Kmetty Zoltán[1]

Temporális és regionális összehasonlítások lehetséges torzításai

Hogyan kezeljük a nem-válaszolást?

DOI: 10.18030/socio.hu.2018.2.95

Absztrakt

A tanulmány arra keresi a választ, hogyan érdemes kezelni az adathiányt abban az esetben, ha időben és/vagy térben elnyúló adataink vannak. Amellett érvelünk, hogy időben és térben kiterjedt adatok esetében több olyan egyedi szempont is felmerülhet, ami eltérő válaszadói struktúrát eredményezhet az adatfelvételek között. Az empirikus elemzésben a *European Social Survey (ESS)* 7 hullámának és 14 országának adatait felhasználva vizsgáltuk meg (N=185 049), hogy egy rendszerellenességet mérő index esetében mekkora volt az adathiány országonként és hullámonként. Az ESS adatok nagy szórást mutattak abban, hogy a vizsgált rendszerellenesség indexet alkotó változók között mekkora volt az adathiány: általában a nyugati és a skandináv országokban magasabb volt az érvényes válaszok aránya, míg keleten és a mediterrán országokban alacsonyabb. A többszintű regressziós modellek azt mutatták, hogy a nem-válaszolás közötti különbség nagyjából 3 százalékáért felel az ország és az adatfelvétel hulláma, míg egyéni szinten a rosszabb szociokulturális környezet magasabb nem-válaszolással jár együtt, akárcsak az, ha a válaszadó nő vagy idősebb. A nem-válaszolás kezelésének több forgatókönyvét is teszteltük, és bár összességében hasonló képet mutattak, de a pótolt adatokat tartalmazó index magasabb rendszerellenességet mutatott, mint a nem-válaszolók kihagyását követő stratégia.

Az adathiány természetes velejárója az elemzéseinknek, és ha tudatosan nem foglalkozunk vele, akkor is kezeljük valahogy. Ez a kezelés sok esetben a használt statisztikai programok működési módjából következik. Ha nem akarjuk, hogy a programok vezessenek minket, alakítsunk ki egyértelmű protokollt a nem-válaszolás kezelésére, mert ez lehet a záloga annak, hogy az eredményeink érvényesek és megbízhatóak legyenek.

Kulcsszavak: nem-válaszolás, adatpótlás, ESS, többszintű elemzés, rendszerellenesség

1 ELTE TáTK Szociológia Intézet, MTA–ELTE Peripato Kutatócsoport.

Possible biases of temporal and regional comparisons

How to handle item non-response?

Abstract

The main question of this paper is how to handle missing data in spatial and temporal analysis. We argue here, that in the case of spatial and temporal data, there may be several unique aspects that can lead to different response structures between datasets. In the empirical analysis, we have investigated the data of seven waves and 14 countries of ESS (N = 185 049) regarding the volume on item non-response per wave and country in the case of anti-regime attitudes. The ESS data has showed great deviation in the volume of missing answers of anti-regime attitudes. In the case of Western and Scandinavian countries higher rates, and in the case of Eastern and Mediterranean countries lower rates of valid answers have been measured. Based on the multi-level analysis countries and waves have been responsible for around 3 percent of the difference in non-response. At the respondent level, a lower response rate was more typical in the case of low social status, female respondents and older people. We have treated non-responses with different methods such us complete case analyis or nearest neighbour imputation. The difference was not extremely high between the methods, but overall the imputed anti-regime index was higher than the basic version.

Missing data is an everyday component of our analysis. The non-treatment of non-response is also some kind of treatment. This non-intentional treatment often comes from the default setting of the statistical programmes used. If we do not want to be led by software we have to develop clear protocols for how to handle missing data. This could be an important foundation of reliable and valid data analysis.

Keywords: non-response, imputation, ESS, multilevel analysis, anti-regime attitude

Kmetty Zoltán

Temporális és regionális összehasonlítások lehetséges torzításai

Hogyan kezeljük a nem-válaszolást?

1. Bevezetés

A legtöbb empirikus adatelemzés *egy* adatfelvételre épít, és abból igyekszik a kutatási kérdéseire válaszolni. Az utóbbi évtizedekben egyre bővülő nemzetközi összehasonlító projektek (például ESS, ISSP, EVS) azonban kibővítették az időbeli és térbeli összehasonlító kutatások lehetőségét. Mind az időbeli, mind a nemzetközi összehasonlításoknál folyamatos dilemmát jelent, hogy mikor lehet érvényesen és megbízhatóan trendeket felrajzolni. Az azonos kérdőívszerkezet evidenciának számít – főleg a nemzetközi kutatások esetében. Ez egy több országot érintő kutatás egyetlen hullámának esetében még viszonylag egyszerűen betartható, de több adatfelvételi hullámon át már nehezebben érvényesíthető, ezért a kérdőíves kontextushatás kiszűrése nagyon tervszerű kérdőív-kialakítást igényel (Sigelmann 1981, Tourangeau et al. 1989). A mintavételi mód és a mintavételi technika azonossága is elvárás (Lynn 1998, Bowling 2005, Jackle et al. 2010, Lugtig 2011), akárcsak az eredeti tartalomhoz legközelebb álló fordítás biztosítása. Ezen felül természetesen vannak nem kontrollálható hatások is, mint az adott adatfelvételt körülvevő kontextusok különbsége, változékonysága.[2]

A szigorú elvárásrendszer ellenére sem jelenthető azonban ki, hogy az adatfelvételek egyforma minőségben zajlanak akár térben, akár időben összehasonlítva. Ezt bizonyítja egy friss ESS elemzés (Beullens–Loosveldt 2016), ami rámutat arra, hogy a különböző országok között viszonylag jelentős különbségek vannak abban, hogy mekkora a kérdezőbiztos hatása az eredményekre. Amellett, hogy a kérdezőbiztos akaratán kívül tudja befolyásolni a kérdezés menetét, abban is kiemelten fontos szerepe van, hogy milyen arányban tudnak főcímen elérni kérdezettet *(unit non-response)*, illetve a kérdőíven belül az egyes kérdéseknél mekkora a tétel nem válaszolás *(item non-response)* aránya. A nem-válaszolás egy kutatási kérdés érvényes vizsgálatát önmagában is el tudja rontani (ha nem teljesen véletlenszerű nem-válaszolásról van szó), de időbeli és térbeli összehasonlításkor ezek a torzulások talán még nagyobb gondot okozhatnak, főleg, ha nem azonos mértékű és irányú a nem-válaszolás az egyes időszakokban és térbeli egységekben. A következő elemzésben ez utóbbi problémát állítjuk fókuszba ESS adatok elemzésén keresztül.

A tanulmány első felében röviden felvázoljuk a nem-válaszolás típusait, majd a tétel nem-válaszolásra koncentrálva végigvesszük, hogy milyen okok járulhatnak hozzá az adathiány kialakulásához. Az elméleti részben röviden kitérünk arra is, hogy milyen módszerek terjedtek el az adathiány kezelésére a társadalomtudományi kutatásokban. Az elemzés részben a *European Social Survey* (ESS) első hét hullámát felhasználva egy index példáján bemutatjuk, hogy milyen különbségek adódhatnak a nem-válaszolás kezelésének eltérő stratégiáiból.

2 Ilyenre példa a 2008-as gazdasági válság, ami az ESS 4. hullámában résztvevő országokat eltérő időben érte el. Az elhúzódó adatfelvételi időszak ráadásul további problémát jelenthet az adatok kiértékelésénél.

2. Elméleti összefoglaló

2.1. A nem-válaszolás két nagy típusa

Az adathiány nem új keletű probléma a társadalomtudományi kutatásokban, ennek ellenére időről időre felmerül, hogyan érdemes kezelni ezt a problémát. Ennek legfőbb oka az, hogy nem létezik általánosan bevett jó eljárás az adathiány kezelésére, illetve a kezelési módszerek gyakran a misztikum homályába vesznek látszólagos statisztikai bonyolultságuk okán.

A nem-válaszolást két nagy csoportra bonthatjuk. Egyrészt adathiányként tekinthetünk arra, ha a mintába beválasztott személy nem kerül végül lekérdezésre a kutatásban. Ennek több oka lehet: következhet elutasításból, abból, hogy nem találták meg a terepmunka során a személyt (pl. nem volt otthon, elköltözött), vagy akár abból is, hogy valami oknál fogva (pl. pszichológia problémák) nem képes válaszolni a kutatási kérdésekre az alany. Ezt nevezi az irodalom eset nem-válaszolásnak *(unit non-response)*. Az ismert eloszlásokra különböző súlyozási eljárásokkal be tudjuk állítani a mintánkat, így ezt a hibát részben tudjuk kezelni. Természetesen sokszor nyitott kérdés marad, hogy a populációs szinten nem ismert változók esetében ez a kalibrálás/súlyozás képes-e kijavítani a minta torzulásait, ezt sok esetben nagyon nehéz megbecsülni, eldönteni. A pontatlan politikai közvélemény-kutatások esetében gyakran hivatkoznak a torzuló mintaösszetételre, a 2015-ös kudarcos brit választási előrejelzéseknél is előkerült ez a magyarázat (Sturgis et al 2016). A súlyozási és kalibrálási módszerek a torzulás bonyolultságától függően igen komplexek is lehetnek, de általában a minták korrekcióját elvégzik az adatfelvétel módszertani részéért felelős statisztikusok, az empirikus elemzést elvégző kutatóknak „csak" arra kell koncentrálni, hogy a súlyozás be legyen kapcsolva, amikor az elemzést végzik. Ez a nem válaszolási probléma tehát nem igazán vetődik fel akkor, amikor empirikus elemzést készítünk (megoldottnak tekintjük).

A másik, a tanulmány szempontjából minket jobban érdeklő nem-válaszolás, a kérdések esetében keletkezett adathiány, azaz a kérdés (tétel) nem válaszolás *(item non-response)*. A kérdés nem-válaszolásnak több oka is lehet, például a kérdezőbiztos figyelmetlensége vagy a kérdezett oldaláról érkező megtagadás, de akár az is, hogy a kérdezett nem tud az adott kérdésre válaszolni.

2.2. Az adathiányt befolyásoló tényezők

Bár az eset szintű nem-válaszolás időbeli/térbeli változása szintén érdekes és releváns kérdés, ezzel most nem foglalkozom ebben a tanulmányban (ilyen jellegű tanulmány kapcsán lásd például György 2004, az ESS esetében pedig Blom 2008, Stoop et al. 2010), helyette a kérdés (tétel) szintű nem válaszolásra fókuszálok.

Az adathiány nagyságát négy faktor befolyásolja: a válaszadó, a kérdezőbiztos, a kérdőív és az adatfelvétel módja (De Leeuw 2001). A válaszadók különböznek egymástól, vannak, akik könnyebben tudnak válaszolni, vannak, akiknél lassabb, nehezebb az a kognitív folyamat, amely során a válasz „elkészül". Át kell gondolni a kérdést, ki kell találni a választ, és azt még elő is kell adni a kérdezőbiztosnak (Koch–Blohm 2009). Idősebb, alacsonyabb végzettségű válaszolóknál jellemzően magasabb az adathiány, de politikai kérdéseknél például ezeken felül az alacsonyabb politikai érdeklődés (és tudás) is nagyobb nem válaszolással járhat együtt (Pickery–Loosveldt 2001).

A kérdezőbiztos a második faktor, ami jelentősen befolyásolhatja a nem-válaszolás nagyságát. Az adatfelvételt végző cégek eltérő módon tréningezhetik a kérdezőbiztosaikat, van, ahol rámenősebb adatgyűjtésre kérik a kérdezőbiztosokat, amit akár inszentívákkal, ösztönzőkkel is erősíthetnek. Túl sok meg nem válaszolt kérdés akár bérlevonáshoz is vezethet, ami arra sarkallhatja a kérdezőket, hogy mindenképp próbáljanak választ kapni minden kérdésre. Persze a kérdezőbiztosok maguk is különböznek, eltérő korúak, neműek, eltérő habitussal és tapasztalattal rendelkeznek. Ezek a tényezők szintén fontosak lehetnek annak kapcsán, hogy mekkora lesz a nem-válaszolás egy kérdőíven belül (Koch–Blohm 2009).

A nem-válaszolás nagysága természetesen nagyban függ magától a kérdőívtől is. A kérdések bonyolultsága, a vizsgált téma kényessége, valamint a témához kötődő elvárt tudás mind nagyon fontos tényező, akárcsak a kérdés típusa vagy a kérdőív hossza (elfáradás). Az sem mindegy, hogy a „nem tudja", „nem válaszol" opciókat rutinból felajánlják a válaszadóknak vagy sem, mivel előbbi stratégia könnyen vezethet magasabb adathiányhoz (Koch–Blohm 2009).

A negyedik fontos faktor pedig a kérdezés módja. A kérdezőbiztossal történő interjú általánosságban magasabb válaszoláshoz vezet, mint az önkitöltés (bár bizonyos nagyon kényes kérdésekben az önkitöltős kérdőívek tudnak jobban működni). A számítógéppel támogatott lekérdezések pedig több érvényes válaszhoz vezetnek, mint a papíros kutatások, mivel a jól programozott kérdőívek segítik a kérdezőbiztost, hogy minden releváns kérdést feltegyen a kérdezettnek, és figyelmeztetnek akkor is, ha valami véletlenül kimaradna (Koch–Blohm 2009, Kmetty 2012).

2.3. Az adathiány típusai

Az adathiány nagyságát meghatározó okok után röviden áttérünk az adathiány típusaira. Az adathiánynak három típusát különbözteti meg a szakirodalom (Little–Rubin 1987).

Teljesen véletlenszerű adathiány *(missing completely at random,* MCAR): nevéből is következik, hogy az adatkiesés teljesen véletlenszerű, a válaszolók és nem válaszolók semmilyen szempontból nem különböznek egymástól. Ez a legkevésbé problematikus eset, a válaszolók véleménye torzítatlanul adja vissza az adott változó eloszlását, tehát nyugodtan figyelmen kívül hagyhatjuk a nem válaszolókat, mindössze a kisebb esetszám okozhat problémát, mivel növeli a standard hiba nagyságát.

Véletlenszerű adathiány *(missing at random,* MAR): az adathiány teljes egészében magyarázható külső változók segítségével, de az adathiány nem függ az adott célváltozótól, amit vizsgálunk. Ebben az esetben már a bent maradó esetekre koncentráló elemzések torzítottak lehetnek, valamilyen módon pótolni kell az adatokat.

Nem véletlenszerű adathiány *(missing not at random,* MNAR): az adathiány nemcsak külső változóktól függ, hanem a vizsgált változótól is. Az adathiány szerkezetének pontos feltárása segíthet ebben az esetben, de ilyen jellegű adathiánynál a torzítás kiküszöbölése nem biztos, hogy megoldható.

Egy politikai példán mutatjuk be a három típust, a pártpreferenciát állítva a fókuszunkba. Ha a pártpreferencia kérdésre *teljesen véletlenszerűen nem válaszolnak* az emberek, akkor a kapott eredmények jól tükrözik a sokasági eloszlást, tehát torzítatlan becslést tudunk adni a pártok arányára (természetesen feltételezve a

megfelelő mintavételt). Ha a nem válaszolás függ egyes külső változóktól, de csak azoktól, akkor *véletlenszerű nem-válaszolással* van dolgunk. Ilyen lehet, ha például a fiatalabbak kevésbé mondják meg a pártpreferenciájukat. *Nem véletlenszerű* a válaszhiány, ha valamely párt szimpatizánsai nagyobb arányban nem válaszolnak a pártpreferencia kérdésre, mint egy másik párt szimpatizánsai, és ez a hatás semmilyen külső változóval nem modellezhető. Utóbbi esetben valószínűleg minden igyekezetünk ellenére csak torzított becslést tudunk adni a pártpreferencia eredményekre.

A következő részben áttérünk a nem-válaszolás problémájára időbeli és térbeli adatok esetében, elsősorban az ESS adataira fókuszálva.

2.4. Nem-válaszolás longitudinális és térben szétterülő adatok esetében

A nem-válaszolás „egyszerű" minták esetében is okozhat olyan torzítást, amit nehéz kiküszöbölni, de országok közötti, illetve időbeli összehasonlítás esetében a probléma még komplexebb. A komplexitás nem abból adódik, hogy bonyolultabb a nem-válaszolási struktúra, hanem abból, hogy az egyes hullámok és térbeli egységek közötti eltérést több mechanizmus is generálhatja. Az ilyen adatokat használó elemzések, ha eltérő nem-válaszolási minták vannak mögöttük, könnyen félrevezetőek lehetnek. A nem-válaszolás lehetséges okaira kitértünk már az előző fejezetben, itt most azt vesszük végig, hogy az országok közötti összehasonlításban ezek az okok hol és miként tudnak felmerülni.

A 2.2 fejezetben négy olyan tényezőt határoztunk meg, amelyek jelentősen befolyásolhatják a nem-válaszolás nagyságát: a válaszadó, a kérdezőbiztos, a kérdőív és az adatfelvétel módja. Röviden végigvesszük, hogy ezek mennyire jelentkezhettek az ESS (vagy más nemzetközi adatfelvétel) kapcsán.

Kérdésekre (jól) válaszolni egy soklépcsős kognitív folyamat, vannak, akik jobban meg tudják ezt oldani, vannak, akiknek ez nehézséget okoz. Alacsonyabb végzettségű, idősebb emberek gyakran nehezebben tudnak válaszolni, ami az ő esetükben megnövelheti a nem-válaszolási arányt. Országonként eltérő nem-válaszolási arányok következhetnek az országok közötti demográfiai különbségekből: az alacsonyabb végzettségű válaszolók magasabb aránya már önmagában jelentős különbségeket okozhat a nem-válaszolási arányban. Ha viszont időben hasonlítunk össze kutatásokat, akkor az idő előrehaladtával megvalósuló oktatási expanzió (és az alacsony végzettségű idős válaszadók kiesése a mintából) vezethet csökkenő nem-válaszoláshoz (Koch–Blohm 2009).

Szintén okozhat különbséget országok között, ha a kérdezőbiztosokat más instrukciókkal küldik ki a nem-válaszolás kezelése kapcsán, vagy eltérő a kérdezőbiztosok szocio-demográfia profilja, tapasztalata az egyes országokban. Az ESS kapcsán igen nagy különbségek vannak az országok között abban, hogy a kérdezőbiztosoknak mekkora hatása van a lekérdezés kimenetére (Beullens–Loosveldt 2016).

Az ESS kérdőív teljesen standardizált, tehát elviekben azt várhatnánk, hogy a kérdőív nem okozhat különbséget az egyes országok között. Azonban országonként eltérő lehet, hogy mi számít kényes kérdésnek, és az is eltérő lehet, hogy milyen kérdéseket tudnak egyszerűbben vagy nehezebben megválaszolni az emberek, hiszen eltérő tapasztalati környezet veszi körbe az egyes országok válaszadóit. Ez egy nehezen kontrolálható hatás, de ezzel is számolni kell, amikor a nem-válaszolás nagyságát próbáljuk modellezni.

Az utolsó lehetséges különbségként az adatfelvétel módját említettük. Az ESS esetében elvárás a személyes kérdezés, de egyes országokban CAPI (laptopos kérdezés), más országokban PAPI módon (papír alapon) kérdeznek. Az utóbbi, papíros kérdezés vezethet magasabb nem-válaszoláshoz (Koch–Blohm 2009).

Koch és Blohm (2009) az ESS első három hullámán vizsgálta meg a nem-válaszolás alakulását. Azt a 17 országot vizsgálták, amely mind a három hullámban részt vett, és a mag modul[3] 75 kérdését vettél alapul. A válaszadók 54–58 százalékánál nem volt adathiány, ami önmagában elég alacsony szám, de a tendencia összességében javuló volt. A skála egyik végén (magas nem-válaszolás) Portugália, Magyarország és Szlovénia állt, míg a legjobban teljesítő országok listáját Norvégia vezette, nem sokkal megelőzve Belgiumot és Finnországot (Koch–Blohm 2009). A többváltozós elemzések azt mutatták, hogy a nem-válaszolás varianciájának 8,7 százaléka keletkezett az országok közötti szinten, és 15,7 százaléka a kérdezőbiztosi szinten. A variancia legnagyobb részéért (76 százalék) az egyéni szint felelt. A nők, az alacsonyabb végzettségűek, az idősebbek, és a politika iránt kevésbé érdeklődők nagyobb arányban nem válaszoltak a kérdésekre. Az eredmények azt is alátámasztották, hogy az országok közötti szocio-demográfia különbségek is szerepet játszanak az eltérő nagyságú adathiány kialakulásában. A számítógéppel támogatott kérdezés pedig felére csökkentette az adathiány nagyságát a papíros lekérdezéshez képest (Koch–Blohm 2009).

2.5. Kezelési lehetőségek

A nem-válaszolás kezelésének hatalmas irodalma van, egy ilyen jellegű és fókuszú tanulmányban mindössze annyira vállalkozhatunk, hogy a főbb irányokat és dilemmákat felrajzoljuk. Ezért ez a rész szándékosan elnagyolt lesz, és elsősorban azoknak szól, akik nem sokat foglalkoztak eddig ezzel a kérdéssel.

A kutatók többsége abból a meglehetősen naív premisszából indul ki, hogy ha nem nyúl az adatokhoz, akkor követi el a legkisebb hibát, ezért az elemzések során gyakran a komplett kérdőívekre koncentrál, és egyszerűen kihagyja a nem válaszolókat. Fontos, hogy a kutatókban tudatosuljon, hogy ez a döntésük is mond valamit arról, hogy mit gondolnak az adathiányról. Ha csak az adathiány nélküli esetekre koncentrálnak, akkor abból a ki nem mondott feltevésből indulnak ki, hogy az adathiány teljesen véletlenszerű. A probléma nem figyelembevétele nem tünteti el a problémát, csak azt okozza, hogy a döntésünket nem kommunikáljuk tisztán. Ha csak az elérhető esetekre hagyatkozunk, akkor is kétfajta megoldás közül választhatunk. Vagy a teljesen komplett eseteket vesszük figyelembe, vagy páronként vizsgáljuk az adathiányt. Utóbbinak lehet értelme bizonyos statisztikai számolásoknál. Például, ha szeretnénk bemutatni három változó egymással való korrelációját, dönthetünk úgy, hogy csak azokat vesszük be az elemzésbe, akik mind a három kérdésre válaszoltak, vagy páronként számoljuk ki a korrelációkat, és ha az adott kérdezett *A* kérdésre nem válaszolt, de *B*-re és *C*-re igen, akkor *B* és *C* korrelációjának kiszámolásakor figyelembe vesszük az esetet.

Ha azt feltételezzük, hogy az adathiány nem teljesen véletlenszerű, akkor az elérhető változókkal való számolás torzított statisztikai becslésekhez fog vezetni. Ilyen esetben az a célszerű stratégia, ha pótoljuk a nem-válaszolást.

3 Azok a kérdések tartoznak a mag modulba, amelyek az összes hullámban azonosan szerepelnek

A pótlások talán legegyszerűbb megoldása, ha a változó átlagát (esetleg mediánját vagy nominális mérési szintnél móduszát) pótoljuk be a hiányzó esetek helyére (Peng et al. 2006). Ez a módszer azonban nem kezeli az adatok torzítását, ha nem véletlenszerű az adathiány, ráadásul a változó szórását is jelentősen csökkenti, cserébe viszont a magasabb esetszám miatt csökkenek a standard hibák. Ha tudjuk, hogy mi az a külső változó, ami felelős az adathiányért, lehet pótolni csoportonkénti átlagokat. Például, ha tudjuk azt, hogy az idősebbek inkább idegenellenesek, és az idegenellenesség változót kell pótolnunk, akkor tehetjük azt, hogy a nem válaszolóknak korcsoportonként becsült átlagokat pótlunk be. Az átlaggal való pótlás persze okozhat anomáliákat, például, ha olyan értékeket pótlunk be, ami az eredeti változóban nem is szerepelt, mondjuk egy 1–7 diszkrét osztályzatokat használó skálán 3,4-es értéket.

Ha tudjuk, hogy az adathiány véletlenszerű, és van arról is sejtésünk, hogy milyen változók befolyásolják az adathiány keletkezését, akkor összetettebb pótlási megoldást érdemes alkalmazni. Itt nagyon sok lehetőség áll előttünk. Kézenfekvő megoldás lehet regressziós pótlás alkalmazása. A pótolni kívánt változót tesszük meg függő változónak, az adathiánnyal összefüggő változókat pedig független változóknak. A felírt regressziós egyenlet becsült értékeivel pótolhatjuk a válaszhiányos eseteket. Itt még inkább előfordulhat, hogy az eredeti változó értéktartományban nem szereplő értékek kerülnek pótlásra (például negatív jövedelem), illetve a pótolt változó szórása is jelentősen csökken. Utóbbin lehet javítani, ha a pótlás során valamilyen véletlen hibataggal egészítjük ki a becslésünket. A pótlást megközelíthetjük *maximum likelihood* módszerekkel is, ahol azt keressük meg, hogy a független változók milyen paraméterei mellett kapjuk meg legnagyobb valószínűséggel a függő változó adott értékét. Erre épül például a bevett módszernek számító várható-érték maximalizálás (Dobi 2015), azaz EM-módszer (*expectation-maximization*). Az EM módszerek többségükben abból az előfeltevésből indulnak ki, hogy az adataink többdimenziós normál eloszlást követnek, ami gyakran túl erős feltevésnek bizonyul. Utóbbi hiányában érdemes lehet nem-paraméteres megoldások felé fordulni, amelyeknél a pótlás során megkeressük, hogy a nem válaszolóhoz melyik válaszoló van a legközelebb a fontosnak gondolt változók mentén. Ezek a módszerek általában valamilyen távolságmetrikán alapulnak, és szomszédokat keresnek. Innen jön a módszer neve is: k-legközelebbi szomszéd módszer. A pótlásnál itt is több stratégia közül választhatunk: bepótolhatjuk a legközelebbi eset értékét, a *k* legközelebbi esetből véletlenszerűen sorsolhatunk valakit vagy akár a legközelebbi esetek átlagát is beilleszthetjük a nem válaszolás helyére.

Fontos megemlíteni azokat a módszereket, amelyek többszörös adatpótlásra építenek, tehát nem egy, hanem több lehetséges értéket pótolnak be, aminek következtében nem egy, hanem több adatbázis fog keletkezni (Rubin 2004, Oravecz 2008). Ez statisztikailag egy jobb megoldás, viszont a későbbi elemzések nehézkesebbé válnak, ezért a módszer a társadalomtudományban nem igazán elterjedt.

A pótló eljárásokat úgy is csoportosíthatjuk, hogy csak belső információkat használunk fel (*hot-deck* módszerek) vagy külső adatforrásokat[4] is igénybe veszünk (*cold-deck* eljárások). Speciális esetet az, amikor panel adataink vannak, hiszen itt a pótlásnál a korábbi felvételek adataihoz is visszanyúlhatunk. Az ilyen pótlásoknál azonban még körültekintőbben kell eljárni, mivel az időbeli összehasonlítást rossz pótlási megoldásokkal el tudjuk lehetetleníteni.

4 Például, ha egy nyugdíjasnál adathiány van a jövedelmi kérdésben, és a kérdőív alapján van valamilyen világos képünk arról, hogy mit dolgozott életében, akkor megpróbálkozhatunk azzal, hogy a Nyugdíjfolyósítótól kapott adatok alapján pontosabbá tesszük a pótolt adatunkat.

Ahogy az alfejezet elején is jeleztük, nagyon vázlatosan foglaltuk csak össze a lehetséges pótlási eljárásokat, ennél jóval részletesebb módszertani összefoglalók magyar nyelven is elérhetők (lásd például Hunyadi–Vita 2002, Oravecz 2008, Dobi 2015).

3. Adatok és módszerek

Elemzésünkben az ESS első hét hullámának adatait használjuk fel.[5] Célunk nem az, hogy a tanulmányban korábban bemutatott Koch és Blohm (2009) elemzést megismételjük egy hosszabb idősoron, inkább egy mindennapi kutatási problémához közelebbi teszt terepet kerestünk. Az ESS felépítéséből adódóan kiválóan alkalmas arra, hogy különféle indexeket állítson elő a kutató. Egy olyan indexet kerestünk a teszteléshez, amit korábban már validáltak, kellően összetett (több komponensből áll), és összefüggésben van a mindennapi politikai tapasztalásokkal. Utóbbi azért fontos szempont, mert teljesülése esetén joggal várhatjuk azt, hogy az index reagálni tud a politikai változásokra, és a különböző politikai oldalak szimpatizánsai eltérő mértékben tartják érzékenynek az indexet alkotó kérdéseket. Ezeknek a feltételeknek az ESS adatokon már tesztelt DEREX index (Juhász–Krekó–Molnár 2014) kiválóan megfelelt, ezért ennek egyik alkomponensét, a rendszerellenességet választottuk az elemzéshez. A rendszerellenesség indexnek négy alindexe van (Juhász–Krekó–Molnár 2014: 29).

1. Elégedetlenség a politikai rendszerrel

A *„Mennyire elégedett a jelenlegi magyar kormány munkájával”* és a *„Mindent összevetve mennyire elégedett Magyarországon a demokrácia működésével”* 0–10 skálán mért kérdésekre mindkét esetben 0-t vagy 1-et válaszoltak (nagyon nem elégedett válaszopciók).

2. Bizalmatlanság a jogrendszerrel és a jogalkalmazókkal szemben

Az *„Ön személy szerint mennyire bízik a magyar jogrendszerben”* és az *„Ön személy szerint mennyire bízik a rendőrségben”* 0–10 skálán mért kérdésekre mindkét esetben 0-t vagy 1-et válaszoltak (nagyon nem bízik válaszopciók).

3. Bizalmatlanság a politikai elittel szemben

Az *„Ön személy szerint mennyire bízik a magyar Országgyűlésben”* és az *„Ön személy szerint mennyire bízik a politikusokban”* 0–10 skálán mért kérdésekre mindkét esetben 0-t vagy 1-et válaszoltak (nagyon nem bízik válaszopciók).

4. Bizalmatlanság a nemzetközi szervezetekkel szemben

Az *„Ön személy szerint mennyire bízik az ENSZ-ben”* és az *„Ön személy szerint mennyire bízik az Európai Parlamentben”* 0–10 skálán kérdésekre mindkét esetben 0-t vagy 1-et válaszoltak (nagyon nem bízik válaszopciók).

Azokat tekinti a DEREX modell rendszerellenesnek, akik a négy alindex közül legalább egy esetben rendszerellenesnek bizonyulnak.

5 Az ESS első hullámát hivatalosan 2002–2003-ban kérdezték le, utána két évente követték egymást a hullámok. Az adatfelvételek pontos idejét az ESS dokumentáció tartalmazza. Lásd: http://www.europeansocialsurvey.org/data/deviations_index.html.

A következő nem-válaszolás kezelési módszereket teszteltük le:

A nem válaszolók nem rendszerellenesek. Az első módszernél azt feltételeztük, hogy a nem válaszolók 0-tól és 1-től eltérő értéket mondtak volna, ha válaszoltak volna. Ez az eredeti tanulmányban használt megoldás, ezt tekintjük referenciának az elemzésünkben.

A nem válaszolók rendszerellenesek. Azt feltételeztük, hogy a nem válaszolók 0-t vagy 1-et válaszoltak volna a bevont kérdésekre.

A nem válaszolók kihagyása. Az indexet csak azokra számoltuk ki, akik az összes indexet alkotó kérdésre válaszoltak. Ezzel azt feltételeztük, hogy a válaszhiány teljesen véletlenszerű.

A nem válaszolók válaszainak pótlása statisztikai módszerekkel a szocio-demográfia változók segítségével. A nem-válaszolásról azt feltételeztük, hogy összefügg bizonyos demográfiai változókkal. Megkerestük, hogy a bevonható demográfiai változók mentén (nem, kor, szubjektív jövedelem, iskolai végzettség, gazdasági aktivitás: munkanélküli) a nem válaszolókhoz ki a legközelebbi válaszoló, akinek nem hiányos a válasza. Ennek a legközelebbi válaszadónak (donor) válaszával pótoltuk az adathiányt. A módszer technikaibb jellegű leírása a mellékletben található.

Ahogy korábban írtuk, a teszteléshez az ESS első hét hullámának adatait használtuk fel. Csak azt a 14 országot vettük be az elemzésbe, amelyek mind a hét vizsgált hullámban részt vettek: Belgium, Dánia, Egyesült Királyság, Finnország, Hollandia, Lengyelország, Magyarország, Németország, Norvégia, Spanyolország, Svájc, Svédország, Szlovénia, Portugália. A minta teljes mérete 185.049 fő volt.

Elemzésünk három lépésből állt. Először megvizsgáltuk, hogy az indexeket alkotó változókban mekkora az adathiány országonként és hullámonként. Ezt követően többszintű modellekkel körüljártuk, hogy a nem-válaszolás milyen demográfiai tényezőktől függ, majd az elemzés utolsó lépcsőjében leteszteltük, hogy a négy korábban felvázolt adathiány-kezelési stratégia milyen mértékű különbségekhez vezet az index kiértékelésekor.

Az ESS adatbázisban alapvetően három féle módon lehetett adathiány a vizsgált kérdésekben: lehetett visszautasítás, lehetett „nem tudja” válasz, vagy egy „nem válaszolnak” jelzett opció is előfordulhatott.[6] Az adathiány legnagyobb hányada (80–90%) a „nem tudja” opcióból érkezik, ezért a másik két kategória külön elemzése nagyon problematikus lett volna az alacsony elemszámok miatt. Az elemzésünkben ebből következően nem választjuk szét a különböző adathiány típusokat, hanem egyben kezeljük őket.

6 Több forrásból is származhat ilyen kódolás. Okozhatja az, ha egy papíros kitöltésnél „kihagyták” a kérdést, de akár az is lehet a magyarázat, hogy anonimitási okokból bizonyos kéréseknél a helyi kutatást vezetők kódolták erre a kategóriára a változót, vagy az is, ha egy szűrés miatt az adott kérdést nem kellett megválaszolnia a kérdezettnek.

4. Eredmények

4.1. A nem-válaszolás területi és időbeli mintái

Elemzésünk első részében megvizsgáltuk, hogy összességében mekkora jelentőségű problémával állunk szemben. Egy-egy kérdés esetében általában jelentéktelen szokott lenni az adathiány, kivéve néhány nagyon speciális kérdést, mint a jövedelem vagy a pártpreferencia. De ezeknél a kérdéseknél tudatosan lehet arra törekedni, hogy úgy alakítsuk ki az adatfelvételi designt, hogy minimalizáljuk a nem-válaszolást.[7] Indexek, tipológiák készítésénél viszont az egy-egy kérdésnél még jelentéktelennek tűnő adathiány gyorsan megnőhet olyan nagyságúra, amit már érdemes az elemzésünknél figyelembe venni. Az általunk vizsgált rendszerellenesség nyolc változóból áll össze. Az 1. táblázat országonként tartalmazza, hogy a hét adatfelvételi hullám alatt a kérdőívre válaszolók hány százalékánál nem volt adathiány, egy kérdésnél volt adathiány, és legalább két kérdésnél volt adathiány.

1. táblázat.
Az adathiány nagysága a 8 kérdésben országonként (a hét hullám adatai összevontan szerepelnek).

	-0 item	1 item	2 vagy több item
Belgium	92,0%	4,2%	3,8%
Dánia	85,9%	7,6%	6,5%
Egyesült Királyság	82,0%	8,8%	9,1%
Finnország	93,0%	3,4%	3,6%
Hollandia	89,6%	5,8%	4,7%
Lengyelország	75,5%	10,3%	14,2%
Magyarország	76,6%	10,2%	13,1%
Németország	87,3%	6,8%	6,0%
Norvégia	84,9%	11,8%	3,3%
Portugália	78,9%	6,8%	14,3%
Spanyolország	80,5%	7,4%	12,1%
Svájc	81,2%	9,4%	9,4%
Svédország	81,3%	11,0%	7,7%
Szlovénia	80,1%	9,1%	10,7%

Ahogy a táblából jól látható, az országok között igen nagy a szórás. Finnországban volt a legjobb az adatfelvétel a nem-válaszolás vonatkozásában, összességében pedig a nyugati és a skandináv országok inkább jól szerepeltek. A lista utolsó öt országa között pedig a mediterrán és a kelet-közép-európai országok találhatók, Magyarországgal és Lengyelországgal a rangsor legvégén. Utóbbi két országban a válaszadók közel negyedénél volt adathiány, és közel 15 százaléknál 2 vagy több item hiánya volt a jellemző.

Az időbeli összehasonlítás szintén hasonló változatosságot mutat (2. táblázat). A belgák 6. és 7. hullámbeli eredménye is nagyon jó ilyen szempontból, az esetek mindössze 4 százalékában volt valamilyen adathiány az indexen. Az időbeli összehasonlítás mutatja, hogy a belga eredményeket az 1. (2002-es) hullám adata húzza csak le, de a tendencia gyakorlatilag folyamatosan javul (2002-ben még papíros kérdezés volt Belgiumban, a további években már számítógéppel támogatott CAPI módszert használtak).

7 Ilyen megoldás lehet, ha a pártpreferencia kérdésre nem közvetlenül kell válaszolni a kérdezőbiztosnak, hanem egy lapon beikszelhetjük a preferenciánkat, majd ezt a lapot egy borítékba tehetjük, hogy a kérdezőbiztos ne lássa a válaszunkat.

2. táblázat. Az esetek hány százalékában volt legalább egy kérdésnél adathiány

	1	2	3	4	5	6	7
Belgium	18,50%	10,10%	7,30%	6,40%	5,00%	3,70%	4,10%
Dánia	18,50%	16,70%	14,00%	14,40%	13,10%	13,20%	9,10%
Egyesült Királyság	13,20%	16,90%	18,80%	13,60%	24,70%	24,20%	13,40%
Finnország	8,50%	8,60%	5,50%	7,20%	6,20%	6,20%	6,90%
Hollandia	12,10%	14,90%	9,60%	8,50%	9,60%	9,30%	8,50%
Lengyelország	31,80%	28,40%	22,10%	22,20%	21,90%	23,20%	19,70%
Magyarország	36,90%	22,00%	25,20%	22,80%	22,70%	17,70%	17,40%
Németország	12,70%	15,40%	14,80%	12,70%	14,50%	11,10%	8,20%
Norvégia	15,50%	14,30%	16,00%	14,70%	16,40%	14,80%	13,40%
Portugália	26,60%	18,00%	24,00%	24,80%	24,80%	13,70%	13,60%
Spanyolország	22,00%	22,20%	20,30%	24,50%	13,40%	13,50%	19,40%
Svájc	17,40%	22,30%	19,10%	21.70%	18,40%	15,90%	15,30%
Svédország	20,00%	16,80%	28,30%	16,00%	19,20%	12,60%	17,20%
Szlovénia	18,50%	22,80%	19,90%	17,20%	18,50%	19,30%	23,00%

A legmagasabb adathiány az 1. hullám magyar adatfelvételénél volt, az esetek közel 40 százalékában nem volt érvényes válasz, ami drámaian magas szám összehasonlítva a többi eredménnyel. Arra, hogy a 2002-es adatfelvételben miért ennyire magas a magyar nem-válaszolási arány, kézenfekvő válasz lehetne, hogy az átlagosnál feszültebb politikai helyzet miatt a kormánnyal és/vagy a demokráciával való elégedettséget mérő kérdéseknél nem mondták el az emberek a véleményüket. Feszült politikai helyzet alatt azt értem, hogy 2002-ben választási év volt, és némi meglepetésre (szemben a közvéleménykutatások várakozásaival) a baloldal tudott kormányt alakítani. Az adatfelvétel viszont már októberben volt, fél évvel a választások után, ami már egy nyugodtabb politikai helyzetet feltételez. Szintén felmerülhet esetleges magyarázatként, hogy a kormánnyal és demokráciával kapcsolatos elégedettség kérdésnél rosszul volt a kérdőívben a skála felcímkézve (bár a válaszlapokon jó volt a címke).[8] Az adatok változónkénti vizsgálata ezzel szemben azt mutatja, hogy az elégedettség kérdésekre viszonylag nagy arányban válaszoltak az emberek, ezzel szemben az ENSZ-szel és Európai Parlamenttel kapcsolatos bizalom kérdésénél extrém magas volt a „nem tudja” válaszok aránya, utóbbi kérdésnél 27 százalék. A jelzett szám annak a fényében még inkább különösen magas, ha összehasonlítjuk az 1999-es (tehát korábbi) *European Values Study* (EVS) EU-val kapcsolatos hasonló (bár máshogy megfogalmazott) kérdésével, ahol a válaszmegtagadás „mindössze” 13,6 százalék volt.

Összességében a legtöbb országban javuló tendenciát lehet látni, és ha átlagoljuk, hogy évente az esetek hány százalékánál volt adathiány, szintén látható a pozitív tendencia. Az első ESS adatfelvétel idején még a válaszolók közel 20 százalékánál volt adathiányos kitöltés, míg ugyanez az arány 13,5 százalékra javult a 7. hullámra. Vannak ezzel részben ellentétes tendenciák is. A brit adatfelvétel például ilyen kilógó eset, ahol az első négy hullámban folyamatosan bőven 20 százalék alatt megjelenő adathiány közel 25 százalékra ugrott fel az 5. és a 6. hullámban, hogy aztán újra visszaessen 13,4 százalékra. Esetükben szintén az ENSZ és Európai Parlament kapcsán felmért bizalom esetében nőtt meg extrém módon a nem tudja válaszok aránya.[9]

Az egyszerűbb leíró statisztikák markánsan mutatják, hogy mind időben, mind területi bontásban jelen-

8 http://www.europeansocialsurvey.org/data/deviations_country.html?year=2002&land=348

9 Az adatfelvétel módja nem változott esetükben, és az ESS dokumentációban sem jeleztek problémát ezeknél a kérdéseknél.

tős eltérések vannak az adathiány nagyságában. Az elemzés következő lépcsőjében annak megértésére fókuszálunk, hogy milyen szocio-demográfia háttérváltozókkal függ össze a nem-válaszolás.

4.2. Többszintű modell a nem-válaszolás struktúrájának megértésére

A nem-válaszolás akkor okoz jelentősebb problémát, ha nem teljesen véletlenszerű. Teljesen véletlenszerű adathiánynál „csak" azzal kell szembenéznünk, hogy az adathiány miatt kisebb mintákon kell számolnunk, ami a becsléseink standard hibáját növeli, de a kapott paraméterek torzítatlanok maradnak. Ha MAR típusú (véletlen) adathiánnyal szembesülünk, a paraméter becsléseink már nagyobb eséllyel lesznek torzítottak. Annak érdekében, hogy az adathiány természetét megértsük, egy regressziós modellt illesztettünk a nem-válaszolásra, pontosabban arra, hogy a nyolc vizsgált kérdésből hány esetben volt adathiány. A speciális adatstruktúra miatt az egyszerűbb OLS lineáris regresszió helyett többszintű modellt illesztettünk, mivel joggal feltételezhettük azt, hogy az adatok térben és időben is klaszterezettek (Koltai 2009). A modell első szintje az ország és az ESS adatfelvételi hullám kombinált változója volt (tehát minden ország minden adatfelvétele külön egységnek számított). Ez 98 adatpontot jelentett. A második szint pedig a válaszadók szintje volt. Megoldásként felmerülhetett volna, hogy egymásba ágyazzuk az idő és területi komponenseket, de külön-külön kevés megfigyelési egységünk lett volna ahhoz, hogy többszintű modellt illesszünk. Közbülső szintként a kérdezők is bekerülhettek volna a modellbe, de mivel ezt a hatást nem tudtuk volna kontrollálni az adatpótláskor, ezért ezt a szintet nem építettük be az elemzésbe.[10]

A modellben összesen öt független változót szereptettünk, a nemet (1: férfi, 2: nő), a kort, az iskolai végzettséget (5 kategóriás), a szubjektív jövedelmet (4 kategóriás, a magasabb érték jobb gazdasági helyzetet jelez), illetve azt, hogy volt-e a kérdezett munkanélküli a kérdezést megelőző egy hétben (0: nem volt, 1: volt). Az összes független változót folytonosként szerepeltettük a modellünkben.

A modellt több lépcsőben illesztettük. Az első lépcsőben csak a konstans került be a regresszióba fix hatásként. Ezt nevezzük referencia modellnek, mivel az összes későbbi modell illeszkedését ehhez tudjuk hasonlítani. Második lépcsőben engedtük, hogy a tengelymetszetek országonként és adatfelvételenként eltérjenek, ami praktikusan azt jelenti, hogy a modell megengedte, hogy az adathiány átlagos szintje országonként és időperiódusonként különböző legyen. Ez a modell segít nekünk azt lemérni, hogy a variancia hány százaléka keletkezik az első elemzési szinten (ország*adatfelvétel). A harmadik lépcsőben hozzáadtuk fix hatásként mind az öt független változót, majd külön-külön modellekben teszteltük, hogy ha az egyes változók meredekségét nem fixáljuk le, hanem engedjük, hogy országonként és adatfelvételenként eltérjenek egymástól, hogyan változik a modellünk illeszkedése.

10 A személyes adatfelvételekben egy kérdezőbiztos praktikusan egymáshoz közeli helyen szokott kérdezni. Ezért nagyon nem triviális, hogy hogyan tudjuk szétválasztani a kérdezőbiztos hatását a válaszadói struktúra területi homogenitásától.

3. táblázat: Az adathiány nagyságára illesztett többszintű lineáris modell

		Random tengely-metszet modell	Random tengelymetszet + fix szocio-demográfiai változók		
		B	B	Standard hiba	Beta
Fix hatás	Konstans	0,349	0,112***	0,018	-0,003
	Nem		0,182***	0,004	0,096
	Kor		0,003***	0,000	0,057
	Szubjektív jövedelmi helyzet		-0,068***	0,003	-0,059
	Iskolai végzettség		-0,107***	0,002	-0,150
	A kérdezett munkanélküli volt az elmúlt 1 héten		-0,039***	0,009	-0,010
Random hatás	reziduális (variancia)	0,876	0,796		
	Ország*hullám szint (variancia)	0,027	0,018		
BIC (fix konstans modell: 506.190)		501.147	475.508		

**** 0,001 szinten szignifikáns hatás*

A csak random tengelymetszetet tartalmazó modell segítségével megbecsülhető, hogy a vizsgált változó (adathiány nagysága) varianciájának mekkora aránya érkezik az első szintből, tehát az országok és hullámok közötti különbségből. A kapott érték mindössze 2,93 százalék, ami jelzi, hogy az adathiány elsősorban egyéni tényezőkkel magyarázható, nem országok és időszakok közötti különbséggel. A szocio-demográfiai változókat is tartalmazó modellben (a változók meredekségének fixen tartása mellett) az első szintre jutó variancia lecsökken 2,2 százaléka, ami arra utal, hogy az országok és hullámok közötti különbséget kismértékben magyarázza az országok és időszakok eltérő demográfiai profilja is (hasonló eredmények kapcsán lásd Koch–Blohm 2009).

A fix hatások a szakirodalomban is tárgyalt módon működtek, a nőknél, az idősebbeknél és a rosszabb anyagi helyzetben lévőknél, illetve munkanélkülieknél magasabb volt az adathiány, akárcsak az alacsonyabb végzettségű válaszolóknál. A legmarkánsabb hatása az iskolai végzettségnek volt, amit a nem változó követett, de nyilvánvalóan az iskolai végzettség és gazdasági helyzet közötti erős pozitív korreláció kontrollálta ezeknek a változóknak az amúgy erősebb nyers hatását. Többszintű modellek esetében nem triviális a megmagyarázott hányad megbecslése, a BIC mutató alapján azonban kiszámolhatunk egy pszeudo R^2 értéket, ami 6 százalék körüli megmagyarázott varianciát mutat.

Fontos kérdés az is, hogy vajon az egyes demográfiai változók eltérően függenek-e össze az adathiány nagyságával országonként és hullámonként. Ennek megvizsgálására egy olyan modellt futtattunk, ahol megengedtük, hogy az első szinten a független változóink eltérő meredekséget vegyenek fel országonként és hullámonként. Az eredmények minden esetben azt mutatták, hogy szignifikánsan javítja a modellek illeszkedését, ha nem fixáljuk az első szinten a B paramétereket, azonban az összes esetben gyenge volt a modell illeszkedések javulása a BIC mutatók szerint. Az előbb említett pszeudo R^2 mutató alapján a legnagyobb növekedés is mindössze 0.3 százalékpontos volt az illeszkedésben, utóbbi egyébként a kor változó esetében.

Mivel a függő változó valójában nem folytonos, hanem diszkrét eloszlást követ, statisztikai szempontból a lineáris becslő modell használata nem feltétlen ad pontos eredményt. Az eredmények robusztusságának tesztelésére egy olyan többszintű modellt is illesztettünk, ahol negatív binomiális[11] eloszlást használtunk (lásd

11 Mivel a függő változó varianciája és átlaga jelentősen eltért, ezért a Poisson modell nem lehet volna megfelelő választás.

melléklet M1. táblázat). A kapott eredmények egy paraméter kivételével megegyeztek a lineáris modellel – a negatív binomiális modellben a munkanélküliség változó nem volt szignifikáns. Olyan további teszteket is végeztünk, amelyekben a kor, az iskolai végzettség és a szubjektív jövedelem változókat kategoriálisan vontuk be a modellbe. Ennek a modellnek az illeszkedése valamivel magasabb volt, mint az alapmodellé. Ez abból következett, hogy a kor esetében a függő változóval való kapcsolat nem lineáris, hanem sokkal inkább U alakú, miszerint a legfiatalabbak és a legidősebbek rendelkeznek a legtöbb válaszhiánnyal és a 40–49 valamint az 50–59 korosztály volt az, aki legkevesebb válaszhiánnyal töltötte ki a kérdőíveket.

Összességében az eredmények alátámasztják, hogy az adathiány nem teljesen véletlenszerű volt, tehát a rendszerellenesség nagyságára feltehetően csak torzított becslést tudunk adni, ha nem vesszük figyelembe az adathiányt.

4.3. Az eredmények eltérései az adathiány kezelésének függvényében

Az előző két részben körbejártuk, hogy időben és országok között milyen különbségek vannak a nem-válaszolási minták között. Ahogy az előző rész is mutatta, a nem-válaszolás nem teljesen véletlenszerű, a vizsgálatba bevont demográfiai változók mentén szignifikánsan eltértek az érvényes válasz arányai. Ez azt indikálja, hogy érdemes lehet tovább foglalkozni azzal, hogy milyen hatása lehet a nem-válaszolásnak a végső eredményekre. Ennek érdekében négy, az adathiány kezelése kapcsán felmerülő, egymástól eltérő megoldást hasonlítottunk össze.

Ha azt feltételeztük, hogy az adathiányos válaszolók nem rendszerellenesek (1-nél magasabb értékkel pótoltuk az adathiányt az indexet alkotó változókban), akkor a rendszerellenesség átlaga az összes ország összes hullámát tekintve 13,9 százalék volt (1. módszer). Ha ezzel szemben a hiányzó értékeket úgy pótoltuk, hogy azok rendszerellenes válaszokra utaljanak, akkor az index értéke felugrott 20,3 százalékra (2. módszer). A harmadik megoldásnál kihagytuk a nem válaszolókat, ez lecsökkentette az index értékét 13,5 százalékra, ami 0,4 százalékponttal kisebb, mint az eredeti verzióban kapott érték (3. módszer). Ha pedig pótoltuk a nem-válaszolást, akkor 14,6 százalékra ment fel az index értéke, ami 0,7 százalékpontos növekmény a referenciának tekinthető 1-es megoldáshoz képest (4. módszer). A 2. módszer nagyon eltérő eredménye mutatja, hogy az a megoldás jelentős torzítást visz az index értékbe, ezért a továbbiakban ezzel nem is foglalkozunk részletesebben.

A mellékletben az összes országra és hullámra lebontva bemutatjuk az eredményeket (Melléklet 3. táblázat) külön pirossal kiemelve azt, ahol 1 százalékpontnál nagyobb az eltérés a 2. és a 3. modell eredményében az 1. modellhez képest.

Az adathiányos esetek kihagyása, ahogy láttuk, összességében alacsonyabb rendszerellenességet hozott ki, de természetesen ez országonként és hullámonként nagyon változó. Magyarországon a negyedik hullám esetében az index értéke 1,7 százalékponttal magasabb lett volna, míg a többi hullámban valamivel csökkent volna az index érte. Összességben nagyon kevés esetben volt 1 százalékpontos különbség az 1. és 3. módszer között, és 2 százalékpontnál több egyedül a portugál 5. hullám esetében.

A 4. módszer pótlásos megközelítése már jelentősebb eltéréseket hozott az 1. referencia módszerhez képest. Egy kivételtől eltekintve (Svédország, 7. hullám) minden országban magasabb lett a rendszerellenesség

az adatpótlás hatására az 1. módszerhez képest. A különbség azokban az országokban volt a legnagyobb, ahol a rendszerellenesség index értéke amúgy is magas volt: Magyarországon, Lengyelországban, Szlovéniában, Spanyolországban és Portugáliában. Ez egyébként pont az az öt ország volt, ahol az adathiány is a legmagasabb volt ezekben a kérdésekben. Tehát azokban az országokban és hullámokban, ahol a rendszerellenesség magas volt, az adathiány is magasabb volt, és a pótlás hatására a rendszerellenesség még magasabb lett volna. A spanyol 7. és a portugál 5. hullám ilyen szempontból a legkirívóbb, mivel itt több mint 2 százalékponttal magasabb rendszerellenességet kaptunk a pótlás hatására. A portugál 5. hullám olyan szempontból még érdekesebb, hogy a 3. és 4. módszer alapján (kihagyás és pótlás) közel 5 százalékpontos különbség van az index értékében (35,0% vs. 39,9%).

5. Szintézis

Aki survey adatokkal dolgozik, el kell, hogy fogadja, hogy az adathiány sajnálatos, de természetes velejárója az adatfelvételnek. Van, amikor ez a probléma olyan jelentéktelen, hogy eltekinthetünk a kezelésétől, vannak azonban olyan esetek, amikor az adathiányt már érdemes számításba vennünk eredményeink elemzésekor. Írásunkban amellett érveltünk, hogy időben és térben kiterjedt adatok esetében több olyan egyedi szempont is felmerülhet, ami az adatfelvételek között, eltérő válaszadói struktúrát eredményezhet. Az ESS adatok nagy szórást mutattak abban, hogy a vizsgált rendszerellenesség indexet alkotó változók között mekkora volt az adathiány. Általában a nyugati és a skandináv országokban magasabb volt az érvényes válaszok aránya, míg keleten és a mediterrán országokban alacsonyabb, és összességben a nem-válaszolási arány hullámról hullámra csökkent. A többszintű regressziós modellek azt mutatták, hogy a nem-válaszolási arányok közötti különbség nagyjából 3 százalékáért felel az ország és az adatfelvétel hulláma, míg egyéni szinten a rosszabb szocio-kulturális környezet magasabb nem válaszolással jár együtt, akárcsak az, ha a válaszadó nő vagy idősebb.

A nem-válaszolás kezelésének több forgatókönyvét is teszteltük. Összességében azt mondhatjuk, hogy drámai változást nem okozott az eltérő kezelési stratégia (kivéve a 2. módszert, ami azonban annyira extrém, hogy nem is foglalkoztunk vele részletesebben). Azonban érdekes és fontos eredmény volt, hogy a nem válaszoló esetek kihagyása általánosságban alacsonyabb rendszerellenességet hozott volna ki, mint az 1. referencia módszer (ahol azt feltételezzük, hogy a nem válaszoló az adott kérdés esetében nem a rendszerellenesnek tekintett véleményt választotta volna). Ez első lépcsőben kontraintuitív eredménynek tűnhet, hiszen logikusabb lenne azt feltételezni, hogy az 1. módszernél kapunk alacsonyabb rendszerellenességet. Az húzódik meg az eredmények mögött, hogy azok az esetek, amelyeket a 3. módszernél kidobunk, a többi bevont változó alapján már inkább rendszerellenesnek mutatkoznak még annak ellenére is, hogy magánál az adathiányos változónál nem feltételezünk rendszerellenességet.

Ezzel szemben a pótlásos stratégia szinte kivétel nélkül növeli a rendszerellenesség nagyságát. A különbség még nagyobb lenne, ha az egyébként a szociológiai gyakorlatban egyik leginkább elterjedt kihagyásos strarégiát (3. módszer) hasonlítjuk össze a pótlásos megoldással. Itt van olyan ország és hullám, ahol közel 5 százalékos különbséget mértünk a két módszer kimenete között. Az, hogy a rendszerellenesség nagyobb a pótlás után, nem lenne triviális, de a mellékletben található egyszerű logisztikus regressziós módszer rávilágít arra,

hogy ez miért van (lásd Melléklet 4. táblázat). Az összes ország összes hullámának bevonásával modelleztük, hogy a pótlásnál figyelembe vett demográfiai változók hogyan függenek össze azzal, hogy valaki rendszerellenes-e. Az eredmények azt mutatták, hogy ugyanazok a változók táplálják a rendszerellenességet, amelyek a nem-válaszolást is. Tehát a rosszabb anyagi helyzetben lévők, a munkanélküliek, az alacsonyabb végzettségűek és az idősebbek azok, akik Európában inkább rendszerellenesek. Mindössze a nem változó működött másként ilyen szempontból, mivel a férfiak inkább számítanak rendszerellenesnek, az adathiány viszont a női válaszadókra jellemzőbb volt. Tehát általánosságban elmondható, hogy a nem válaszolók demográfiai profilja közel áll a rendszerellenesek demográfiai profiljához, aminek hatására a pótláskor majdnem minden vizsgált adatfelvételben megnőtt a rendszerellenesség mértéke. Ez ráadásul azt eredményezte, hogy az egyes országok közötti távolságok markánsabbak lettek, mint amit a referencia modell mutatott.

Természetesen felmerülhet kritikaként, hogy csak egyetlen pótlási eljárást teszteltünk. A regressziós és EM módszerek természetükből adódóan átlaghoz közelebbi értékeket pótolnak, ami az index logikája szerint az 1. módszerhez hasonló eredményeket adott volna. De mivel a vizsgált index nem az alkérdések átlagaira, hanem azok extrémebb értékeire koncentrált (0–1 választás a skálákon), ezért véleményünk szerint ezek a pótlásos módszerek ebben az esetben nem lettek volna adekváltak.

A bevont változók köre is viszonylag szűk volt, itt is lehetett volna valószínűleg más változókat is keresni. A szűk változószett azonban egy nagyon tudatos döntés volt, olyan szocio-demográfiai kérdések kerültek felhasználásra, amelyek a nem-válaszolás szempontjából korábban nemzetközileg már tesztelve lettek. A munkanélküliség lóg ki ebből a sorból, utóbbit viszont a rendszerellenesség kapcsán gondoltuk fontosnak. Az is fontos elem volt a bevont független változókban, hogy ezekben a kérdésekben nagyon alacsony legyen az adathiány, ne kerüljünk végeláthatatlan adatpótlási spirálba. Az összesített adathiány kevesebb, mint 2 százalék volt ezekben a változókban.

A pótlás persze sosem lehet tökéletes, nem tudhatjuk, hogy kimaradt-e bármilyen fontos változó. A sejtésünk erre általában az lehet, hogy igen. Egy rendszerellenes indexnél például logikus lehet annak a feltételezése, hogy az adott időszakban kormányon lévő párt hívei inkább elégedettek a rendszerrel, az ellenzéki szimpatizánsok inkább elégedetlenek, és ha az ellenzéki szavazók még a válaszolásnál is rejtőzködőbbek, akkor ennek a dimenziónak a mentén is torzul az eredményünk. Ezt a hatást viszont nehéz modellezni, mert a pártpreferencia is nehezen mérhető, magas a meghiúsulás a változó esetében, ezért nehéz rajta keresztül pótlásokat végezni. Azt a hatást pedig végképp nem tudjuk kontroll alatt tartani, ha a vizsgált változó olyan adathiányt rejt, ami külső hatásokkal nem modellezhető (nem véletlen adathiány, NMAR).

Az adathiányt a tanulmányban egyben kezeltük, de logikusan felmerülhet az is, hogy külön kezeljük a „nem tudja” és a „nem válaszol?” opciókat, mivel ezeknek eltérő lehet a generáló mechanizmusa. De az is egy potenciális további iránya lehetne a munkának, hogy az eset (unit) nem-válaszolással is kiegészítsük a vizsgálódásainkat.

Egy ilyen jellegű módszertani tanulmány végén érdemes lehet az eredmények alapján követendő stratégiát felvázolni, azonban az elemzés is mutatta, nincsenek triviálisan jó megoldások, több alternatíva is lehet,

amit érdemes tesztelni. Két potenciális tanácsot lehet talán megfogalmazni. Az egyik, hogy nagyobb adattömegek mozgatása után, főleg, ha időbeli és térbeli összehasonlításra is vállalkozunk, érdemes megvizsgálni, hogy mekkora az adathiány nagysága, és ha jelentősebb, azt is érdemes letesztelni, hogy vannak-e olyan változók, amik összefüggenek az érvényes válaszok arányával. Amennyiben ezekre a kérdésekre igen választ kapunk, az még nem feltétlenül jelenti azt, hogy mindenképpen adatpótlást kell alkalmaznunk. Azonban ebben az esetben érdemes akár kisebb tesztek segítségével megvizsgálni az eredményeink robusztusságát, és adott esetben érdemes lehet beavatkozni. Az adathiány vizsgálata azért is különösen fontos lehet, mert segít nekünk képet alkotni arról is, hogy milyen volt az adatfelvétel minősége. A nagyon magas visszautasítás („nem szeretnék válaszolni" opció) jelezheti azt, hogy a kérdező és válaszoló közötti bizalmi légkör nem tudott létrejönni, a „nem tudja" válaszok magas aránya pedig a kérdőív bonyolultságát, nehézségét jelentheti.

Egy olyan gondolattal zárjuk az elemzést, amit már korábban is megfogalmaztunk. Az adathiány természetes velejárója az elemzéseinknek, és ha tudatosan nem foglalkozunk vele, akkor is kezeljük valahogy. Ez a kezelés sok esetben a használt statisztikai programok működési módjából következik. Ha nem akarjuk, hogy a programok vezessenek minket, alakítsunk ki egyértelmű protokollt a nem-válaszolás kezelésére, mert ez lehet a záloga annak, hogy az eredményeink érvényesek és megbízhatóak legyenek.

Irodalomjegyzék

Beullens, K. – Loosveldt, G. (2016) Interviewer Effects in the European Social Survey. *Survey Research Methods,* 10(2), 103–118. http://dx.doi.org/10.18148/srm/2016.v10i2.6261

Blom, A. G. (2008) Measuring nonresponse cross-nationally. *ISER Working Paper Series*, No. 2008–41.

Bowling, A. (2005) Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27(3), 281–291. http://dx.doi.org/10.1093/pubmed/fdi031

De Leeuw, E. D. (2001) Reducing missing data in surveys: An overview of methods. *Quality & Quantity*, 35(2), 147–160. http://dx.doi.org/10.1023/A:1010395805406

Dobi B. (2015) *A többszörös imputálás gyakorlati alkalmazása*. Eötvös Loránd Tudományegyetem, Szakdolgozat.

György E. (2004) A nemválaszolás elemzése a munkaerő felvételben. *Statisztikai Szemle,* 82(8), 747–772.

Hunyadi L. – Vita L. (2002) *Statisztika közgazdászoknak*. Budapest: Központi Statisztikai Hivatal.

Jäckle, A. – Roberts, C. – Lynn, P. (2010) Assessing the effect of data collection mode on measurement. *International Statistical Review*, 78(1), 3–20. http://dx.doi.org/10.1111/j.1751-5823.2010.00102.x

Juhász A. – Krekó P. – Molnár Cs. (2014) A szélsőjobb iránti társadalmi kereslet változása Magyarországon. *Socio.hu Társadalomtudományi Szemle,* 4(4), 25–55. http://dx.doi.org/10.18030/socio.hu.2014.4.25

Kmetty Z. (2012) A telefonos kutatások speciális problémái. *Statisztikai Szemle,* 90(1), 41–63.

Koch, A. – Blohm, M. (2009) Item Non-response in the European Social Survey. *ASK. Research and Methods*, *18*(1), 45–65.

Koltai, J. (2009) A nyugdíjrendszerrel kapcsolatos igazságossági attitűdök – egy bonyolult struktúrájú adatbázis feldolgozási lehetőségei. In Némedi D. – Szabari V. (szerk.) *Kötő-jelek 2009: Az Eötvös Loránd Tudományegyetem Társadalomtudományi Kar Szociológiai Doktori Iskolájának Évkönyve*. Budapest: ELTE TÁTK Szociológia Doktori Iskola, 137–169.

Little, R. J. A. – Rubin, D. B. (1987) *Statistical analysis with missing data.* New York: John Wiley & Sons.

Lynn, P. (1998) Data collection mode effects on responses to attitudinal questions. *Journal of Official Statistics*, *14*(1), 1–14.

Lugtig, P. J. – Lensvelt-Mulders, G. J. L. – Frerichs, R. – Greven, A. (2011) Estimating nonresponse bias and mode effects in a mixed mode survey. *International Journal of Market Research*, *53*(5), 669–686. http://dx.doi.org/10.2501/IJMR-53-5-669-686

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

Peng, C. Y. J. – Harwell, M. – Liou, S. M. – Ehman, L. H. (2006) Advances in missing data methods and implications for educational research. In Sawilowsky, S. S. (szerk.) *Real data analysis.* Charlotte, North Carolina: Information Age Publishing, 31–78.

Pickery, J. – Loosveldt, G. (2001) An exploration of question characteristics that mediate interviewer effects on item nonresponse. *Journal of Official Statistics*, *17*(3), 337–350.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. Hoboken, New Jersey: John Wiley & Sons.

Sigelmann, L. (1981) Question-order effects on presidential popularity. *Public Opinion Quarterly*, *45*(2), 199–207. http://dx.doi.org/10.1086/268650

Stoop, I. – Billiet, J. – Koch, A. – Fitzgerald, R. (2010) *Improving survey response: Lessons learned from the European Social Survey*. Chichester: John Wiley & Sons.

Sturgis, P. – Baker, N. – Callegaro, M. – Fisher, S. – Green, J. – Jennings, W. – Kuha, J. – Lauderdale, B. – Smith, P. (2016) *Report of the Inquiry into the 2015 British general election opinion polls.* London: Market Research Society and British Polling Council.

Tourangeau, R. – Rasinski, K. A. – Bradburn, N. – D'Andrade, R. (1989) Carryover effects in attitude surveys. *Public Opinion Quarterly*, *53*(4), 495–524. http://dx.doi.org/10.1086/269169

Felhasznált ESS adatbázisok

ESS Round 7: European Social Survey Round 7 Data (2014). Data file edition 2.1. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 6: European Social Survey Round 6 Data (2012). Data file edition 2.3. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 5: European Social Survey Round 5 Data (2010). Data file edition 3.3. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 4: European Social Survey Round 4 Data (2008). Data file edition 4.4. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 3: European Social Survey Round 3 Data (2006). Data file edition 3.6. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 2: European Social Survey Round 2 Data (2004). Data file edition 3.5. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

ESS Round 1: European Social Survey Round 1 Data (2002). Data file edition 6.5. NSD- Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.

Melléklet

1. táblázat. Az adathiány nagyságára illesztett többszintű negatív binomiális modell

		Random tengelymetszet + fix szocio-demográfia változók		
		B	Standard hiba	Sig.
Fix hatás	Konstans	-1,661	0,064	0,000
	Nem	0,592	0,014	0,000
	Kor	0,004	0,000	0,000
	Szubjektív jövedelmi helyzet	-0,169	0,009	0,000
	Iskolai végzettség	-0,391	0,006	0,000
	A kérdezett munkanélküli volt az elmúlt 1 héten	-0,037	0,029	0,207
	reziduális (variancia)			
	Ország*hullám szint (variancia)	0,238		
BIC		942.723		

2. táblázat. Az adathiány nagyságára illesztett többszintű negatív binomiális modell kategoriális független változókkal

			Random tengelymetszet + fix szocio-demográfiai változók		
			B	Standard hiba	Sig.
Sig. Fix hatás	Konstans		-2,355	0,068	0,000
	Nem		0,587	0,014	0,000
	Szubjektív jövedelem (Referencia: Nagyon nehezen jönnek ki a jövedelmükből)	Kényelmesen megélnek	-0,550	0,034	0,000
		Kijönnek a jövedelmükből	-0,399	0,031	0,000
		Nehezen, de megélnek	-0,179	0,033	0,000
	Kor (Referencia: 60+)	15–29	-0,009	0,020	0,000
		30–39	-0,230	0,023	0,000
		40–49	-394	0,023	0,000
		50–59	-0,397	0,022	0,000
	Iskolai végzettség (Referencia: Diploma)	8 általános vagy kevesebb	1,484	0,027	0,000
		Szakmunkás	0,949	0,024	0,000
		Érettségi	0,464	0,022	0,000
		Felsőfokú technikum, továbbképzés	0,128	0,037	0,000
	A kérdezett munkanélküli volt az elmúlt 1 héten		-0,003	0,030	0,925
Random hatás		Reziduális (variancia)	1		
		Ország*hullám szint (variancia)	0,256		
BIC			932.821		

3. táblázat. A rendszerellenesség nagysága a különböző adatkezelési stratégiák esetében

Ország	Hullám	Módszer 1	Módszer 2	Módszer 3	Módszer 4	Módszer 1 – Módszer 3	Módszer 1 – Módszer 4
Belgium	1	12,10%	20,20%	11,40%	13,20%	-0,70%	1,10%
	2	10,40%	14,30%	9,90%	10,60%	-0,50%	0,20%
	3	9,60%	11,80%	9,30%	9,80%	-0,30%	0,20%
	4	11,40%	13,10%	11,40%	11,40%	0,00%	0,00%
	5	13,00%	14,60%	13,30%	13,10%	0,30%	0,10%
	6	9,40%	10,80%	9,20%	9,50%	-0,20%	0,10%
	7	13,20%	14,30%	13,30%	13,30%	0,10%	0,10%
Svájc	1	4,20%	10,80%	4,70%	4,60%	0,50%	0,40%
	2	4,70%	14,40%	5,30%	5,20%	0,60%	0,50%
	3	4,80%	12,00%	5,30%	5,20%	0,50%	0,40%
	4	5,00%	15,40%	5,20%	5,40%	0,20%	0,40%
	5	6,50%	13,60%	6,80%	6,50%	0,30%	0,00%
	6	4,50%	11,10%	4,30%	5,00%	-0,20%	0,50%
	7	6,10%	10,80%	5,80%	6,20%	-0,30%	0,10%
Német-ország	1	12,40%	16,40%	12,30%	12,70%	-0,10%	0,30%
	2	15,60%	21,70%	15,90%	16,10%	0,30%	0,50%
	3	16,20%	22,40%	16,20%	16,70%	0,00%	0,50%
	4	11,30%	15,90%	11,00%	11,50%	-0,30%	0,20%
	5	15,10%	20,20%	15,60%	15,80%	0,50%	0,70%
	6	8,90%	12,90%	8,50%	9,30%	-0,40%	0,40%
	7	11,20%	14,00%	11,30%	11,40%	0,10%	0,20%
Dánia	1	3,60%	12,20%	3,70%	3,90%	0,10%	0,30%
	2	2,70%	9,40%	2,30%	2,80%	-0,40%	0,10%
	3	3,70%	9,20%	3,50%	3,90%	-0,20%	0,20%
	4	3,40%	8,50%	3,30%	3,40%	-0,10%	0,00%
	5	4,30%	9,60%	4,20%	4,50%	-0,10%	0,20%
	6	3,30%	8,70%	3,10%	3,40%	-0,20%	0,10%
	7	5,10%	7,90%	4,80%	5,10%	-0,30%	0,00%
Spanyolország	1	12,90%	25,30%	13,80%	14,30%	0,90%	1,40%
	2	10,30%	20,40%	9,40%	11,30%	-0,90%	1,00%
	3	10,70%	22,80%	11,00%	11,30%	0,30%	0,60%
	4	10,10%	24,20%	10,20%	10,90%	0,10%	0,80%
	5	17,70%	24,10%	18,20%	18,40%	0,50%	0,70%
	6	32,80%	38,60%	33,70%	33,90%	0,90%	1,10%
	7	28,00%	37,40%	28,90%	30,10%	0,90%	2,10%
Finnország	1	5,10%	8,30%	4,70%	5,30%	-0,40%	0,20%
	2	4,30%	7,40%	4,10%	4,40%	-0,20%	0,10%
	3	4,40%	6,70%	4,20%	4,40%	-0,20%	0,00%
	4	3,70%	6,00%	3,50%	3,70%	-0,20%	0,00%
	5	6,00%	8,70%	5,70%	6,10%	-0,30%	0,10%
	6	3,40%	5,60%	3,20%	3,40%	-0,20%	0,00%
	7	6,20%	9,10%	5,90%	6,50%	-0,30%	0,30%
Egyesült Királyság	1	12,80%	17,00%	12,70%	13,20%	-0,10%	0,40%
	2	15,50%	22,70%	15,70%	16,40%	0,20%	0,90%
	3	15,90%	22,70%	16,10%	16,50%	0,20%	0,60%
	4	17,40%	22,70%	17,70%	18,00%	0,30%	0,60%
	5	18,90%	29,10%	18,40%	20,30%	-0,50%	1,40%
	6	14,50%	25,60%	14,20%	15,80%	-0,30%	1,30%
	7	19,60%	25,00%	19,50%	20,40%	-0,10%	0,80%

Ország	Hullám	Módszer 1	Módszer 2	Módszer 3	Módszer 4	Módszer 1 – Módszer 3	Módszer 1 – Módszer 4
Magyarország	1	12,40%	31,50%	12,00%	14,20%	-0,40%	1,80%
	2	22,30%	30,50%	21,60%	23,30%	-0,70%	1,00%
	3	32,60%	42,70%	32,20%	34,50%	-0,40%	1,90%
	4	45,50%	54,60%	47,20%	47,40%	1,70%	1,90%
	5	20,90%	31,20%	20,40%	22,10%	-0,50%	1,20%
	6	28,90%	34,40%	28,50%	29,70%	-0,40%	0,80%
	7	22,20%	28,60%	22,60%	23,20%	0,40%	1,00%
Hollandia	1	6,80%	11,40%	6,50%	7,00%	-0,30%	0,20%
	2	8,30%	12,70%	8,30%	8,50%	0,00%	0,20%
	3	5,20%	9,10%	5,20%	5,60%	0,00%	0,40%
	4	4,70%	7,80%	4,80%	4,90%	0,10%	0,20%
	5	5,30%	8,60%	5,00%	5,30%	-0,30%	0,00%
	6	5,70%	8,60%	5,30%	5,80%	-0,40%	0,10%
	7	6,80%	9,60%	6,90%	7,00%	0,10%	0,20%
Norvégia	1	4,10%	6,40%	3,90%	4,30%	-0,20%	0,20%
	2	5,50%	7,20%	5,20%	5,60%	-0,30%	0,10%
	3	4,20%	6,70%	3,90%	4,20%	-0,30%	0,00%
	4	4,10%	6,80%	4,00%	4,10%	-0,10%	0,00%
	5	3,50%	6,50%	3,10%	3,50%	-0,40%	0,00%
	6	2,60%	5,10%	2,70%	2,80%	0,10%	0,20%
	7	3,00%	5,20%	3,00%	3,00%	0,00%	0,00%
Lengyelország	1	21,80%	37,60%	21,80%	23,70%	0,00%	1,90%
	2	38,10%	46,80%	37,20%	39,60%	-0,90%	1,50%
	3	32,30%	40,70%	32,50%	33,70%	0,20%	1,40%
	4	27,30%	35,90%	26,70%	28,80%	-0,60%	1,50%
	5	22,10%	31,40%	21,40%	23,40%	-0,70%	1,30%
	6	29,30%	38,30%	29,80%	31,10%	0,50%	1,80%
	7	35,70%	42,80%	35,50%	37,50%	-0,20%	1,80%
Portugália	1	16,10%	32,10%	16,20%	17,30%	0,10%	1,20%
	2	28,00%	36,20%	26,80%	29,90%	-1,20%	1,90%
	3	22,70%	36,00%	21,70%	24,30%	-1,00%	1,60%
	4	27,10%	39,30%	25,80%	28,80%	-1,30%	1,70%
	5	37,40%	47,00%	35,00%	39,90%	-2,40%	2,50%
	6	40,90%	46,00%	40,40%	42,00%	-0,50%	1,10%
	7	38,70%	43,90%	38,70%	40,10%	0,00%	1,40%
Svédország	1	4,70%	11,00%	4,30%	5,00%	-0,40%	0,30%
	2	8,30%	13,80%	8,00%	8,50%	-0,30%	0,20%
	3	5,10%	12,70%	5,40%	5,30%	0,30%	0,20%
	4	4,80%	9,80%	4,40%	4,90%	-0,40%	0,10%
	5	3,10%	10,80%	3,10%	3,20%	0,00%	0,10%
	6	5,10%	8,80%	5,10%	5,20%	0,00%	0,10%
	7	4,40%	8,00%	4,00%	4,30%	-0,40%	-0,10%
Szlovénia	1	19,40%	28,60%	20,00%	20,20%	0,60%	0,80%
	2	18,30%	28,80%	19,20%	20,10%	0,90%	1,80%
	3	16,80%	25,50%	16,50%	17,50%	-0,30%	0,70%
	4	14,70%	22,40%	13,90%	15,10%	-0,80%	0,40%
	5	35,00%	41,30%	36,10%	36,80%	1,10%	1,80%
	6	33,70%	40,60%	33,40%	34,90%	-0,30%	1,20%
	7	37,60%	44,00%	38,90%	39,30%	1,30%	1,70%

4. táblázat. A rendszerellenességet vizsgáló binomiális logisztikus regressziós modell (összes ország, összes hullám egyben kezelve – N=185043)

	B	S.E.	Sig.	Exp(B)
Nem	-0,123	0,014	0,000	0,884
Kor	0,003	0,000	0,000	1,003
Szubjektív jövedelmi helyzet	-0,604	0,008	0,000	0,546
Iskolai végzettség	-0,181	0,006	0,000	0,834
A kérdezett munkanélküli volt az elmúlt 1 hétben	0,246	0,026	0,000	1,279
Konstans	-2,536	0,040	0,000	0,079
Nagelkerke R2	0,086			

Az adatpótlás technikai megvalósítása (4. módszer)

- Nem vettük figyelembe az adatpótláskor azokat az eseteket, ahol valamelyik bevont demográfiai változóban adathiány volt.
- Az öt pótláshoz használt változót (nem, kor, szubjektív jövedelem, iskolai végzettség, munkanélküli) országonként és hullámonként standardizáltuk.
- A pótlást országokon és hullámokon belül végeztük.
- Minden nem válaszoló esetében megkerestük, hogy az öt vizsgált szocio-demográfiai változó mentén melyik adathiány nélküli válaszoló van a legközelebb, őt tekintettük donor válaszolónak.
- A távolság meghatározásakor négyzetes euklidészi távolságot használtunk.
- Ha több lehetséges donor válaszoló is volt, akkor véletlenszerűen választottunk közülük.
- Nem korlátoztuk annak a számát, hogy ki hányszor lehet donor.
- Azokban az indexet alkotó kérdésekben, ahol adathiány volt, kipótoltuk a donor válaszadó adott kérdésre adott válaszával az adathiányt.
- A pótolt kérdések alapján újraszámoltuk a rendszerellenesség indexet.