetén hiányzik	Random	Random
X változó kapcsolata Y hiányzó értékeivel		Nincs kapcsolat	Csak X = 1 esetén hiányzik	Nincs kapcsolat

Forrás: *Enders (2010)* alapján saját példa, saját szerkesztés.

1.2. Az adathiány mértéke

Az adathiány struktúrájának feltérképezése mellett fontos néhány szóban kitérni az adathiány mértékére is. Bár az elfogadhatóság határértéke függ az adatok szerkezetétől, a hiányzási mechanizmustól és az alkalmazott statisztikai módszerektől, a szakirodalom számos iránymutatást kínál.

Általánosan elterjedt nézet, hogy amennyiben az adathiány mértéke kevesebb mint 1%, az adathiány elhanyagolható. Ha a hiány kevesebb, mint 5%, és a hiány mechanizmusa MCAR, az eredmények torzítása továbbra is minimális *(Schafer, 1999)*, és a standard statisztikai módszerek megbízhatónak tekinthetők *(Little–*

Rubin, 2002). *Máder (2005)* alapján az 5 és 15% közötti adathiány már mindenképpen komolyabb statisztikai módszerekkel kezelendő. A tudományos konszenzus szerint a 10%-nál kisebb adathiány általában nem okoz jelentős torzítást *(Bennett, 2001; Hair et al., 2010)*. Ha az adathiány 10–20% közötti, akkor *Schafer és Graham (2002)* alapján az imputáció vagy más korrekciós módszerek alkalmazása válik szükségessé, de *Máder (2005)* arra is felhívja a figyelmet, hogy a 15% feletti adathiány már súlyos interpretálhatósági problémákat is felvethet. 20%-os arány felett már komplexebb statisztikai módszerek (például többszörös imputáció, Bayes-alapú imputáció, maximum likelihood becslés) alkalmazása szükséges. Egyes kutatások szerint azonban akár 30%-os hiányzási arány is kezelhető megfelelő imputációs technikák alkalmazásával, különösen, ha a hiány MAR vagy MCAR típusú *(Dong–Peng, 2013)*.

2. Az adathiány kezelése

Az adathiány kezelésekor *Oravecz (2008)*, valamint *Little és Rubin (2002)* alapján négy, részben átfedő csoportba sorolhatjuk a rendelkezésre álló módszereket: teljesen megfigyelt vagy elérhető egységek elemzésén alapuló eljárások, átsúlyozás, imputációalapú eljárások és modellalapú eljárások. A teljesen megfigyelt vagy elérhető egységek elemzésén alapuló eljárások közé sorolhatjuk például a törlési eljárásokat, amelyekről a 2.1. alfejezet beszél részletesen, az imputációs eljárásokat pedig a 2.2. alfejezet ismerteti. Átsúlyozás esetében nem becsüljük vissza a hiányzó mintabeli értékeket, hanem a súlyok alakításával pontosítjuk a sokasági paraméter-becsléseinket, így jelen tanulmány esetében az e módszertan alá tartozó eljárások nem relevánsak. A modellalapú eljárások pedig általában többváltozós modellekben működnek jól, ahol a megfigyelt változók segíthetnek a hiányzók becslésében. Ennek oka, hogy e módszerek egy olyan modellt definiálnak, ahol a becsléseket a posterior valószínűségekre vagy a likelihoodra alapozzák *(Oravecz, 2008)*. Paneladatok esetében tehát akkor működhetnek a modellalapú becslések, ha például idősorelemzési technikákat alkalmazunk. Ez utóbbira lehet példa egy dinamikus lineáris modell (DLM) *(Laine, 2019)* is, ez az irány azonban túlmutat jelen tanulmány tartalmán. Tanulmányunk egyik lényeges korlátja, hogy a terjedelmi keretek miatt nem nyílik lehetőség valamennyi releváns, a gyakorlatban is alkalmazott imputációs eljárás részletes bemutatására. Ezért a továbbiakban elsősorban azon módszerek bemutatására összpontosítunk, amelyek a tanulmány empirikus része kapcsán relevánsak, és a következő alfejezetekben ezek mélyrehatóbb ismertetésére törekszünk.

2.1. Törlési módszerek

A hiányzó adatok kezelésének egyik legegyszerűbb, de gyakran problémás megközelítése a *listwise* (teljes esetek törlése) és a *pairwise* (páronkénti) törlés. A *listwise* törlés esetében minden olyan megfigyelési egység kikerül az elemzésből, amelynél legalább egy adat hiányzik, míg a *pairwise* törlés az adott elemzéshez szükséges változók alapján szelektálja az elérhető adatokat. Bár ezek a módszerek egyszerűen alkalmazhatók, és megőrzik az adatok eredeti skáláját, komoly hátrányuk, hogy jelentős információvesztést okozhatnak *(Sajtos–Mitev, 2007)*, különösen, ha a hiányzó adatok aránya magas.

Területi szinten aggregált vagy területi egységekre rendelkezésre álló adatok esetében a *listwise* törlés különösen problematikus lehet, mivel bizonyos régiók vagy települések teljes mértékben kieshetnek az elemzésből, ami torzított következtetésekhez vezethet, ha a hiány nem véletlenszerű (MAR vagy MNAR esetén). A *pairwise* törlés ugyan mérsékeli az adatvesztést, de az eltérő mintanagyság miatt az eredmények összehasonlíthatósága csökkenhet. Ezek a módszerek ezért jellemzően csak akkor indokoltak, ha a hiányzó értékek MCAR (teljesen véletlenszerű) módon fordulnak elő, vagy ha az adathiány mértéke minimális. Ha csak egy változó áll rendelkezésre (például egy régió GDP-je, munkanélküliségi rátája stb.), akkor a *listwise* (teljes esetek törlése) és a *pairwise* (páronkénti) törlés alkalmazhatósága jelentősen korlátozott. Ekkor ugyanis az egyetlen változó hiányzó értékei miatt az adott megfigyelési egységeket teljesen elveszítjük – azaz gyakorlatilag ugyanoda jutunk mind a teljes eset törlése, mind a páronkénti törlés esetében, így az utóbbi eljárás egyetlen változó esetén, mondhatjuk, hogy egyáltalán nem releváns.

A törlési eljárások (*listwise* és *pairwise* törlés) nem tekinthetők imputációs módszereknek. Az imputációs módszerek célja a hiányzó adatok pótlása, míg a törlési eljárások a minták csökkentésére szolgálnak a hiányzó értékek eltávolításával. Mivel egyváltozós területi adatok esetében a törlési eljárások jelentős adatvesztéssel és torzítással járhatnak, helyettük imputációs módszereket érdemes alkalmazni *(Allison, 2009)*.

2.2. Imputációs módszerek

Imputációs módszerek terén különbséget tehetünk egyszerűbb és komplexebb – modell alapú – technikák között. Mivel e tanulmány egyváltozós, idősoros panel-adatokra fókuszál, így a komplexebb, sok esetben egyéb változókhoz vagy gépi tanuláshoz kötött módszerek alkalmazása ilyen típusú adatfeldolgozáskor általában nem kivitelezhető. Tanulmányunk célja továbbá olyan imputációs módszerek

azonosítása, amelyek a későbbiekben széles körben alkalmazhatók, és amelyek implementálása nem igényel összetett számítási erőforrásokat. A 2.2.1.–2.2.4. alfejezetben röviden ismertetjük azon egyszerűbb, a gyakorlatban széles körben leggyakrabban alkalmazott módszereket, amelyek a tanulmányunk szempontjából relevánsnak tekinthetők.

2.2.1. Középértékkel való helyettesítés

Az adathiány problémájának egyik legegyszerűbb megoldása a hiányzó értékek valamilyen középértékkel való helyettesítése. A helyettesítés ilyenkor jellemzően a megfigyelt értékek átlagával, esetleg móduszával vagy mediánjával történik *(Máder, 2005)*. Az eljárás előnye annak egyszerűsége értelmezési és számításigényesség szempontjából egyaránt. Hátránya, hogy minden megfigyelést azonos értékkel helyettesít. Ennek következtében az adott ismérv eloszlása erősen torzulhat *(Alwateer et al., 2024)*, mivel az értékek szórása jelentősen csökkenhet *(Kang, 2013; Malhotra, 1987)*.

Mivel az adataink paneljelleggel kerültek megadásra, a középértékkel való helyettesítés esetében három könnyen interpretálható módon is meghatározható, hogy pontosan milyen szint középértékével helyettesítünk: kézenfekvő lehatárolni a mátrix teljes egésze, annak adott sora vagy adott oszlopa esetében mért középértéket. Ennek következtében az átlaggal való helyettesítés formalizálva az alábbi három módon oldható meg:

$$\bar{m} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} m_{ij}}{\sum_{i=1}^{a}\sum_{j=1}^{b}(1-f_{ij})} \tag{5}$$

$$\bar{m}_{i.} = \frac{\sum_{j=1}^{b} m_{ij}}{\sum_{j=1}^{b}(1-f_{ij})} \tag{6}$$

$$\bar{m}_{.j} = \frac{\sum_{i=1}^{a} m_{ij}}{\sum_{i=1}^{a}(1-f_{ij})}, \tag{7}$$

ahol a korábbiakban bevezetett jelölésen túl $\bar{m}$ a teljes megfigyelésre vonatkoztatott átlag imputációs helyettesítő érték, $\bar{m}_{i.}$ az i-edik sor esetében a sor alapú átlaggal való helyettesítő érték, $\bar{m}_{.j}$ a j-edik oszlop esetében az oszlop alapú átlaggal való helyettesítő érték. A fentiekhez hasonlóan más középértékek esetében is elvégezhető mindhárom szemléletmód szerint a helyettesítő értékek kiszámítása, amely aztán három különböző imputációs eljárás értékeiként szolgál. A sor- és az oszlopalapú helyettesítés esetében a módszer alkalmazhatóságának feltétele, hogy minden előbbi esetben minden megfigyelési egységnek legalább egy időperiódusban, utóbbi esetben pedig minden időegység esetében legalább egy megfigyelési egységre rendelkeznünk kell megfigyelt adattal. Kutatásunk során a középértékkel való helyettesítés két fajtája, az átlag- és a mediánimputáció esetében mindhárom

típusú eljárást alkalmaztuk, ami tanulmányunk szempontjából összesen hat, külön értékelt imputálási eljárást jelent.

2.2.2. Utolsó megfigyelés továbbvitele

Egy másik fajta, idősorok és panelek esetében jól alkalmazható imputációs eljárás, egy kijelölt megfigyelés továbbvitele, amely a soralapú középértékkel való helyettesítésre hasonlít abban, hogy az adott megfigyelési egységeket önállóan vizsgálja. Előnye, hogy nem egy teoretikus értéket, hanem az adott megfigyelési egység esetében egy valós, megfigyelt értéket helyettesítünk be a hiányzó adat helyére. Az eljárás számításigényesség szempontjából viszonylag egyszerű, és az átlaggal való helyettesítéshez képest a helyettesítő érték dinamikusabb módon kezelhető. Hátránya lehet, hogy erős torzítást okozhat, ha az adatok nagyon fluktuálnak, vagy erős növekvő vagy csökkenő trendet mutatnak. A módszer továbbá fokozottan érzékeny arra, ha egy megfigyelési egység esetében egymást követően több időperiódusra is hiányzó adatokat tapasztalunk *(Alwateer et al., 2024)*. A kijelölt érték statikusan és dinamikusan is megválasztható. Statikus kijelölt érték esetében egy év, jellemzően egy idősor kezdeti értéke kerül kijelölésre, és minden hiányzó adatnál az adott megfigyelési egység első megfigyelt értékét helyettesítik a hiányzó adatpont helyére. Kutatásunkban a dinamikus kijelölt érték továbbvitelének egy esetét, az utolsó megfigyelt érték továbbvitelét (*last observation carried forward*, LOCF) alkalmaztuk, általános esetben:

$$m_{ij} = m_{ij-1} | f_{ij} = 1 \text{ és } f_{ij-1} = 0 \quad (8)$$

Az eljárás halmozott hiányzó értékek esetében, vagyis amikor egy adott m_{ij-1} és m_{ij} pár egyaránt hiányzik, további specifikációra szorul. Elemzéseink esetében ilyenkor a hiányzó adatok feltöltését $j = 1, 2, 3 \ldots b$ sorrendben végeztük el, vagyis a fenti példánál maradva az m_{ij-1} és az m_{ij} értékét egyaránt m_{ij-2} értékével töltöttük fel. Amennyiben i megfigyelési egység esetében $f_{i1} = 1$, vagyis az első megfigyelt időszak esetében adathiányt tapasztaltunk, úgy a (9) egyenlet alapján jártunk el:

$$m_{i1} = m_{i2} | f_{i1} = 1 \text{ és } f_{i2} = 0, \quad (9)$$

azaz a hiányzó értéket i megfigyelési egység első megfigyelt értékéből becsültük. A módszernek a soralapú középértékkel való helyettesítéshez hasonlóan szintén követelménye, hogy minden megfigyelési egység legalább egy időperiódusban rendelkezzen megfigyelt értékkel.

2.2.3. Idősoros/regressziós becsléssel való helyettesítés

Ahogy az utolsó megfigyelés továbbvitele esetében láthattuk, bizonyos eljárások képesek figyelembe venni a paneladatok idősoros jellegét. Hasonló elven működik

az idősoros regressziós eljárással való becslés is. Ebben az esetben minden megfigyelési egységre felírható egy

$$y_{ij} = \beta_0 + \beta_1 j + \varepsilon \tag{10}$$

regressziós egyenlet, így a legkisebb négyzetek módszerével megbecsülhető egy $\hat{y}_{ij}$ érték minden egyes megfigyelési egységre minden időperiódusban. Ezt követően az **M** mátrix hiányzó elemei helyére ezen értékeket helyettesítve a hiányzó értékektől mentes kiegyensúlyozott panel adattáblánk felírható a következőképpen:

$$m_{ij} = \begin{cases} m_{ij} \mid f_{ij} = 0 \\ \hat{y}_{ij} \mid f_{ij} = 1 \end{cases} \tag{11}$$

A regressziós eljárás előnye és hátránya is, hogy képes trendeket megragadni az adatokban. A trendek figyelembevétele javíthat az eljárás pontosságán, továbbá az utolsó megfigyelt érték alapú eljáráshoz viszonyítva kevésbé érzékeny az egymásután hiányzó adatokra. Amennyiben azonban az adatokat leíró trend nem lineáris, esetleg a trendben strukturális törés van, vagy nem is jellemzi trendszerűség, akkor a hiányzó adatok becslése pontatlan lehet *(Alwateer et al., 2024)*. Az előző eljáráshoz hasonlóan szintén feltétele a módszer alkalmazhatóságának, hogy megfigyelési egységenként rendelkezésre álljon megfigyelt érték, azonban ebben az esetben legalább két megfigyelt érték szükséges.

2.2.4. Peremeloszlások alapján történő becslés

Az eddig bemutatott imputálási eljárásokban legfeljebb korlátozott mértékben tudtuk figyelembe venni az adatok panel jellegét. Ugyan a sor- és az oszlopalapú helyettesítés, illetve a regressziós és az utolsó megfigyelt érték alapú módszerek egyaránt kihasználják, hogy az adataink sorai egyazon megfigyelési egység időben egymást követő értékei, vagy azt, hogy az oszlopokban lévő értékek azonos időperiódus különböző alanyok esetében tapasztalt adatai, a két tulajdonság együttes figyelembevételére azonban egyetlen eddig bemutatott módszer sem volt alkalmas. Ennek a problémának az áthidalására alkalmazható egy kifejezetten paneladatok hiányának imputációjára szolgáló eljárás *(Kleinke et al., 2011)*, a peremeloszlások alapján történő becslés.

Az eljárás öt lépésből áll. Első lépésben szükséges azonosítani az időszaki átlagos értékeket a területi egységek összessége esetében. Ez praktikusan megegyezik az oszlopalapú átlaggal való helyettesítés helyettesítő értékével: a (7) egyenlet alapján tehát kiszámítható az $\bar{m}_{.j}$ oszlopátlagok értéke minden egyes j időegység esetére.

Második lépésben szükséges meghatározni a vizsgált mutatók esetében az adott régió és az aggregált oszlopátlagok értékének hányadosát. Ez a mutató abszolút vagy százalékos értékben adja meg, hogy az adott régió hogyan teljesít a teljes

megfigyelt terület (tanulmányunk esetében Magyarország) átlagához viszonyítva, formalizálva pedig esetünkben a százalékos eltérés a következőképpen írható fel:

$$d_{ij} = \frac{m_{ij}}{\bar{m}_{.j}} \tag{12}$$

A továbbiakban jelöljük a d_{ij} általános elemű mátrixot **D**-vel. Fontos észrevenni, hogy amennyiben az m_{ij} az **M** mátrix esetében hiányzó adatpont volt, úgy a **D** mátrix d_{ij} eleme is hiányzik, és fordítva, ha az **M** mátrix esetében m_{ij} egy megfigyelt adatpontot jelöl, akkor d_{ij} sem hiányozhat a **D** mátrix esetében.

A harmadik lépésben e százalékos értékek átlagolásra kerülnek, olyan módon, hogy egy mutató esetében az adott régió összegzett értéke a megfigyelt (vagyis nem hiányzó) években kapott d_{ij} hányadosainak átlaga, amelyek praktikusan a **D** mátrix hiányzó adatokat nélkülöző sorátlagai.

$$\bar{d}_{i.} = \frac{\sum_{j=1}^{b} d_{ij}}{\sum_{j=1}^{b} (1 - f_{ij})} \tag{13}$$

A következő, negyedik lépésben elkészítünk egy becsült adatpanelt, amelyhez a peremszámokat, vagyis **M** oszlopátlagait és **D** sorátlagait használjuk fel, így megkapva **N** becsült mátrixot, amelynek általános eleme n_{ij} az alábbiak szerint írható fel:

$$n_{ij} = \bar{m}_{.j} \bar{d}_{i.} \tag{14}$$

Végül az adathiányok behelyettesítése az utolsó lépés, amely a korábbiakhoz hasonlóan formalizálható:

$$m_{ij} = \begin{cases} m_{ij} \mid f_{ij} = 0 \\ n_{ij} \mid f_{ij} = 1 \end{cases} \tag{15}$$

Bizonyos paneltípusoknál előfordulhat, hogy egyes lépések egyszerűsíthetők vagy pontosíthatók. Az elemzéseinkben felhasználthoz hasonló területi elemzések esetében gyakran előfordul, hogy az adott megfigyelési egységre nem, viszont magasabb aggregálási szinten elérhetők az adatok. Ilyenre lehet példa, ha egy ország esetében regionális vagy települési szinten hiányzó adatokkal szembesülünk, viszont az adott ország viszonylatában az idősor hiánytalan. Ebben az esetben a (7) egyenlet alapján becsült oszlopátlagok helyett ezen ismert általgokat is felhasználhatjuk. További pontosítás lehet, ha a (7), illetve a (12)–(14) egyenletben bemutatott módszert nem a terület egészére, hanem annak egy alegységére végezzük el, például az országos helyett vármegyei éves átlagokhoz viszonyítva számítjuk ezen egyenleteinket. Ahogy a korábbi módszerek esetében, itt is van egy feltétel, amelynek szükséges a teljesülése: ahhoz, hogy a módszer alkalmazható legyen, minden megfigyelési egységre és minden időszakra legalább egy megfigyelt adattal rendelkezni kell.

Elemzésünkben összesen tizenegy imputálási módszert alkalmaztunk: az átlag és a medián helyettesítésének a fent bemutatott három (teljes panel alapú, településeket figyelembe vevő sor alapú és az éveket figyelembe vevő oszlop alapú) válfaját mindkét középérték esetén, az utolsó megfigyelés továbbvitelét, valamint a regressziós helyettesítéssel való becslést a fent bemutatottak szerint, továbbá a peremek alapján való imputáció három alfaját: az alapeljárást, egy olyan változatot, ahol az oszlopátlagok ismertek, valamint egy olyan verziót, ahol a (7) és a (12)–(14) egyenletet a települések esetében várvármegyék szerinti bontásban számoltuk ki.

3. Módszertan

Tanulmányunkban az imputációs módszerek hatékonyságának tesztelésére az alábbi, hatlépéses eljárást alkalmaztuk. Először ahhoz, hogy a folyamat további lépéseit megfelelően tudjuk végrehajtani, olyan adatbázist kerestünk, amelyben paneltípusú adatok konzisztensen, hiánymentesen, számos megfigyelési egységre és viszonylag hosszabb időtávra rendszeresen rendelkezésre állnak. Ilyen adatbázis az Országos Területfejlesztési és Területrendezési Információs Rendszer (TeIR), amely Magyarországon települések szintjén tartalmaz különböző hivatalos forrásokból származó adatokat. Előnye, hogy Magyarország közel minden településére tartalmaz adatot[1] számos terület különböző indikátorai esetében éves bontásban, így összesen 3153 település került a mintánkba, minden választott ismérv esetében, a 2011–2019 közötti teljes időszakra.

Az adatok kiválasztása során több szempontot figyelembe vettünk. Ahhoz, hogy általánosítható iránymutatással szolgálhassunk a hiányzó adatok kezelése terén, fontosnak tartottuk, hogy az adataink között szerepeljenek olyanok, amelyek esetében az értékek markáns, növekvő vagy csökkenő trendet követnek, illetve olyanok is, amelyek jellemzően stagnálnak. Ezenfelül az évenkénti változékonyságot figyelembe véve olyan mutatókat is bevontunk az elemzésbe, amelyek évről évre lassú ütemben változnak, de olyan is akad, amely esetében az értékek évente kisebb vagy akár extrém mértékben is váltakozni képesek. Ez alapján a 2. táblázatban bemutatott öt ismérvet vontuk be elemzésünkbe. Az évek közötti trend meredekségének megállapításában a 3153 település évenként vett mediánértékének

[1] Ez alól kivétel Balatonkenese, amelyből 2014-ben kivált Balatonakarattya, így 2014 előtt egy, míg ezt követően két településként szerepelnek az adatbázisban. Mivel a két mai település esetében az adatok időben nem konzisztensek, ezért ez a két település kizárásra került a mintánkból.

relatív szórása adta az alapot. Ezen túlmenően az egyes települések esetében az adatok fluktuációját is elemeztük, amely alapjául a települések idősorára illesztett lineáris trendvonalának illeszkedése szolgált. Amennyiben az LNM-módszer alapján legjobban illeszkedő trendegyenes a település értékeiben a variancia legalább 70%-át megmagyarázta, úgy szinte egyáltalán nem fluktuálóként, kiegyenlítettként azonosítottuk az ismérvet. Amennyiben az R^2 értéke ezen egyenes esetében 30% és 70% között volt, akkor az ismérvet enyhén, míg legfeljebb 30%-os magyarázott variancia esetén erősen fluktuáló ismérvnek kategorizáltuk.

2. táblázat

A tanulmányban elemzett ismérvek karakterisztikái
Characteristics of the variables analyzed in the study

Ismérv	Mértékegység	Relatív szórás (medián), %	Lineáris trend R^2	Trend iránya	Fluktuáció
65 év felettiek aránya a lakosságban	%	0,79	0,495	Stagnáló	Enyhe
Egy lakosra jutó jövedelem[a)]	Forint/fő	28,72	0,929	Növekvő	Kiegyenlített
Ezer lakosra jutó élveszületések száma	Születés/ 1000 fő	4,69	0,163	Stagnáló	Erős
Házasságkötések száma	Darab	13,36	0,318	Növekvő	Enyhe
Őstermelők száma	Fő	5,96	0,483	Csökkenő	Enyhe

a) A mutató az egy lakosra jutó szja-adóalapot képező belföldi jövedelem (forint) és az adott év január 1-jei állapot szerinti állandó népesség (fő) hányadosát mutatja meg.
Megjegyzés: minden ismérv esetében N = 3153, és a megfigyelési időtáv 2011–2019. A trend és a fluktuáció rendre az idősor mediánértékeinek relatív szórása, valamint az adatokra településenként illesztett trendegyenes R^2 értéke alapján történt. 5% alatti relatív szórás esetében stagnáló, annál nagyobb esetben növekvő vagy csökkenő trendű ismérvről beszélünk, míg a fluktuáció esetében a magyarázott variancia 70% feletti, akkor kiegyenlített, 30–70% közötti R^2 esetében enyhe fluktuáció, 30% alatti R^2 esetében erős fluktuáció jellemzi az adott ismérvet.
Forrás: saját szerkesztés.

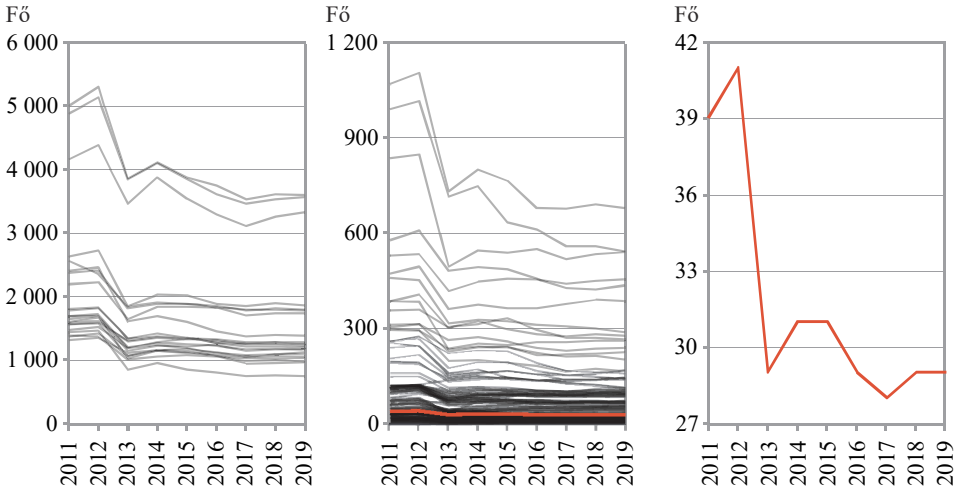
A 2. táblázatból látható, hogy a mediánérték relatív szórása két esetben 5% alatt maradt, így ezeket az ismérveket stagnálóként azonosítottuk. A relatív szórás mellett a trendek meghatározására a Függelék F1. ábrái is segítséget nyújtottak. Ezen információk felhasználásával beazonosíthattuk, hogy a maradék három ismérvből kettő esetében növekvő, míg egy esetben csökkenő trendet fedezhetünk fel. Az így kapott kategóriákat a függelékben szemléltetett lefutásdiagramok is megerősítik. Ezek alapján az egy lakosra jutó jövedelem esetében minimális fluktuációt tapasztaltunk, vagyis a települések jelentős hányada egy jól látható trendet követett az értékek növekedése során. Enyhén fluktuálóként azonosítottuk a 65 év felettiek aránya a lakosságban, valamint a házasságkötések száma ismérveket, amelyek esetében évről évre legfeljebb kisebb változások figyelhetők meg az egyes települések

értékeiben, azonban a 9 éves időtartam alatt már bizonyos települések viszonylag nagy eltéréseket is képesek voltak mutatni. Erős fluktuáció figyelhető meg az ezer lakosra jutó születésszámban, ahol a települések nagy része esetében akár egyik évről a másikra is jelentős változás következhetett be a mutató értékében. Az őstermelők ismérv esetében az idősorok lefutása a fenti kategorizálás alapján csökkenő és enyhén fluktuáló, azonban a települések idősorait szemléltetve további érdekességre lehetünk figyelmesek (1. ábra). Látható, hogy az időperiódusunk harmadik évében egy jelentős esés következett be az idősorokban, de az értékek ez előtt és ez után is nagyjából stagnálni, enyhén csökkenni látszanak. Továbbá ez a jelenség nemcsak néhány, hanem szinte minden település esetében megfigyelhető.

1. ábra

Az őstermelők ismérv idősorainak lefutása

The time series trends of the variable primary producers



Megjegyzés: az ábrán az évenkénti legnagyobb értékeket mutató települések mértéke (bal oldali panel), a bal oldali panel által nem érintett településekből vett véletlenszerű 100 településes minta értékei (középső panel) és a 3153 település éves medián értékei (középső panel narancssárga vonal és jobb oldali panel) láthatók.
Forrás: saját szerkesztés.

Az eljárásunk második lépéseként az egyes adattáblákból véletlenszerűen töröltük a megfigyelések 5%-át, aminek eredményeként kialakult **F** adathiányindikátor-mátrixunk is. Harmadik lépésként a 2.2. alfejezetben bemutatott tizenegy imputálási eljárás alapján megbecsültük a hiányzó adatok értékét. Negyedik lépésként az eredeti adattábla azon elemeit, amelyek a második lépésben törlésre kerültek, egy vektorba (***e***) rendeztük, majd az imputált mátrix becsült, helyettesítő elemeit szintén egy vektorba (***h***) rendezve használtuk tovább. E két vektor értelemszerűen azonos méretű lett, és egyes elemeik ugyanazon megfigyelési egység–év

pár eredeti, illetve imputált értékeit reprezentálják. E tulajdonságot kihasználva kiszámítottunk egy harmadik, $\boldsymbol{k}$ különbségvektort, amely az imputálási eljárásunk hiányzó adatpontonként fellépő hibáját hivatott mérni:

$$\mathbf{k} = \mathbf{e} - \mathbf{h} \tag{16}$$

Az ötödik lépésben az $\boldsymbol{e}$ és a $\boldsymbol{h}$, illetve a $\boldsymbol{k}$ vektort különböző illeszkedés vagy eltérést mérő mutatók segítségével összehasonlítottuk. Az eltérés karakterisztikáit az $\boldsymbol{e}$ és a $\boldsymbol{h}$ vektor esetében korrelációelemzéssel és párosított mintás t-próbával vizsgáltuk. Az eltéréseket az alapján értékeltük, az eredeti és a helyettesített értékek mennyire mozogtak együtt, milyen szoros korrelációs kapcsolat állt fent közöttük, valamint hogy az eredeti és az imputált értékek különbsége hány esetben bizonyult 0-tól eltérőnek 5, illetve 1%-os szignifikanciaszinten. A különbségvektorok elemzésénél a regressziós modellezésben gyakran alkalmazott illeszkedésmutatókat, az átlagos abszolút hiba (*mean absolute error*, MAE), valamint az átlagos négyzetes hiba (*mean squared error*, MSE) mutatókat *(Alwateer et al., 2024)* adaptáltuk, és a $\boldsymbol{k}$ vektor esetében kiszámítottuk a

$$\frac{\sum|\mathbf{k}|}{\sum f_{ij}}, \tag{17}$$

valamint a

$$\frac{\sum(\mathbf{k})^2}{\sum f_{ij}} \tag{18}$$

értékeket, vagyis azt, hogy a $\boldsymbol{k}$ vektor elemei abszolút értékének és négyzeteinek átlaga mekkora. Emellett kiszámítottuk, hogy mekkora a $\boldsymbol{k}$ vektor terjedelme, vagyis mekkora volt a legnagyobb és a legkisebb eltérés közötti különbség az $\boldsymbol{e}$ és a $\boldsymbol{h}$ vektor páronként vett értékei között.

A 2–5. lépéseket ismérvenként összesen 1000 alkalommal végeztük el, aminek köszönhetően minden illeszkedést és különbséget mérő mutató esetében ismérvenként 1000 értéket kaptunk. Hatodik lépésben ezen értékeket elemeztük, ennek tárgyalását a 4. fejezet tartalmazza.

4. Eredmények

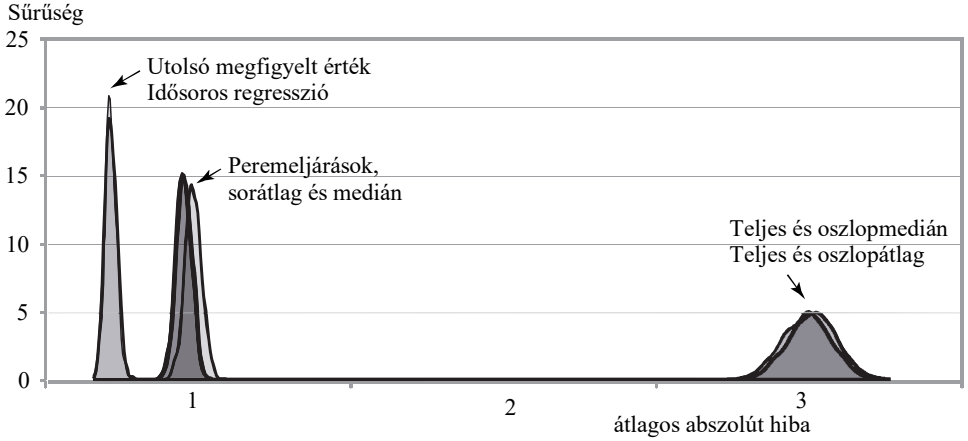
Az alábbiakban a 2. táblázatban bemutatott öt ismérv esetében részletesen elemezzük a módszertani fejezetben bevezetett illeszkedésmutatókat, feltárjuk a hasonlóságokat és különbségeket az egyes eljárások alkalmazása kapcsán, valamint javaslatot is teszünk a szimuláció eredményei alapján általánosan alkalmazható eljárásokra.

A 65 éven felüliek aránya ismérv esetében az átlagos abszolút hiba (2. ábra), az átlagos négyzetes hiba, valamint a hibák szórása esetében is látható, hogy a regressziós becslés, valamint az utolsó megfigyelés továbbvitele módszerei rendelkeznek a legjobb értékekkel. Az illeszkedésmutatók minimumából és maximumából látható (Függelék F1. táblázat), hogy a két módszer lényegesen alacsonyabb eltérésértékekkel rendelkezik mindhárom mutató esetében, mint a többi eljárás. Ezt követően szintén e három illeszkedésmutató alapján öt további módszer közepes lemaradással következik. E módszerek a három peremalapú eljárás, valamint a két sor középértékkel (átlag és medián) való helyettesítés. Ezt követi a többi középértékkel való helyettesítési módszer, ám ezek lemaradása minden metrika alapján jelentős. A hibák terjedelme tekintetében szintén ez a sorrend látszik kirajzolódni, a legkisebb különbségek az imputált és az eredeti értékek között átlagosan az utolsó megfigyelt értékek továbbvitele, valamint a regressziós becslésen alapuló eljárás esetében voltak.

2. ábra

Az átlagos abszolút hiba (MAE) mutató értékei alapján, 1000 szimulációt követően összeállított sűrűségdiagram a 65 éven felüliek aránya ismérvhez

Density plot based on mean absolute error (MAE) values after 1000 iterations for the variable of the proportion of individuals aged 65 and over



Forrás: saját szerkesztés.

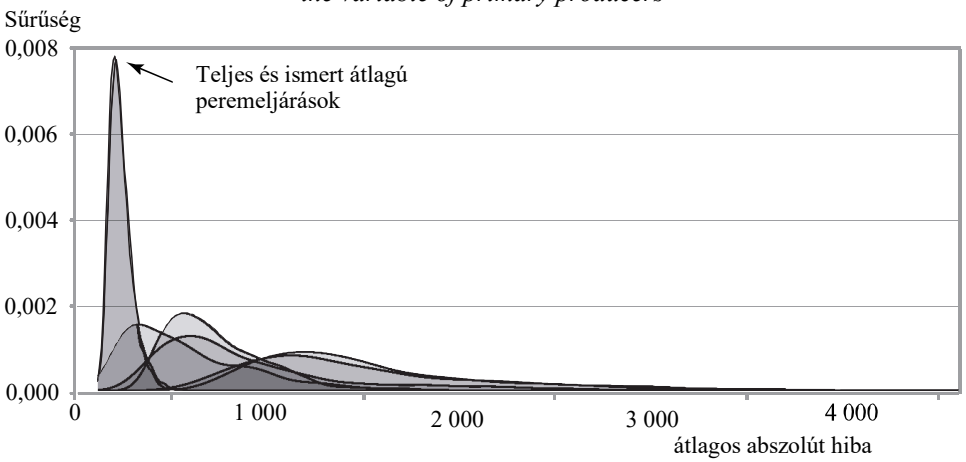
A többi ismérv esetében az átlagos abszolút hibát, az átlagos négyzetes hibát, valamint a hibák szórását vizsgálva jellemzően a peremalapú eljárások teljesítettek legjobban. Az ezer lakosra jutó élveszületések száma esetében e három módszer mellett a soralapú átlag és a sormedián eljárás volt még kiemelkedő, míg a házasságkötések száma ismérv esetében e módszerek ugyan megfelelően teljesítettek a hibaalapú illeszkedésmutatók alapján, azonban az utolsó megfigyelt érték továbbvitele és az idősoros regresszió eljárások is jó értékeket mutattak. Az egy lakosra jutó

jövedelem, illetve az őstermelők száma ismérvek vonatkozásában a peremeljárások mellett szintén az utolsó megfigyelt érték továbbvitele és az idősoros regresszió módszerei voltak még kiemelkedők, azonban ezen esetekben a soralapú eljárásokkal kapcsolatban a többi középértékkel való helyettesítéses módszerekhez hasonlóan gyenge illeszkedésmutatókat tapasztalhattunk (3. ábra). A leggyengébben minden ismérv esetében a teljes panelt, valamint az éves oszlopokat figyelembe vevő átlag és medián helyettesítés módszerei teljesíttetek. A peremeljárások közül a vármegyei átlagokból becsült csoportosított módszer teljesítménye hullámzó volt a másik két peremeljáráshoz képest, az egy lakosra jutó jövedelem esetében enyhén ugyan, de jobb illeszkedésmutatókat figyelhettünk meg, míg a többi ismérv esetében jellemzően elmaradt a teljesítménye a másik két peremeljárás mögött.

3. ábra

Az átlagos négyzetes hiba (MSE-) mutató értékei alapján, 1000 szimulációt követően összeállított sűrűségdiagram az őstermelők ismérvhez

Density plot based on mean squared error (MSE) values after 1000 iterations for the variable of primary producers



Forrás: saját szerkesztés.

Az eredeti, törölt értékek és az imputált értékek közötti Pearson-féle korrelációs együttható a 65 év feletti lakosok aránya, az ezer lakosra jutó élveszületések száma és a házasságkötések száma ismérvek esetében is hasonló. Ezen ismérveknél az oszlopalapú átlaggal és mediánnal való helyettesítés esetében nagyon alacsony (0,1 alatti) értéket vett fel, míg minden további imputálási eljárás[2] esetében ez az érték közepesen erős vagy erős, pozitív irányú kapcsolatot mutatott. Az egy lakosra jutó jövedelem ismérv esetében a peremalapú három eljárás, valamint az

[2] A teljes adattábla átlagával és mediánjával való helyettesítés esetében a $\boldsymbol{h}$ vektorunk konstans, így e két módszer esetében a korrelációs együtthatók nem számíthatók ki.

utolsó megfigyelt érték továbbvitele és az idősoros regressziós eljárások is magas és pozitív korrelációs együtthatóval bírnak (R > 0,9), viszont a középértékkel való helyettesítési eljárások esetében a Pearson-féle korrelációs együttható értéke 0,6 és 0,7 közötti értéket vesz fel. A nagyon fluktuáló élveszületések száma esetében nem meglepően globálisan alacsony korrelációs együtthatókat láthatunk. Ezen ismérv esetében az R értéke az oszlopalapú helyettesítési eljárások esetében 0,1 alatti, az utolsó megfigyelés továbbvitele esetében átlagosan 0,3 körüli, míg a többi eljárás esetében átlagosan 0,5 körüli értéket vesz fel.

Elemzésünk során a párosított mintás t-próbákat is megvizsgáltuk minden ismérvre. Azt teszteltük, hogy mind $\alpha = 5$, mind $\alpha = 1\%$-os szignifikanciaszint mellett az eredeti és az imputált értékek átlagának különbsége az 1000 szimulációból – indikátoronként – hány százalékában bizonyult statisztikailag szignifikánsnak.[3] A soralapú átlag, valamint az általános és az ismert átlaggal számoló peremeljárások esetén minden ismérv tekintetében elmondható, hogy alacsony arányban lettek szignifikánsak a párosított mintás t-próbáink. E három eljárás $\alpha = 5\%$-os szinten az esetek 4,1 és 7,2% közötti mértékben, míg $\alpha = 1\%$-os szignifikanciaszint mellett 0,5 és 2,2% közötti mértékben mutatott szignifikánsan eltérő átlagot az eredeti és az imputált értékek között. Hasonlóan jól, 10% alatti szignifikáns t-próba aránnyal a többi módszer csak bizonyos ismérvek esetében teljesített: a csoportos peremeljárás az 1000 lakosra jutó élveszületések száma, az egy lakosra jutó jövedelem, valamint a 65 év feletti lakosok aránya esetében illeszkedett jól, az oszlop- és a panelalapú átlag eljárás a házasságkötések száma kivételével a többi négy ismérv esetében alacsony arányban produkált szignifikáns t-teszt-értékeket. Az utolsó megfigyelt érték továbbvitele eljárás csak az 1000 lakosra jutó élveszületések száma, valamint a 65 év feletti lakosok aránya esetében mutatott kevés esetben eltérést, az idősoros regressziós eljárás pedig a házasságkötések száma, illetve az 1000 lakosra jutó élveszületések száma esetében bizonyult megfelelő eljárásnak a t-tesztek alapján. A medián típusú módszerek mindegyikéről elmondható, hogy a többi mutatóhoz hasonlóan a t-tesztek alapján is rossz becslési eredményeket mutattak (3. táblázat).

[3] Fontos megjegyezni, hogy mivel a t-teszt a két vektor átlagának összehasonlítására szolgál, az átlagalapú imputációknál teljesen véletlenszerű adathiány esetében vélhetően valamelyest e mutatók „javára” torzítanak a tesztek.

3. táblázat

Az alkalmazott illeszkedésmutatók eredményei ismérvenként és imputációs módszerenként

Goodness of fit measures for each variable and imputation method

Eljárás	Átlagos abszolút hiba	Átlagos négyzetes hiba	Hibák szórása	Korrelációs együttható	Hibák tartománya
65 éven felüliek aránya					
Sorátlag	**0,8541**	**1,6415**	**1,2800**	**0,9534**	**17,4909**
Oszlopátlag	2,9703	18,0373	4,2444	0,0165	42,2606
Teljes átlag	2,9706	18,0393	4,2446		42,1933
Sormedián	0,8823	1,8304	**1,3510**	**0,9482**	**18,8369**
Oszlopmedián	2,9475	18,1893	4,2443	0,0168	42,2597
Teljes medián	2,9483	18,1899	4,2446		42,1933
Perem	**0,8497**	**1,6317**	**1,2761**	**0,9537**	**17,5875**
Perem, ismert átlag	**0,8493**	**1,6304**	**1,2756**	**0,9537**	**17,5821**
Perem vármegyénként	**0,8453**	**1,6112**	**1,2681**	**0,9542**	**17,5169**
Utolsó megfigyelés	**0,6057**	**0,8912**	**0,9422**	**0,9753**	**15,5673**
Regressziós becslés	**0,6025**	**0,8961**	**0,9448**	**0,9750**	**15,6227**
Egy lakosra jutó jövedelem					
Sorátlag	2,150E + 05	6,934E + 10	2,633E + 05	0,6414	1,907E + 06
Oszlopátlag	2,019E + 05	6,753E + 10	2,598E + 05	0,6467	2,353E + 06
Teljes átlag	2,681E + 05	1,161E + 11	3,406E + 05		2,482E + 06
Sormedián	2,182E + 05	7,409E + 10	2,673E + 05	0,6240	1,995E + 06
Oszlopmedián	2,015E + 05	6,784E + 10	2,598E + 05	0,6467	2,354E + 06
Teljes medián	2,646E + 05	1,190E + 11	3,406E + 05		2,482E + 06
Perem	**4,249E + 04**	**4,945E + 09**	**6,975E + 04**	**0,9790**	**1,434E + 06**
Perem, ismert átlag	**4,248E + 04**	**4,943E + 09**	**6,974E + 04**	**0,9790**	**1,433E + 06**
Perem vármegyénként	**4,061E + 04**	**4,637E + 09**	**6,745E + 04**	**0,9801**	**1,407E + 06**
Utolsó megfigyelés	8,818E + 04	1,252E + 10	8,453E + 04	**0,9747**	**1,375E + 06**
Regressziós becslés	6,506E + 04	8,121E + 09	8,936E + 04	**0,9649**	**1,337E + 06**
1000 lakosra jutó élveszületések aránya					
Sorátlag	**3,1618**	**22,9764**	**4,7819**	0,4910	69,0702
Oszlopátlag	3,6987	29,2782	5,4014	0,0756	**58,9741**
Teljes átlag	3,7138	29,4430	5,4166		**58,5212**
Sormedián	**3,1769**	25,2398	**4,9963**	0,4544	73,4405
Oszlopmedián	3,6720	29,4904	5,4013	0,0758	**58,8470**
Teljes medián	3,6873	29,6706	5,4166		**58,5212**
Perem	**3,1381**	**22,7050**	**4,7533**	0,4993	**68,6363**
Perem, ismert átlag	**3,1371**	**22,6907**	**4,7518**	0,4997	**68,6196**
Perem vármegyénként	**3,1416**	**22,7238**	**4,7553**	0,4993	**68,4884**
Utolsó megfigyelés	4,0246	40,6232	6,3592	0,3117	102,9915
Regressziós becslés	**3,3307**	26,3224	5,1191	0,4453	80,2009

(A táblázat a következő oldalon folytatódik)

(folytatás)

Eljárás	Átlagos abszolút hiba	Átlagos négyzetes hiba	Hibák szórása	Korrelációs együttható	Hibák tartománya
Házasságkötések száma					
Sorátlag	3,6488	462,7095	16,0991	**0,9857**	579,1225
Oszlopátlag	18,3753	2,562E + 04	120,2168	0,0310	3,728E + 03
Teljes átlag	18,5105	2,561E + 04	120,2314		3,721E + 03
Sormedián	3,8084	542,6386	17,8483	**0,9841**	643,7765
Oszlopmedián	13,1382	2,573E + 04	120,1979	0,0399	3,723E + 03
Teljes medián	13,2062	2,574E + 04	120,2314		3,721E + 03
Perem	**2,9453**	1,027E + 03	20,7036	**0,9949**	734,7052
Perem, ismert átlag	**2,6203**	**276,8080**	**12,1242**	**0,9958**	**433,5688**
Perem vármegyénként	4,8349	9,235E + 03	65,2737	**0,9746**	2,207E + 03
Utolsó megfigyelés	**3,1730**	**209,7128**	**10,1747**	**0,9930**	**310,7350**
Regressziós becslés	**2,7991**	**190,7410**	**10,1284**	**0,9941**	**343,7912**
Őstermelők száma					
Sorátlag	13,5776	1,513E + 03	38,2446	**0,9834**	988,6135
Oszlopátlag	97,0494	4,368E + 04	206,5599	0,0629	3,632E + 03
Teljes átlag	97,4053	4,384E + 04	206,9493		3,611E + 03
Sormedián	12,0247	1,553E + 03	38,1873	**0,9845**	933,4655
Oszlopmedián	77,9272	4,740E + 04	206,7013	0,0651	3,618E + 03
Teljes medián	78,0897	4,749E + 04	206,9493		3,611E + 03
Perem	**6,7704**	**236,7722**	**15,2590**	**0,9972**	**360,6282**
Perem, ismert átlag	**6,7067**	**229,1890**	**15,0201**	**0,9973**	**353,7087**
Perem vármegyénként	**7,5753**	597,0320	**23,3942**	**0,9955**	633,5320
Utolsó megfigyelés	**7,9928**	929,8822	29,1234	**0,9914**	798,8630
Regressziós becslés	9,8326	738,6560	26,7383	**0,9919**	720,8337

Megjegyzés: félkövérrel az adott változó adott illeszkedésmutatója szerinti legkedvezőbb értékeket jelöltük. Az értékek az 1000 szimuláció esetében mért illeszkedésmutatók átlagai.
Forrás: saját szerkesztés.

Ahol nagyobb arányban bizonyultak szignifikánsan eltérőnek a t-tesztek alapján az eredeti és az imputált értékeket tartalmazó vektorok, ott érdemes megtekinteni, hogy milyen jellegű különbségek adódnak. A 4. táblázatban a párosított mintás t-próba átlagra vonatkozó oszlopa[4] alapján megállapítható, hogy a legtöbb esetben az értékek alulbecslése történt. Az utolsó megfigyelés továbbvitele módszer esetében nem meglepő, hogy növekvő trend (egy lakosra jutó jövedelem) esetében a szisztematikus tévedések alul-, csökkenő trend (őstermelők száma) esetében pedig túlbecsülték az eredeti értékeket. A többi eljárás esetében általánosan megál-

[4] A párosított mintás t-próbákat H_0: $\mathbf{e} - \mathbf{h} = 0$ módon formalizált nullhipotézis tesztelésére végeztük el, ez alapján a pozitív empirikus tesztértékek az imputációs eljárás rendszerszintű alulbecslésére utalhatnak, azaz arra, hogy az eredeti adatokat tartalmazó $\boldsymbol{e}$ vektor elemei nagyobbak, mint az imputált értékeket tartalmazó $\boldsymbol{h}$ vektor elemei.

lapítható, hogy amennyiben magas volt a t-tesztek közül a statisztikailag szignifikáns eredmények aránya, úgy az alulbecslésre utaló pozitív átlagos t-tesztértékkel találkozhattunk.

Összességében látható, hogy a 65 éven felüliek aránya ismérv esetében a tizenegy eljárás közül az utolsó megfigyelés továbbvitele a legjobb, a többi ismérvnél viszont az illeszkedésmutatóink körének egészét figyelembe véve a peremalapú eljárások bizonyultak a legjobbnak. Általánosan jellemző, hogy a középértékkel való helyettesítés a többi eljárással szemben gyengébb teljesítményt nyújtott, ami alól néhány esetben az adott megfigyeléseket figyelembe vevő, soralapú középértékkel való helyettesítés mutatott kivételt.

4. táblázat

Az alkalmazott hipotézisellenőrzések eredményei ismérvenként és imputációs módszerenként

The results of the t-tests for each variable and imputation method

Eljárás	Párosított mintás t-próba átlaga	5%-on szignifikáns t-próbák aránya	1%-on szignifikáns t-próbák aránya
		%	
65 éven felüliek aránya			
Sorátlag	–0,0298	**5,0**	**0,8**
Oszlopátlag	–0,0161	**6,6**	**1,5**
Teljes átlag	–0,0153	**6,5**	**1,5**
Sormedián	0,9515	16,3	5,9
Oszlopmedián	3,4684	93,3	81,1
Teljes medián	3,4398	93,1	80,3
Perem	–0,0213	**5,5**	**0,9**
Perem, ismert átlag	–0,0316	**5,2**	**0,9**
Perem vármegyénként	0,1056	**5,5**	**1,1**
Utolsó megfigyelés	–0,2321	**7,3**	**1,6**
Regressziós becslés	0,6916	10,3	**2,2**
Egy lakosra jutó jövedelem			
Sorátlag	0,0062	**4,1**	**1,0**
Oszlopátlag	0,0285	**6,4**	**1,5**
Teljes átlag	0,0142	**5,0**	**1,9**
Sormedián	7,2226	100,0	100,0
Oszlopmedián	2,5090	70,0	48,0
Teljes medián	5,9869	100,0	100,0
Perem	–0,0039	**5,4**	**1,0**
Perem, ismert átlag	–0,0168	**5,7**	**1,2**
Perem vármegyénként	0,1410	**5,1**	**0,8**
Utolsó megfigyelés	32,7592	100,0	100,0
Regressziós becslés	3,8050	96,4	88,8

(A táblázat a következő oldalon folytatódik)

(folytatás)

Eljárás	Párosított mintás t-próba átlaga	5%-on szignifikáns t-próbák aránya	1%-on szignifikáns t-próbák aránya
		%	
1000 lakosra jutó élveszületések aránya			
Sorátlag	–0,0549	**6,0**	**1,3**
Oszlopátlag	–0,0566	**6,1**	**1,7**
Teljes átlag	–0,0580	**5,9**	**1,7**
Sormedián	2,7582	79,1	58,9
Oszlopmedián	3,2117	90,5	72,8
Teljes medián	3,3060	91,8	76,9
Perem	–0,0512	**5,7**	**1,6**
Perem, ismert átlag	–0,0535	**5,7**	**1,6**
Perem vármegyénként	–0,0189	**5,9**	**1,4**
Utolsó megfigyelés	0,4245	**8,5**	**1,7**
Regressziós becslés	–0,3318	**7,4**	**2,2**
Házasságkötések száma			
Sorátlag	–0,0821	**4,4**	**0,5**
Oszlopátlag	–1,5071	42,6	31,7
Teljes átlag	–1,5041	42,3	31,4
Sormedián	1,3850	33,8	15,0
Oszlopmedián	5,5580	98,3	69,9
Teljes medián	5,6412	98,6	70,4
Perem	0,8707	**7,2**	**2,2**
Perem, ismert átlag	0,2830	**5,2**	**1,3**
Perem vármegyénként	2,0909	58,1	31,0
Utolsó megfigyelés	4,6485	95,4	91,3
Regressziós becslés	0,4724	**7,5**	**1,0**
Őstermelők száma			
Sorátlag	–0,0137	**5,1**	**1,1**
Oszlopátlag	–0,1074	**6,1**	**2,6**
Teljes átlag	–0,1054	**6,0**	**2,1**
Sormedián	5,0054	100,0	100,0
Oszlopmedián	11,1699	100,0	100,0
Teljes medián	11,1415	100,0	100,0
Perem	0,3835	**7,1**	**1,4**
Perem, ismert átlag	0,0366	**6,5**	**1,2**
Perem vármegyénként	2,6211	84,2	53,5
Utolsó megfigyelés	–5,0666	99,1	98,1
Regressziós becslés	1,5174	31,6	15,0

Megjegyzés: félkövérrel az adott változó adott illeszkedésmutatója szerinti legkedvezőbb értékeket jelöltük. Az értékek az 1000 szimuláció esetében mért illeszkedésmutatók átlagai.
Forrás: saját szerkesztés.

Az általános és az ismert átlaggal működő peremalapú eljárások minden esetben legalább elfogadhatóan teljesítettek az illeszkedésmutatók széles körén, ráadásul a legtöbb, ötből négy ismérv esetében ezen eljárások bizonyultak összességében is a legjobbnak. A 65 év felettiek aránya ismérvnél a legjobban az előző megfigyelés továbbvitele mutatkozott, azonban ezen eljárás, valamint az idősoros trendvonal alapú regresszió néhány ismérv, és kifejezetten a t-próbák alapján gyengébb módszernek bizonyult. A középértékkel való helyettesítési megoldásoknál jellemzően a szofisztikáltabb technikák jobban becsülték vissza a hiányzó adatainkat.

A peremeljárások közül teljesen véletlenszerűen hiányzó adatok esetében az ismert sokasági átlag általában nem, vagy csak enyhén javított az illeszkedésmutatókon. Ez alól kivétel a házasságkötések száma ismérv, ahol a részletes táblázatok alapján látható (Függelék F1. táblázat), hogy míg az 1000 szimuláció közül a kisebb hibák mértéke a két eljárásnál hasonló mértékű, addig azon szimulációk esetében, amikor ezen eljárások viszonylag nagyot tévedtek, a tévedés mértéke az ismert átlagú eljárás esetében lényegesen kisebb volt. Ennek következtében fordulhatott elő az, hogy míg ezen ismérv esetében a két peremeljárás átlagos négyzetes hibája az 1000 szimulációra vetítve átlagosan nagyságrendekkel eltérő volt, addig az 1000 szimuláció mediánja közel azonos volt mindkét érintett peremeljárás esetében. A vármegyékkel alkotott csoportok alapján számított peremeloszlás esetében az illeszkedésmutatók jelentősen eltértek a másik két módszertől. Egyes esetekben, ahol a csoportosítás vélhetően nagyobb hatással volt az ismérv által felvett értékre, a csoportos verzió valamivel jobban teljesített, de általánosságban elmondható, hogy gyengébb illeszkedésmutatókat eredményezett ezen eljárás, mint a két másik hasonló módszer. Érdemes lehet a jövőben további, homogénebbnek tekinthető csoportosítási alapokat is górcső alá venni, ilyenek lehetnek a kisebb területi egységek vagy a településtípus alapján alkotott csoportosítás.

Tesztjeinket robusztusságellenőrzés céljából 10%-os arányú adathiány esetében is lefuttattuk, és az eredményeink közel azonosak, a végkövetkeztetésünk pedig ezen adatok esetében is teljesen azonos volt. Jövőbeli célunk az adathiány további mértékei mellett az elemzett eljárások pontosságának ismételt vizsgálata, hiszen minél nagyobb az adathiány mértéke, annál nagyobb torzításokra számíthatunk egyes eljárások esetén *(Kleinke et al., 2011)*. Fontos jövőbeli kutatási irányként azonosítjuk továbbá vizsgálatunk elvégzését egyéb hiányzási struktúrák (MAR és MNAR) mellett is.

5. Konklúzió

Tanulmányunkban a hiányzó adatok imputációs technikákkal történő visszabecslésének hatékonyságát vizsgáltuk azon esetekben, amikor mindössze egy változó, azonban annak kapcsán több időbeli megfigyelés áll rendelkezésre. Ez az eset a nem felmérésen (*non-survey*) alapuló adatok esetében igen gyakori kiindulási pont számos kutatás esetében. *Alwateer és szerzőtársai (2024)* alapján a szakirodalomban kiemelkedően fontos, szükséges pontként jelenik meg az imputációs technikák körültekintő értékelése, amely jelenleg kutatási résként azonosítható irány. Tanulmányunkban összesen öt, különböző adatstruktúrájú ismérv 9 éves paneladatait vetettük alá tizenegy különböző, egyszerűen kivitelezhető és széles körben alkalmazott imputációs eljárásnak. Eredményeinket egy olyan szimulációs vizsgálatra építettük, amely alapja minden változó és minden módszer esetében egy 1000-szer ismételt szimulációs, véletlenszerű adattörlés volt, először 5, majd 10%-os adathiányt generálva. Az imputált adatok helytállóságának tesztelésére a szakirodalom javaslatai alapján *(Alwateer et al., 2024; Máder, 2005)* szintén diverz mutatótárral vállalkoztunk: több mutató mentén elemeztük az eredeti és az imputált adatok közti különbségeket, azaz a visszabecslések hibáit, valamint korrelációanalízissel és párosított mintás t-próbákkal is alátámasztottuk konklúzióinkat.

Eredményeink alapján a peremeloszlásra építő imputációs módszer általánosságban kiemelkedett a többi vizsgált eljárás közül, azaz általánosan jó megoldásnak bizonyul az ismérv karakterisztikáitól (például stagnáló vagy trendszerű, fluktuáló vagy stabil) függetlenül. Megállapítottuk továbbá, hogy ezen eljárások esetén bár az ismert országos vagy éves átlagok rendelkezésre állása némileg növeli a pontosságot, hiányuk esetén sem csökken jelentősen az eljárás hatékonysága. Tanulmányunk további konklúziója, hogy a csoportosítást alkalmazó eljárások esetében kulcsfontosságú, hogy a csoportképzés valóban releváns karakterisztikákon alapuljon, ez az eredmény összhangban van *Kleinke és szerzőtársai (2011)* megállapításaival is. Amennyiben a csoportosítás nem megfelelően történik, az az eljárás hatékonyságának romlásához vezethet attól függetlenül, hogy a csoportosítás során minden esetben kisebb, homogénebbnek vélt egységek elemzése érvényesül. Összességében eredményeink rámutatnak arra, hogy a szofisztikáltabb imputációs módszerek minden vizsgált ismérv esetén jelentősen jobban teljesítenek, mint a középértékeket (átlag vagy medián) alapul vevő, egyszerű helyettesítési eljárások, amely konklúzió szintén összhangban van a szakirodalommal *(Máder, 2005)*. Kutatásunk korlátai közé tartozik az adathiány mértékének 10%-os maximálása, valamint az adatstruktúrák közül a teljesen véletlenszerű (MCAR-) eset kizárólagos vizsgálata. Célunk a jövőben mind a hiány mértékének, mind annak

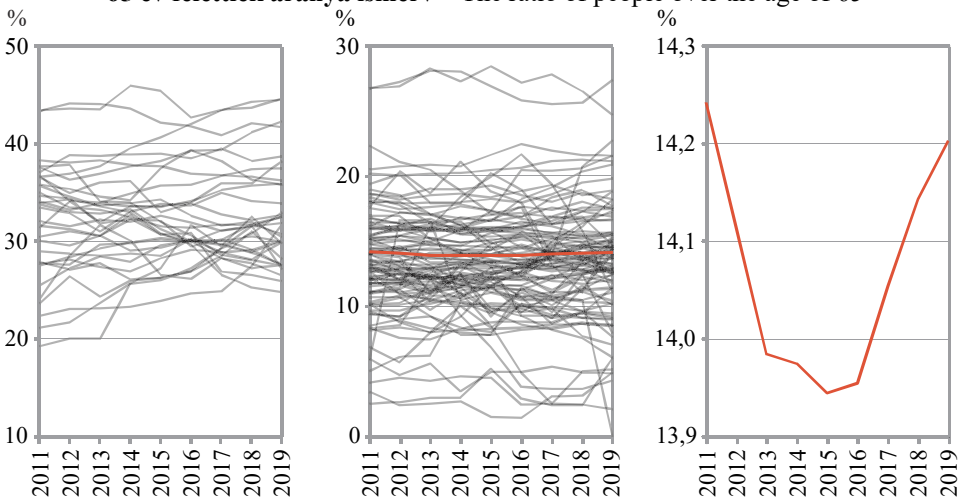
struktúrájának módosítsa után szimulációs elemzéseink ismételt elvégzése, és jövőbeli eredményeink e tanulmányban bemutatottakkal történő szintetizálása után olyan módszerek javaslata, amelyek minden felmerülő adathiány-karakterisztika esetén megoldási alternatívaként szolgálhatnak a kutatóknak. Jelen kutatásunk legfőbb újdonságértékeként pedig azon eredményünket azonosítjuk, miszerint a peremeloszlások figyelembevételével történő imputálási módszerek 5 és 10%-os adathiánymérték mellett, MCAR-adathiány struktúrában minden vizsgált ismérv-karakterisztika (például trend, fluktuáció) megléte mellett jól teljesítettek, így azoknak alkalmazását kiemelten javasoljuk hiányzó adatok becslésére minden olyan esetben, ahol az adatok rendelkezésre állása panelformában is adott.

Függelék

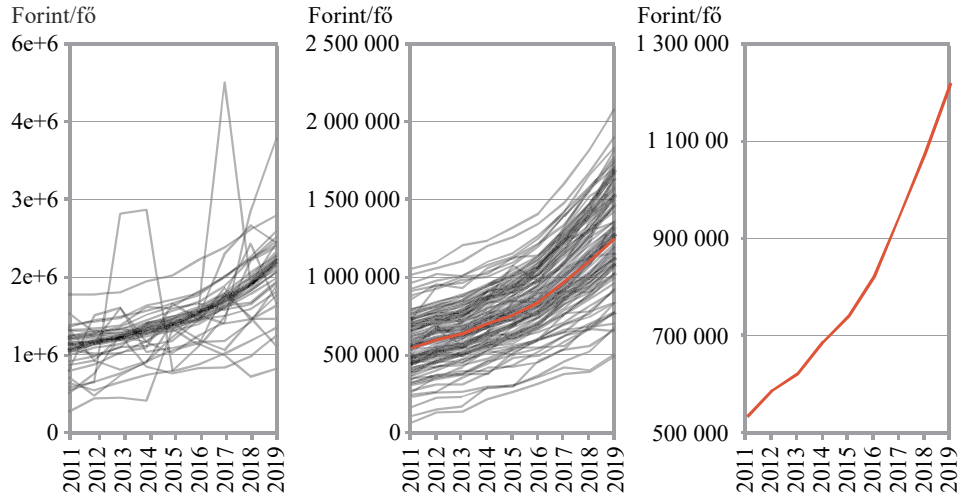
F1. ábra

A további alkalmazott ismérvek idősorainak lefutása
The time series trends of the remaining variables

65 év felettiek aránya ismérv – The ratio of people over the age of 65



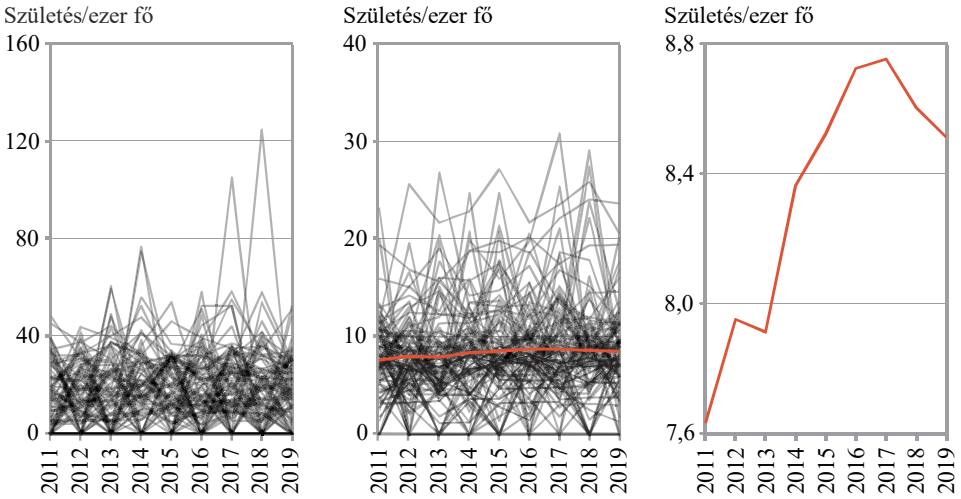
Egy lakosra jutó jövedelem ismérv – Income per inhabitant after tax



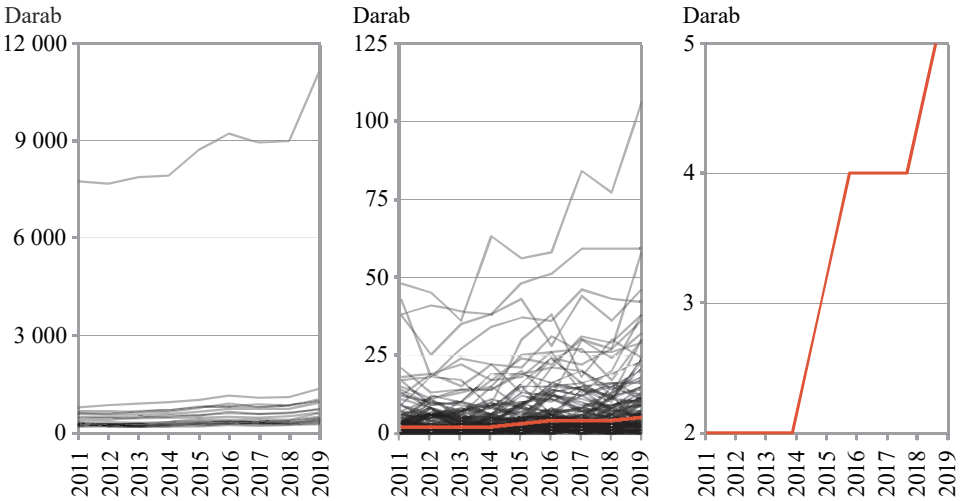
(Az ábrák a következő oldalon folytatódnak)

(folytatás)

Ezer lakosra jutó élveszületések száma – Live births per 1000 inhabitants



Házasságkötések száma – Number of marriages



Forrás: saját szerkesztés.

F1. táblázat

Az ismérvek illeszkedésmutatóinak minimuma, maximuma, átlaga és mediánja az 1000 szimuláció során

The minimum, maximum, mean and median of the goodness of fit measures after 1000 iterations

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
Őstermelők száma ismérv					
Átlagos abszolút hiba	Sorátlag	10,6037	17,0083	13,5776	13,5454
	Oszlopátlag	84,0953	113,5024	97,0494	96,8637
	Teljes átlag	84,5659	114,3669	97,4053	97,1856
	Sormedián	9,0719	15,3608	12,0247	11,9716
	Oszlopmedián	61,47	98,0659	77,9272	77,7304
	Teljes medián	61,4898	98,0677	78,0897	77,8957
	Perem	5,6498	7,9388	6,7704	6,755
	Perem, ismert átlag	5,7009	7,8709	6,7067	6,6894
	Perem vármegyénként	5,9196	11,4245	7,5753	7,5121
	Utolsó megfigyelés	6,2995	10,8097	7,9928	7,8795
	Regressziós becslés	7,9985	12,0435	9,8326	9,8155
Átlagos négyzetes hiba	Sorátlag	641,9888	4063,3886	1512,5801	1367,6311
	Oszlopátlag	17 512,094	109 276,23	43 676,855	41 466,905
	Teljes átlag	17 647,97	109 578,94	43 840,76	41 626,813
	Sormedián	512,9257	4610,1064	1552,789	1335,2078
	Oszlopmedián	19 718,909	115 182,19	47 395,645	45 269,505
	Teljes medián	19 753,22	115 299,91	47 489,61	45 348,732
	Perem	113,1507	543,4402	236,7722	223,6191
	Perem, ismert átlag	111,9826	495,8325	229,189	216,8314
	Perem vármegyénként	133,252	3269,776	597,032	503,0379
	Utolsó megfigyelés	270,487	4211,1128	929,8822	746,2999
	Regressziós becslés	319,3155	2697,7006	738,656	676,247
Hibák szórása	Sorátlag	25,3349	63,7271	38,2446	36,983
	Oszlopátlag	131,758	330,1594	206,5599	203,6817
	Teljes átlag	132,2399	330,6208	206,9493	204,0859
	Sormedián	22,5377	67,4819	38,1873	36,188
	Oszlopmedián	131,9468	330,3876	206,7013	203,8269
	Teljes medián	132,2399	330,6208	206,9493	204,0859
	Perem	10,6408	23,317	15,259	14,9477
	Perem, ismert átlag	10,5853	22,2751	15,0201	14,7273
	Perem vármegyénként	11,5136	56,9907	23,3942	22,3945
	Utolsó megfigyelés	16,1214	64,6055	29,1234	27,0883
	Regressziós becslés	17,8747	51,9295	26,7383	25,9721

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
Őstermelők száma ismérv					
Korreláció	Sorátlag	0,9668	0,9938	0,9834	0,9836
	Oszlopátlag	–0,0270	0,1445	0,0629	0,062
	Teljes átlag				
	Sormedián	0,971	0,994	0,9845	0,9846
	Oszlopmedián	–0,0188	0,1472	0,0651	0,064
	Teljes medián				
	Perem	0,9936	0,9989	0,9972	0,9973
	Perem, ismert átlag	0,9937	0,9989	0,9973	0,9974
	Perem vármegyénként	0,9886	0,9981	0,9955	0,9958
	Utolsó megfigyelés	0,9768	0,9972	0,9914	0,992
	Regressziós becslés	0,9842	0,9964	0,9919	0,9921
Hibák tartománya	Sorátlag	409	2024,1071	988,6135	934,2411
	Oszlopátlag	1181,1306	5334,5234	3631,6927	3800,1902
	Teljes átlag	1178	5332	3611,329	3768
	Sormedián	348	1938,5	933,4655	829,25
	Oszlopmedián	1180	5332	3618,1345	3779
	Teljes medián	1178	5332	3611,329	3768
	Perem	147,8536	786,0665	360,6282	334,7392
	Perem, ismert átlag	131,0885	747,1714	353,7087	324,6818
	Perem vármegyénként	170,2806	1412,3522	633,532	622,7287
	Utolsó megfigyelés	232	2336	798,863	707,5
	Regressziós becslés	283,9709	1876,8834	720,8337	684,8079
Párosított t-próba átlaga	Sorátlag	–3,6423	3,0874	–0,0137	0,0469
	Oszlopátlag	–4,2570	2,8518	–0,1074	–0,0365
	Teljes átlag	–4,3507	2,8081	–0,1054	–0,0330
	Sormedián	2,5925	7,5846	5,0054	5,0132
	Oszlopmedián	8,1452	14,4781	11,1699	11,0848
	Teljes medián	8,1308	14,3829	11,1415	11,0482
	Perem	–2,6877	3,4436	0,3835	0,4487
	Perem, ismert átlag	–3,1053	3,1567	0,0366	0,071
	Perem vármegyénként	0,2403	5,4126	2,6211	2,6307
	Utolsó megfigyelés	–8,2094	–0,9612	–5,0666	–5,1389
	Regressziós becslés	–2,2786	4,9362	1,5174	1,5451

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
65 év felettiek aránya ismérv					
Átlagos abszolút hiba	Sorátlag	0,8079	0,911	0,857	0,8565
	Oszlopátlag	2,7818	3,1487	2,9714	2,9721
	Teljes átlag	2,7822	3,1497	2,9716	2,9719
	Sormedián	0,8337	0,9511	0,8872	0,8865
	Oszlopmedián	2,7418	3,131	2,9489	2,9489
	Teljes medián	2,741	3,1327	2,9496	2,9497
	Perem	0,8009	0,9077	0,8528	0,8519
	Perem, ismert átlag	0,8003	0,9069	0,8522	0,8512
	Perem vármegyénként	0,7915	0,9006	0,8511	0,8508
	Utolsó megfigyelés	0,5699	0,6668	0,6193	0,6188
	Regressziós becslés	0,5755	0,6486	0,6093	0,6093
Átlagos négyzetes hiba	Sorátlag	1,3921	2,0061	1,6555	1,649
	Oszlopátlag	15,1464	20,9499	18,0159	17,9861
	Teljes átlag	15,1605	20,9546	18,0168	17,9875
	Sormedián	1,576	2,2539	1,854	1,8451
	Oszlopmedián	15,1451	21,2966	18,1672	18,1246
	Teljes medián	15,1457	21,3061	18,1671	18,128
	Perem	1,388	2,0066	1,646	1,6389
	Perem, ismert átlag	1,3854	2,0033	1,6443	1,6379
	Perem vármegyénként	1,398	1,9894	1,6302	1,6224
	Utolsó megfigyelés	0,703	1,2561	0,9343	0,9282
	Regressziós becslés	0,7338	1,2629	0,9233	0,9186
Hibák szórása	Sorátlag	1,1801	1,4166	1,2861	1,2843
	Oszlopátlag	3,8863	4,5759	4,2433	4,2414
	Teljes átlag	3,8882	4,5764	4,2434	4,2415
	Sormedián	1,2551	1,5015	1,3605	1,3579
	Oszlopmedián	3,8884	4,5767	4,2431	4,2412
	Teljes medián	3,8882	4,5764	4,2434	4,2415
	Perem	1,1783	1,4168	1,2823	1,2802
	Perem, ismert átlag	1,1772	1,4156	1,2817	1,2796
	Perem vármegyénként	1,1826	1,4103	1,2762	1,2736
	Utolsó megfigyelés	0,8385	1,1207	0,9656	0,9633
	Regressziós becslés	0,8568	1,1239	0,9598	0,9581

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
65 év felettiek aránya ismérv					
Korreláció	Sorátlag	0,9413	0,9629	0,953	0,9531
	Oszlopátlag	–0,0387	0,0713	0,0156	0,0158
	Teljes átlag				
	Sormedián	0,9352	0,958	0,9475	0,9476
	Oszlopmedián	–0,0383	0,0734	0,0163	0,0164
	Teljes medián				
	Perem	0,9414	0,963	0,9533	0,9533
	Perem, ismert átlag	0,9415	0,963	0,9533	0,9534
	Perem vármegyénként	0,9415	0,9625	0,9537	0,9538
	Utolsó megfigyelés	0,9643	0,9804	0,9741	0,9742
	Regressziós becslés	0,9641	0,9803	0,9742	0,9744
Hibák tartománya	Sorátlag	11,7714	29,8829	20,3794	20,1654
	Oszlopátlag	35,94	46,4466	43,9909	44,25
	Teljes átlag	35,94	46,04	43,8562	44,18
	Sormedián	13,875	32,3	21,9042	21,595
	Oszlopmedián	35,94	46,39	43,9774	44,33
	Teljes medián	35,94	46,04	43,8562	44,18
	Perem	11,8227	30,3782	20,5205	20,2574
	Perem, ismert átlag	11,7926	30,3656	20,5148	20,2566
	Perem vármegyénként	12,051	30,0258	20,4727	20,2905
	Utolsó megfigyelés	10,14	28,14	18,8459	18,575
	Regressziós becslés	10,1146	30,6231	19,1058	19,1201
Párosított t-próba átlaga	Sorátlag	–3,4953	3,0815	–0,0639	–0,1440
	Oszlopátlag	–3,2867	3,555	–0,0135	0,0089
	Teljes átlag	–3,2496	3,5869	–0,0131	0,016
	Sormedián	–1,8267	4,6308	1,2874	1,2739
	Oszlopmedián	1,692	7,9626	4,9121	4,9113
	Teljes medián	1,8618	7,973	4,8699	4,8802
	Perem	–3,4022	3,0718	–0,0486	–0,1063
	Perem, ismert átlag	–3,3758	3,0662	–0,0628	–0,1284
	Perem vármegyénként	–3,3974	3,4852	0,1376	0,1041
	Utolsó megfigyelés	–3,6937	3,5782	–0,3256	–0,3363
	Regressziós becslés	–2,4778	5,7526	1,0059	0,9676

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
1000 lakosra jutó élveszületések száma ismérv					
Átlagos abszolút hiba	Sorátlag	2,8642	3,4835	3,1618	3,1624
	Oszlopátlag	3,4177	4,0603	3,6987	3,6993
	Teljes átlag	3,4416	4,0656	3,7138	3,7141
	Sormedián	2,8706	3,5298	3,1769	3,1749
	Oszlopmedián	3,3925	4,0122	3,672	3,6709
	Teljes medián	3,4039	4,0226	3,6873	3,6862
	Perem	2,8614	3,464	3,1381	3,1392
	Perem, ismert átlag	2,861	3,4638	3,1371	3,1383
	Perem vármegyénként	2,8612	3,4686	3,1416	3,1443
	Utolsó megfigyelés	3,5512	4,4662	4,0246	4,0243
	Regressziós becslés	3,029	3,655	3,3307	3,3299
Átlagos négyzetes hiba	Sorátlag	17,0841	38,9249	22,9764	22,2872
	Oszlopátlag	22,3071	43,8965	29,2782	28,6645
	Teljes átlag	22,409	43,9737	29,443	28,7893
	Sormedián	18,3123	42,2772	25,2398	24,3923
	Oszlopmedián	22,1398	44,0573	29,4904	28,8536
	Teljes medián	22,2675	44,1756	29,6706	29,0403
	Perem	16,8381	39,2034	22,705	22,037
	Perem, ismert átlag	16,8279	39,1713	22,6907	22,0198
	Perem vármegyénként	16,9164	39,315	22,7238	22,0164
	Utolsó megfigyelés	27,9805	67,5205	40,6232	39,4912
	Regressziós becslés	19,1938	41,9029	26,3224	25,6571
Hibák szórása	Sorátlag	4,1326	6,24	4,7819	4,7215
	Oszlopátlag	4,7098	6,6278	5,4014	5,3543
	Teljes átlag	4,7194	6,6336	5,4166	5,3661
	Sormedián	4,2736	6,4846	4,9963	4,9274
	Oszlopmedián	4,7058	6,6251	5,4013	5,3549
	Teljes medián	4,7194	6,6336	5,4166	5,3661
	Perem	4,103	6,2622	4,7533	4,694
	Perem, ismert átlag	4,1017	6,2597	4,7518	4,6923
	Perem vármegyénként	4,112	6,2708	4,7553	4,6928
	Utolsó megfigyelés	5,2903	8,22	6,3592	6,2843
	Regressziós becslés	4,3781	6,4729	5,1191	5,0656

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
1000 lakosra jutó élveszületések száma ismérv					
Korreláció	Sorátlag	0,2934	0,6022	0,491	0,4939
	Oszlopátlag	0,0088	0,1544	0,0756	0,0748
	Teljes átlag				
	Sormedián	0,253	0,5796	0,4544	0,4576
	Oszlopmedián	0,0071	0,1496	0,0758	0,075
	Teljes medián				
	Perem	0,2948	0,6107	0,4993	0,5018
	Perem, ismert átlag	0,2955	0,6116	0,4997	0,5023
	Perem vármegyénként	0,2937	0,6098	0,4993	0,5022
	Utolsó megfigyelés	0,1315	0,4676	0,3117	0,3125
	Regressziós becslés	0,2553	0,5664	0,4453	0,4472
Hibák tartománya	Sorátlag	39,9879	145,6388	69,0702	63,6574
	Oszlopátlag	31,5546	125,4194	58,9741	53,0668
	Teljes átlag	31,11	125	58,5212	52,63
	Sormedián	40,7	153,99	73,4405	65,205
	Oszlopmedián	31,495	125,22	58,847	52,9675
	Teljes medián	31,11	125	58,5212	52,63
	Perem	38,4821	147,3516	68,6363	63,3799
	Perem, ismert átlag	38,4378	147,2344	68,6196	63,4168
	Perem vármegyénként	38,1025	146,6643	68,4884	63,0672
	Utolsó megfigyelés	54,19	230,26	102,9915	92,625
	Regressziós becslés	41,6238	169,7	80,2009	72,3105
Párosított t-próba átlaga	Sorátlag	–3,1897	3,0208	–0,0549	0,0002
	Oszlopátlag	–3,6271	2,7431	–0,0566	0,0283
	Teljes átlag	–3,5506	2,8037	–0,0580	–0,0028
	Sormedián	–0,3692	6,3009	2,7582	2,8071
	Oszlopmedián	–0,2169	5,7609	3,2117	3,2227
	Teljes medián	0,0264	5,884	3,306	3,3339
	Perem	–3,3523	3,1519	–0,0512	–0,0064
	Perem, ismert átlag	–3,3524	3,1512	–0,0535	–0,0055
	Perem vármegyénként	–3,3692	3,0085	–0,0189	0,0322
	Utolsó megfigyelés	–2,5877	3,5058	0,4245	0,427
	Regressziós becslés	–3,9494	2,4496	–0,3318	–0,2576

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
Házasságkötések száma ismérv					
Átlagos abszolút hiba	Sorátlag	2,8668	6,6333	3,6488	3,4996
	Oszlopátlag	13,5294	40,2256	18,3753	16,5892
	Teljes átlag	13,645	40,3773	18,5105	16,7138
	Sormedián	2,9341	6,8376	3,8084	3,6445
	Oszlopmedián	7,9359	35,6653	13,1382	11,284
	Teljes medián	8,0451	35,7421	13,2062	11,3605
	Perem	2,0859	8,6133	2,9453	2,4475
	Perem, ismert átlag	2,0923	4,8272	2,6203	2,4365
	Perem vármegyénként	2,2566	20,0805	4,8349	3,3534
	Utolsó megfigyelés	2,642	5,3333	3,173	3,0923
	Regressziós becslés	2,3051	4,2766	2,7991	2,6949
Átlagos négyzetes hiba	Sorátlag	31,5452	6061,0644	462,7095	109,8786
	Oszlopátlag	540,3633	200 005,14	25 622,705	2581,9426
	Teljes átlag	548,45	199 880,72	25 614,941	2582,1697
	Sormedián	32,8584	8104,6755	542,6386	127,5375
	Oszlopmedián	556,1445	200 458,9	25 733,558	2662,6268
	Teljes medián	561,976	200 466,14	25 741,18	2669,4405
	Perem	11,3591	13 169,394	1027,1626	29,3336
	Perem, ismert átlag	11,3921	3283,7581	276,808	29,7021
	Perem vármegyénként	19,2201	83 053,639	9234,9911	306,8739
	Utolsó megfigyelés	19,926	3747,5391	209,7128	50,8883
	Regressziós becslés	15,6146	2183,1928	190,741	38,4358
Hibák szórása	Sorátlag	5,6166	77,8494	16,0991	10,4806
	Oszlopátlag	22,6686	446,752	120,2168	50,7704
	Teljes átlag	22,8349	446,6118	120,2314	50,7946
	Sormedián	5,7287	89,9941	17,8483	11,2806
	Oszlopmedián	22,7342	446,6225	120,1979	50,7712
	Teljes medián	22,8349	446,6118	120,2314	50,7946
	Perem	3,367	114,6903	20,7036	5,4114
	Perem, ismert átlag	3,373	57,2846	12,1242	5,4499
	Perem vármegyénként	4,3661	287,8393	65,2737	17,4904
	Utolsó megfigyelés	4,3942	61,1615	10,1747	7,0689
	Regressziós becslés	3,9457	46,7187	10,1284	6,1987

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
Házasságkötések száma ismérv					
Korreláció	Sorátlag	0,9432	0,9998	0,9857	0,9839
	Oszlopátlag	–0,0515	0,1276	0,031	0,0378
	Teljes átlag				
	Sormedián	0,9413	0,9997	0,9841	0,9816
	Oszlopmedián	–0,0416	0,1296	0,0399	0,0426
	Teljes medián				
	Perem	0,9845	0,9999	0,9949	0,9949
	Perem, ismert átlag	0,9843	0,9999	0,9958	0,9956
	Perem vármegyénként	0,9278	0,9973	0,9746	0,9756
	Utolsó megfigyelés	0,9683	0,9999	0,993	0,9923
	Regressziós becslés	0,9737	0,9999	0,9941	0,9936
Hibák tartománya	Sorátlag	96,125	3461	579,1225	345,3393
	Oszlopátlag	277,1733	11159,383	3727,6749	1034,5327
	Teljes átlag	273	11159	3721,402	1028
	Sormedián	97,5	3461	643,7765	380,5
	Oszlopmedián	274	11159	3722,9155	1030
	Teljes medián	273	11159	3721,402	1028
	Perem	49,6595	3044,564	734,7052	159,4737
	Perem, ismert átlag	50,8285	2315,5164	433,5688	159,2252
	Perem vármegyénként	65,6548	6564,8992	2206,5675	455,5425
	Utolsó megfigyelés	59	2431	310,735	175
	Regressziós becslés	58,4446	2037,9661	343,7912	182,0655
Párosított t-próba átlaga	Sorátlag	–2,9375	2,5093	–0,0821	–0,1187
	Oszlopátlag	–8,8442	1,9926	–1,5071	–1,4936
	Teljes átlag	–8,8750	1,9946	–1,5041	–1,4684
	Sormedián	–1,5815	4,7282	1,385	1,4836
	Oszlopmedián	1,662	10,7582	5,558	6,6384
	Teljes medián	1,6878	10,8924	5,6412	6,7286
	Perem	–2,7136	3,3263	0,8707	0,9877
	Perem, ismert átlag	–2,9595	3,3356	0,283	0,33
	Perem vármegyénként	0,7095	4,1388	2,0909	2,2244
	Utolsó megfigyelés	1,2697	7,7064	4,6485	4,8376
	Regressziós becslés	–3,2483	3,7331	0,4724	0,6091

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
Egy lakosra jutó jövedelem ismérv					
Átlagos abszolút hiba	Sorátlag	2,011E + 05	2,295E + 05	2,150E + 05	2,152E + 05
	Oszlopátlag	1,895E + 05	2,138E + 05	2,019E + 05	2,019E + 05
	Teljes átlag	2,526E + 05	2,854E + 05	2,681E + 05	2,681E + 05
	Sormedián	2,028E + 05	2,337E + 05	2,182E + 05	2,183E + 05
	Oszlopmedián	1,886E + 05	2,137E + 05	2,015E + 05	2,015E + 05
	Teljes medián	2,475E + 05	2,823E + 05	2,646E + 05	2,645E + 05
	Perem	3,870E + 04	4,699E + 04	4,249E + 04	4,244E + 04
	Perem, ismert átlag	3,869E + 04	4,696E + 04	4,248E + 04	4,242E + 04
	Perem vármegyénként	3,699E + 04	4,501E + 04	4,061E + 04	4,058E + 04
	Utolsó megfigyelés	8,259E + 04	9,373E + 04	8,818E + 04	8,813E + 04
	Regressziós becslés	6,069E + 04	7,131E + 04	6,506E + 04	6,500E + 04
Átlagos négyzetes hiba	Sorátlag	5,943E + 10	8,137E + 10	6,934E + 10	6,922E + 10
	Oszlopátlag	5,865E + 10	8,100E + 10	6,753E + 10	6,715E + 10
	Teljes átlag	1,019E + 11	1,357E + 11	1,161E + 11	1,159E + 11
	Sormedián	6,197E + 10	8,667E + 10	7,409E + 10	7,387E + 10
	Oszlopmedián	5,872E + 10	8,145E + 10	6,784E + 10	6,741E + 10
	Teljes medián	1,019E + 11	1,403E + 11	1,190E + 11	1,188E + 11
	Perem	3,154E + 09	1,160E + 10	4,945E + 09	4,530E + 09
	Perem, ismert átlag	3,156E + 09	1,159E + 10	4,943E + 09	4,527E + 09
	Perem vármegyénként	2,838E + 09	1,175E + 10	4,637E + 09	4,211E + 09
	Utolsó megfigyelés	9,937E + 09	2,144E + 10	1,252E + 10	1,179E + 10
	Regressziós becslés	6,293E + 09	1,489E + 10	8,121E + 09	7,709E + 09
Hibák szórása	Sorátlag	2,432E + 05	2,852E + 05	2,633E + 05	2,630E + 05
	Oszlopátlag	2,422E + 05	2,847E + 05	2,598E + 05	2,591E + 05
	Teljes átlag	3,182E + 05	3,681E + 05	3,406E + 05	3,405E + 05
	Sormedián	2,468E + 05	2,879E + 05	2,673E + 05	2,670E + 05
	Oszlopmedián	2,422E + 05	2,847E + 05	2,598E + 05	2,591E + 05
	Teljes medián	3,182E + 05	3,681E + 05	3,406E + 05	3,405E + 05
	Perem	5,617E + 04	1,077E + 05	6,975E + 04	6,731E + 04
	Perem, ismert átlag	5,619E + 04	1,076E + 05	6,974E + 04	6,727E + 04
	Perem vármegyénként	5,329E + 04	1,084E + 05	6,745E + 04	6,491E + 04
	Utolsó megfigyelés	7,193E + 04	1,240E + 05	8,453E + 04	8,045E + 04
	Regressziós becslés	7,926E + 04	1,219E + 05	8,936E + 04	8,732E + 04

(A táblázat a következő oldalon folytatódik)

(folytatás)

Mutató	Eljárás	Minimum	Maximum	Átlag	Medián
Egy lakosra jutó jövedelem ismérv					
Korreláció	Sorátlag	0,5973	0,6784	0,6414	0,642
	Oszlopátlag	0,5846	0,6923	0,6467	0,6469
	Teljes átlag				
	Sormedián	0,5799	0,6649	0,624	0,6248
	Oszlopmedián	0,5848	0,6927	0,6467	0,6468
	Teljes medián				
	Perem	0,9524	0,987	0,979	0,9806
	Perem, ismert átlag	0,9524	0,987	0,979	0,9806
	Perem vármegyénként	0,9525	0,9879	0,9801	0,9818
	Utolsó megfigyelés	0,9368	0,9838	0,9747	0,9783
	Regressziós becslés	0,9387	0,9722	0,9649	0,9664
Hibák tartománya	Sorátlag	1,260E + 06	3,838E + 06	1,907E + 06	1,714E + 06
	Oszlopátlag	1,461E + 06	4,580E + 06	2,353E + 06	2,158E + 06
	Teljes átlag	1,761E + 06	4,428E + 06	2,482E + 06	2,334E + 06
	Sormedián	1,253E + 06	3,902E + 06	1,995E + 06	1,836E + 06
	Oszlopmedián	1,461E + 06	4,587E + 06	2,354E + 06	2,158E + 06
	Teljes medián	1,761E + 06	4,428E + 06	2,482E + 06	2,334E + 06
	Perem	5,388E + 05	3,425E + 06	1,434E + 06	1,271E + 06
	Perem, ismert átlag	5,360E + 05	3,420E + 06	1,433E + 06	1,271E + 06
	Perem vármegyénként	5,311E + 05	3,474E + 06	1,407E + 06	1,233E + 06
	Utolsó megfigyelés	5,100E + 05	4,908E + 06	1,375E + 06	1,033E + 06
	Regressziós becslés	5,255E + 05	4,023E + 06	1,337E + 06	1,153E + 06
Párosított t-próba átlaga	Sorátlag	–3,0089	3,3685	0,0062	0,0203
	Oszlopátlag	–3,0819	3,4222	0,0285	0,0494
	Teljes átlag	–3,3220	3,432	0,0142	0,0135
	Sormedián	4,6505	10,6098	7,2226	7,2006
	Oszlopmedián	–0,4721	5,8482	2,509	2,5173
	Teljes medián	2,9742	9,1162	5,9869	5,9992
	Perem	–3,7049	3,274	–0,0039	0,036
	Perem, ismert átlag	–3,7753	3,2183	–0,0168	0,0231
	Perem vármegyénként	–3,5637	3,6728	0,141	0,1711
	Utolsó megfigyelés	21,4395	38,7543	32,7592	33,8603
	Regressziós becslés	0,1923	7,0153	3,805	3,8155

Forrás: saját szerkesztés.

Irodalom

Allison, P. D. (2009): Missing data. In: Millsap, R. E. – Maydeu-Olivares, A. (eds.): *The SAGE handbook of quantitative methods in psychology.* SAGE, London, pp. 72–89.

Alwateer, M. – Atlam, E. S. – Abd El-Raouf, M. M. – Ghoneim, O. A. – Gad, I. (2024): Missing data imputation: A comprehensive review. *Journal of Computer and Communications,* 12(11), 53–75. https://doi.org/10.4236/jcc.2024.1211004

Anselin, L. (1988): *Spatial econometrics: Methods and models.* Springer, Dordrecht.

Baltagi, B. H. (2021): *Econometric analysis of panel data.* Springer, Cham. https://doi.org/10.1007/978-3-030-53953-5

Békés, G. – Kézdi, G. (2021): *Data analysis for business, economics, and policy.* Cambridge University Press, Cambridge.

Bennett, D. A. (2001): How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health,* 25(5), 464–469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

van Buuren, S. (2012): *Flexible imputation of missing data,* 10: b1182, CRC Press, Boca Raton.

Dong, Y. – Peng, C. Y. J. (2013): Principled missing data methods for researchers. *SpringerPlus,* 2(1), 1–17. https://doi.org/10.1186/2193-1801-2-222

Enders, C. K. (2010): *Applied missing data analysis.* Guilford Press, New York

Gondara, L. – Wang, K. (2018): MIDA: Multiple imputation using denoising autoencoders. In: Phung, D. – Tseng, V. – Webb, G. – Ho, B. – Ganji, M. – Rashidi, L. (eds.): *Advances in knowledge discovery and data mining.* Springer, Cham, pp. 260–272. https://doi.org/10.1007/978-3-319-93040-4_21

Hair, J. F. – Black, W. C. – Babin, B. J. – Anderson, R. E. (2010): *Multivariate data analysis.* Pearson, Upper Saddle River.

Jakobi Á. (2009): Felületmodellek és lejtők a társadalomföldrajzban, avagy térbeli interpoláció társadalomföldrajzi adatokon. In: Hegedűs A. (szerk): *Geoinformatika és domborzatmodellezés.* A HunDEM 2009 és a GeoInfo 2009 konferencia és kerekasztal válogatott tanulmányai.

Kang, H. (2013): The prevention and handling of the missing data. *Korean Journal of Anesthesiology,* 64(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kleinke, K. – Stemmler, M. – Reinecke, J. – Lösel, F. (2011): Efficient ways to impute incomplete panel data. *AStA Advances in Statistical Analysis,* 95, 351–373. https://doi.org/10.1007/s10182-011-0179-9

Laine, M. (2019): Introduction to dynamic linear models for time series analysis. In: Montillet, J-P. – Bos, M. S. (eds.): *Geodetic time series analysis in earth sciences.* Springer, Cham, pp. 139–156.

Little, R. J. A. – Rubin, D. B. (2002): *Statistical analysis with missing data.* Wiley, Hoboken.

Máder M. P. (2005): Az imputálási eljárások hatékonysága. *Statisztikai Szemle,* 83(7), 628–643.

Malhotra, N. K. (1987): Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research,* 24(1), 74–84. https://doi.org/10.2307/3151755

Oravecz B. (2008): Hiányzó adatok és kezelésük a statisztikai elemzésekben. *Statisztikai Szemle,* 86(4), 365–384.

Rubin, D. B. (1976): Inference and missing data. *Biometrika,* 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sajtos L. – Mitev A. (2007): *SPSS kutatási és adatelemzési kézikönyv.* Alinea Kiadó, Budapest.

Schafer, J. L. (1997): *Analysis of incomplete multivariate data.* Chapman & Hall/CRC, Boca Raton.

Schafer, J. L. (1999): Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3–15. https://doi.org/10.1177/096228029900800102

Schafer, J. L. – Graham, J. W. (2002): Missing data: Our view of the state of the art. *Psychological Methods,* 7(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147