

HIÁNYZÓ ADATOK KEZELÉSE

- TANULMÁNY A 2002-ES MAGYAR PARLAMENTI VÁLASZTÁSOK
ELŐREJELZÉSEINEK TÉVEDÉSE KAPCSÁN -

MISSING DATA ANALYSIS

- CASE STUDY ON THE INACCURATE PROGNOSIS REGARDING TO
THE 2002 PARLIAMENTARY ELECTIONS IN HUNGARY -

Fellner Zita

Veit Krisztina

Kézirat lezárva: 2012.03.25.

TARTALOMJEGYZÉK

1.	BEVEZETÉS (VK)	- 3 -
2.	A 2002-ES PARLAMENTI VÁLASZTÁSOK (VK)	- 4 -
2.1.	Előrejelzések és eredmények	- 4 -
2.2.	Hibák csoportosítása	- 5 -
2.3.	Az inkonzisztencia lehetséges okai	- 6 -
2.3.1.	<i>Utánfutó effektus</i>	- 8 -
2.3.2.	<i>Mobilizációs készség</i>	- 9 -
2.3.3.	<i>Rejtőzködők – látens közvélemény</i>	- 10 -
3.	HIÁNYZÓ ADATOK STATISZTIKAI KEZELÉSE (FZ)	- 11 -
3.1.	Hiányzó adatok típusai	- 11 -
3.2.	Hiányzó adatok kezelése	- 13 -
3.2.1.	<i>Teljesen megfigyelt, vagy elérhető egységek elemzésén alapuló eljárások</i>	- 13 -
3.2.2.	<i>Átsúlyozás</i>	- 13 -
3.2.3.	<i>Imputáció alapú eljárások</i>	- 14 -
3.2.4.	<i>Modell alapú eljárások</i>	- 15 -
4.	EMPIRIKUS KUTATÁS (VK, FZ)	- 16 -
4.1.	Adatbázis (VK)	- 16 -
4.2.	Hiányzó adatok kezelése (VK)	- 18 -
4.2.1.	<i>Adathiányt tartalmazó esetek törlése (listwise) (VK)</i>	- 20 -
4.2.2.	<i>Elérhető adatok elemzése (available case analysis) (VK)</i>	- 21 -
4.2.3.	<i>Átsúlyozás (FZ)</i>	- 22 -
4.2.4.	<i>Pótlás középértékkel (FZ)</i>	- 25 -
4.2.5.	<i>Hot deck imputáció (FZ)</i>	- 26 -
4.2.6.	<i>Többszörös imputáció Monte-Carlo szimulációval (VK, FZ)</i>	- 27 -
4.2.7.	<i>Modell a rejtőzködő és impulzus szavazókra (VK)</i>	- 29 -
4.2.8.	<i>Döntési (klasszifikációs) fa (FZ)</i>	- 33 -
5.	ÖSSZEFOGLALÁS (FZ)	- 35 -
6.	IRODALOMJEGYZÉK	- 37 -
7.	MELLÉKLETEK	- 38 -
7.1.	1. számú melléklet – A Társi által használt kérdőív	- 38 -
7.2.	2. számú melléklet – Kolosi-Tóth (2002) 354. oldal	- 40 -

TÁBLÁZATOK JEGYZÉKE

1. táblázat: Az első forduló előtti előrejelzések és mintanagyságaik	4 -
2. táblázat: Az előrejelzések eltérései a tényleges eredmény viszonylatában	5 -
3. táblázat: Az első forduló előtti és a 8 napos moratórium előrejelzéseinek összehasonlítása	17 -
4. táblázat: Iskolai végzettség kategorizálása	18 -
5. táblázat: A pártok kódolása a kérdőív alapján.....	18 -
6. táblázat: A hiányzó adatok mértéke a pártlistára adott válaszok alapján	19 -
7. táblázat: Pártlistákra adott válaszok relatív gyakorisága a listwise módszerrel	20 -
8. táblázat: Pártlisták relatív gyakorisága az elérhető adatok elemzése módszerrel.....	21 -
9. táblázat: A nem és <i>partlist</i> változók közötti kapcsolat szorossága és szignifikanciája	22 -
10. táblázat: A nem szerinti átsúlyozással kapott eredmények összesítése pártonként	23 -
11. táblázat: Az életkor és <i>partlist</i> változók közötti kapcsolat szorossága és szignifikanciája	23 -
12. táblázat: Az életkor szerinti átsúlyozással kapott eredmények összesítése pártonként	23 -
13. táblázat: A településtípus és <i>partlist</i> változók közötti kapcsolat szorossága és szignifikanciája	24 -
14. táblázat: A településtípus szerinti átsúlyozással kapott eredmények összesítése pártonként.....	24 -
15. táblázat: A végzettség és <i>partlist</i> változók közötti kapcsolat szorossága és szignifikanciája	24 -
16. táblázat: A végzettség szerinti átsúlyozással kapott eredmények összesítése pártonként.....	25 -
17. táblázat: A középértékkel (módusz) való pótlás eredményeinek összesítése	25 -
18. táblázat: A móduszok bemutatása a négy dimenzió mentén	26 -
19. táblázat: A hot deck imputációval kapott eredmények pártonként.....	27 -
20. táblázat: A Monte-Carlo szimulációval kapott eredmények bemutatása pártonként	28 -
21. táblázat: A rejtőzködők és impulzus szavazók becslésével kapott eredmények pártonként	32 -
22. táblázat: Döntési fa a demográfiai ismérvek mentén.....	34 -
23. táblázat: Az egyes módszerekkel kapott eredmények összegzése I.	35 -
24. táblázat: Az egyes módszerekkel kapott eredmények összegzése II.	35 -

1. BEVEZETÉS (VK)

A 2002-es magyar parlamenti választások még a politikusok számára is nagy meglepetést okoztak. A közvéleménykutató-cégek kivétel nélkül a Fidesz-MDF győzelmét jelezték előre - viszonylag nagy fölényrel - nem csak a 8 napos moratórium, még a választások napján végzett felmérések alapján is. A végeredmény éppen ellenkezőleg alakult, 1%-kal ugyanis az MSZP megelőzte nagy ellenfelét és ő alakíthatott kormányt. Ez a téves előrejelzés lesújtotta a közvéleménykutató-cégeket, akik visszatekintve a korábbi választásokra fokozatosan egyre nagyobb pontossággal tudták prognosztizálni a választási eredmények alakulását. A kis pártok esetében most sem tévedtek, éppen ezért is érdekes, hogyan történhetett, hogy valamennyien egy irányban jelezték előre a végeredményben ellenzékként a parlamentbe kerülő koalíció győzelmét. Neves kutatók számos tanulmányban keresték a választ erre a kérdésre és próbálták feltárni a probléma forrását, nagyjából hasonló eredménnyel. Ezek részben találgatások, részben pedig komolyan megalapozott, társadalomtudományi elméletek, melyeket nem lehet egzaktan, számszerűen bizonyítani, létezésük ugyanakkor cáfolhatatlan.

Dolgozatunk első részében ismertetjük a 2002-es országgyűlési választások első fordulója előtti előrejelzéseket, majd összevetjük a tényleges eredményekkel. Mivel a kettő eltérése nagy visszhangot váltott ki, és kihatott a közvélemény-kutatások megítélésére is, a társadalomelméleti szakirodalom áttekintését fontos kiindulópontnak tekintettük. Számos elmélet született ennek a tévedésnek a magyarázatára, ugyanakkor vizsgálatunkban inkább statisztikai eszközökkel próbálunk korrekciókat végezni, és az egyes módszerek bemutatása után a lehető legjobb közelítést adni a közvélemény-kutatás eredményeiből a választási eredményekre. Ezt megelőzően röviden bemutatjuk a lehetséges statisztikai módszereket az adathiányok kezelésére, melyeket igen releváns elemzési eszköznek tartunk a téma kapcsán.

Ezúton szeretnénk köszönetet mondani a Tárki Adatbanknak, hogy a rendelkezésünkre bocsátott adatbázis segítségével megvalósulhatott empirikus kutatásunk.

2. A 2002-ES PARLAMENTI VÁLASZTÁSOK (VK)

A közvélemény-kutatás történelmi hátterünk következtében nem tekint vissza hosszú múltra. A rendszerváltás óta eltelt 6 választási ciklus gyakran tartogatott meglepetéseket, a tapasztalatok azonban évről-évre beépültek az előrejelzések módszertanába és hozzájárultak a pontosabb előrejelzéshez. Kivételt képezett ez alól a 2002-es év, amikor megszakadt a korábbi évek tendenciája és minden korábbinál nagyobb hibaszázalékkal prognosztizálták az országgyűlési választások eredményeit a közvéleménykutató-cégek. [Karácsony-Lakatos (2006)]

A fokozódó versenyben egyre erőteljesebben reflektorfénybe kerültek az előrejelzők, főként a közvélemény növekvő érdeklődésének köszönhetően. Az 1990-es választásokhoz képest – melyeken még „csak” három intézet vállalkozott az előrejelzések elkészítésére – a 2002-es évben már öt, 2006-ban pedig hét intézet tette közzé eredményeit.

2.1. Előrejelzések és eredmények

A következő táblázat összefoglalja a tényleges, valamint az egyes közvéleménykutató-cégek által prognosztizált választási eredményeket a 2002-es évben. Az adatok az első forduló előtti felmérések eredményeit mutatják.

Intézet	MSZP	Fidesz-MDF	SZDSZ	MIÉP	Centrum	Minta
Tényleges eredmény	42,05	41,07	5,57	4,37	3,90	-
Gallup	38,50	46,50	6,50	3,50	3,50	2 000
Medián	42,00	44,00	5,00	4,00	2,00	3 000
Szonda Ipsos	40,00	45,00	4,50	4,50	3,00	15 000
Tárki	38,00	46,00	4,50	4,50	3,50	15 000

1. táblázat: Az első forduló előtti előrejelzések és mintanagyságaik

A 8 napos moratórium előtt valamennyi intézet kis mintákkal végezte a felméréseket, a fenti táblázatból azonban jól kivehető, hogy két intézet nagy, kettő pedig kis mintákkal becsülte az eredményeket az első fordulót megelőzően.

A kis mintáknál belátható, hogy kisebb a nem mintavételi hibából adódó tévedés lehetősége. Ezt részben indokolja példaként a kérdezőbiztosok problematikája. Valószínűsíthető, hogy minél magasabb arányú a munkaerő bevonás egy nagyobb minta eléréséhez, annál alacsonyabb a képzettség, fennáll a lehetősége annak, hogy a szignifikánsan növekvő költségek ellentételezésére kevésbé alkalmas munkaerőt alkalmaznak. Összességében úgy közelíthetnénk meg a problémát, hogy a mennyiség a minőség rovására megy, de ez csupán egyetlen összetevő.

Nézzük a következő összefoglaló táblát, melyben az eltéréseket és azok négyzetösszegeit foglaltuk össze, hasonlóképp a 2002-es választások kapcsán.

Intézet	MSZP	Fidesz-MDF	SZDSZ	MIÉP	Centrum	Eltérés négyzetösszeg
Tényleges eredmény	42,05	41,07	5,57	4,37	3,90	-
Gallup	- 3,55	5,43	0,93	- 0,87	- 0,40	43,87
Medián	- 0,05	2,93	- 0,57	- 0,37	- 1,90	12,66
Szonda Ipsos	- 2,05	3,93	- 1,07	0,13	- 0,90	21,62
Tárki	- 4,05	4,93	- 1,07	0,13	- 0,40	42,03

2. táblázat: Az előrejelzések eltérései a tényleges eredmény viszonylatában

Ami leginkább szembetűnő, hogy a két nagy párt esetében mindegyik intézet azonos irányban tévedett az eredményeket illetően, ezek azonban intézetenként jelentősen eltérnek egymástól. Nyilvánvaló, hogy mivel minden intézet más módszert alkalmaz kutatásai során, ez a különbség megmutatkozhat az eredmények vonatkozásában. A kisebb pártok esetében viszont nem szignifikáns az eltérés, valamennyi kutató cég hibahatáron belül prognosztizálta az eredményeket.

2.2. Hibák csoportosítása

Mi lehetett a hiba forrása? Ezt a kérdést tették fel a közvéleménykutató-cégek is maguknak, felgöngyölítve a potenciális tényezőket. Karácsony-Lakatos (2006) tanulmányában a lehetséges hibákat három csoportra bontja. Ennek megfelelően megkülönböztetünk mintavételi hibát, nem mintavételi hibát, valamint az elméleti problémák gyakorlati kezeléséből adódó hibalehetőséget.

A mintavételi hiba természetesen abból ered, hogy a felmérés nem teljeskörű. A nem mintavételi hiba magában foglalja a kódolási, kérdezői hibát, esetleges csalásokat, a nem válaszolókat, vagy nem őszintén válaszolókat miatti torzításokat. A harmadik kategória pedig azt jelenti, hogy a megismert véleményből mennyire tudunk következtetni a várható viselkedésre.

Nem lehet azonban figyelmen kívül hagyni, hogy a választási eredmények becslése mellett egy másik fontos tényező is kiemelt szerepet játszik: a választási részvételi arány. Hiszen míg a 2002-es évet megelőzően az egyes pártok eredményeit viszonylag jó közelítéssel tudták előrejelezni a közvéleménykutatók, a részvételi arány becslése igen pontatlan volt. Érdekes azonban, hogy 2002-ben erre az arányra egészen pontos becslést készítettek és ennek ellenére ilyen nagy mértékű tévedéssel jelezték előre a konkrét pártok eredményeit.

2.3. Az inkonzisztencia lehetséges okai

Valamennyi tanulmányban [Angelusz (2002), Kolosi-Tóth (2002), Karácsony-Lakatos (2006)], amely a témával foglalkozott, számos különböző elmélet került felsorakoztatásra arra vonatkozóan, hogy miért nem sikerült a hasonló évekhez viszonylag pontosan előrejelezni a választások kimenetelét. Ugyanakkor sem globális viszonylatban, sem pedig a társadalomtudományok terén egyáltalán nem számított újszerűnek ez a jelenség.

Angelusz (2002) sorra vette azon országokat, ahol ugyan különböző okokból és mértékben, de tévedtek a közvélemény-kutatók – főként a kormányt alakítók tekintetében. Ilyen eset volt az 1948-as amerikai elnökválasztás, ahol az előrejelzések kétséget kizáróan Thomas E. Dewey győzelmét mutatták, a szavazatok alapján azonban a harmincharmadik amerikai elnök a jól ismert Harry S. Truman lett. Hasonló szituáció alapult ki az 1970-es és 1992-es angol parlamenti választásokon is, melyeken a Munkáspárt előnyét prognosztizálták, ugyanakkor mindkét évben a konzervatívoké lett a győzelem. Az 1990-es nicaraguai választásokon pedig az aktuálisan kormányon lévő elnökjelölt fölényes győzelmét jelezték előre, ugyanakkor az ellenzéki jelölt majdnem akkora előnyre tett szert a tényleges adatok alapján, mint amekkorát eredetileg a kormányon lévő jelölt eredményeként prognosztizáltak.

Ahhoz, hogy megértsük milyen okok húzódnak a téves előrejelzések hátterében, elengedhetetlen, hogy segítségül hívjuk a szakirodalomban tárgyalt főbb jelenségeket. Miként a társadalomtudományhoz és közvéleményhez szorosan kapcsolódó téma a választások előrejelzése, valamennyi jelenség mögött az emberi viselkedés, annak sajátosságai húzódnak.

Azt már érintőlegesen tárgyaltuk, hogy milyen típusú módszertani hibák fordulhatnak elő, amelyek torzíthatják az előrejelzéseket, de a legtöbbhöz tudunk ellenérveket felsorakoztatni. Egyrészt az a tény, hogy minden intézet más mintavételi, interjúztatási módszerrel végzi az adatgyűjtést és ennek ellenére igen hasonló eredményre jutottak a 2002-es előrejelzések kapcsán, nagyon lényeges pont. (lásd 1-2. táblázat) Másrészt a korábbi évek tapasztalatai pozitívak voltak, az egyes pártok eredményeit jó közelítéssel jelezték előre annak ellenére is, hogy a részvételi arányt nem tudták egészen pontosan megbecsülni. Ehhez képest 2002-ben a részvételi arány viszonylag pontos becslése mellett az egyes pártok eredményeit jelezték előre tévesen. Más indokot kellett tehát keresni, hiszen a mintavételi eljárás, az adatgyűjtési módszerek a korábbi évekhez hasonlóan történtek, reprezentatív, véletlen mintán alapulva, esetlegesen a korábbi évek tapasztalatainak beépítésével.

Ebben a fejezetben a továbbiakban tehát viselkedéstudományi, pszichológiai szempontból foglalkozunk a torzítás lehetséges okaival. Nyilvánvaló, hogy ez a számszerűsített, egzaktan mérhető statisztikai adatokhoz képest sokkal nehezebben vizsgálható és bizonyítható. Tény azonban, hogy a szakirodalomban felsorakoztatott viselkedési elméletek létezése egy ilyen konkrét döntési szituációban, mint egy parlamenti választás, megkérdőjelezhetetlen. A legfontosabb kérdés ezzel kapcsolatban tehát az, hogy milyen mértékben van jelen és hogyan befolyásolhatja a választási előrejelzéseket, azok viszonyát a tényleges eredményekhez. Erre a kérdésre nehéz pontos választ adni. Hiszen az individuum várható viselkedését befolyásoló számtalan tényezőt figyelembe venni szinte lehetetlen, nemhogy azt kiterjeszteni egy egész sokaságra, még akkor is, ha valamely ismérvek mentén csoportokat képzünk. Az viszont nyilvánvaló és a tapasztalatok is egyértelműen azt mutatják, hogy foglalkozni kell a magatartásbeli hatásokkal, nem lehet figyelmen kívül hagyni őket, mert olyan meglepetésekkel szembesülhetünk, mint a korábban említett angol, amerikai vagy a jelen dolgozat középpontjában álló magyar választás.

A közvéleménykutató-cégeknek egy olyan felmérés elkészítése, mint egy parlamenti választás bizonyos szemszögből tekintve hálás feladat. Hiszen nem sok hasonló példát tudunk mondani, amelyben az előrejelzések közzététele után rövid időn belül, azokat a tényleges eredményekhez viszonyíthatjuk. Tehát nagyon gyors a visszacsatolási mechanizmus, amelynek során az intézetek pontos és valós képet kapnak arról, hogy mennyire sikerült módszereikkel jól közelíteni a választási eredményeket. A visszacsatolás során nyert információkból pedig tapasztalatokat szerezhetnek, fejlesztéseket eszközölhetnek, hogy javítsák előrejelzési modellüket.

2.3.1. Utánfutó effektus

Az utánfutó – másnéven „Bandwagon” - effektussal a szakirodalomban először Lazarsfeld és társai művében találkozhatunk. [Lazarsfeld et al. (1944), idézi Angelusz (2002)] A jelenség lényege, hogy a bizonytalan szavazók - akiknek nincsen konkrét pártkötődésük – nagyobb valószínűséggel arra teszik le voksukat, akiről úgy vélik, hogy meg fogja nyerni a választásokat, mivel jobban szeretnek a (potenciális) győztes táborhoz tartozni. Ez egy érdekes eset, hiszen legtöbbször a közvélemény az intézetek által folyamatosan közzétett népszerűségi, vagy éppen a választások előtti előrejelzésekből informálódik, amely közvetlenül, vagy közvetett csatornákon keresztül jut el a felhasználóhoz. Belátható, hogy az intézetekre milyen komoly felelősség hárul, hiszen az ő előrejelzésüktől függ ebben az esetben, hogy a bizonytalan szavazók végül melyik táborot fogják erősíteni. Továbbá a médiának is erős befolyása lehet, annak függvényében, hogy hogyan interpretálják a kutatási eredményeket.

Lazarsfeld (1944) művében számos vizsgálat és példa támasztja alá ennek a hatásnak a létezését, a későbbiekben pedig Noelle-Neumann (1980) -idézi Angelusz (2002) - ugyanezt az effektust észlelte és vizsgálta a 60-as 70-es években az NSZK-beli választások során. Azt azonban fontos leszögezni, hogy ez a hatás fordítva is megjelenhet, hiszen természetes, hogy minden egyén eltérő preferenciákkal rendelkezik. Főként azok esetében, akik a nem bizonytalan szavazók körébe tartoznak, valószínűsíthetően nem- vagy kevésbé fogja befolyásolni az előrejelzések eredménye a voksokat, másik szélsőséges esetben pedig pont ellentétesen is működhet a mechanizmus. Ez alatt azt értjük, hogy pontosan az előrejelzések készítetik az egyént a vesztesnek prognosztizált fél támogatására.

Fontos tehát, hogy óvatosan kezeljék ezeket a különböző hatásokat, ezt szemlélteti a következő idézet is Angelusz Róberttől (2002):

„A Bandwagon-hatás tehát nem feltétlenül érvényesül, s akik rendíthetetlenül bíznak működésében, könnyen kudarcot vallhatnak. Az utánfutó-hatás feltételezése és az ezzel járó magabiztosság ráadásul a pluralizmus ignoranciája következtében kifejezetten hibás kampányakciókat válthat ki.”

Ez az idézet szorosan kapcsolódott a szerző művében az 1970-es és 1992-es angol parlamenti választások kapcsán cáfolt utánfutó-hatáshoz.

2.3.2. Mobilizációs készség

A mobilizációs készség az utánfutó effektushoz hasonlóan kiemelt szerepet játszik egy parlamenti választás esetében. A mobilizációs készség – ahogyan nevéből is következik - lényege, hogy milyen mértékben tudják az egyes pártok mozgósítani potenciális szavazóikat. Ezzel szoros összefüggésben áll az előbbi fejezetben már említett kvázi-ellentéte az utánfutó-hatásnak, melyet underdog-effectnek hívnak. Hiszen arra is épülhet a politikai kampány stratégiája, hogy a vesztesnek látszó pozícióval motiválja a választót a szavazásra.

A másik fél részéről ez azonban komoly veszélyekkel járhat. A túlzott magabiztosság és erőfölényfitogtatás - amellyel potenciálisan az utánfutó-hatásra apellál egy párt – a mobilizációs készség szempontjából éppen ellentétes hatást válthat ki. Magyarán inaktívvá teheti az adott párt biztos választóinak egy részét, akik úgy vélik felesleges az urnához járulni, mert nélkülük is győzelmet arat majd preferált pártjuk, ez pedig nagy meglepetéseket okozhat egy hibás kampánystratégia megválasztásánál.

Angelusz (2002) tanulmányában a mobilizációs készség kapcsán azt írja, elsősorban az alacsonyabb státusú választók mobilizációjára kell fókuszálni, mert jellemzően ebben a csoportban alacsonyabb a részvételi hajlandóság. Ezzel azonban úgy véljük, szemben áll az a politikai underclass elmélet, mely szerint van egy olyan rétege a társadalomnak, amely sem pártpreferenciával, sem pedig szavazási hajlandósággal nem rendelkezik és a tendenciák alapján a későbbiekben sem fog [Angelusz (2006)]. Abban az esetben, ha a két csoport között átfedés van, akkor fokozott körültekintéssel kell a mobilizációs fókusz csoportjának kiválasztását kezelni.

2.3.3. *Rejtőzködők – látens közvélemény*

Az egyik legkomolyabb probléma és legnehezebben mérhető viselkedési morál, amely torzíthatja az előrejelzéseket, az az ún. látens közvélemény. Szinte lehetetlen kiküszöbölni azt a jelenséget, amikor az egyén magán- és nyilvános véleménye valamely okból kifolyólag eltér. Elsőként feltehetnénk a kérdést, hogy miért létezik egy ilyen effektus egy demokratikus rendszerben, ahol szabad véleménynyilvánítás van, mindenféle szankciók nélkül. A válasz keresése kapcsán természetesen visszajutunk a viselkedéstudományok ingoványos területére. Hazánk esetében az egyik legerősebb befolyásoló tényező kétségkívül a történelmi eseményekre, korszakokra vezethető vissza. Hiszen a választásra jogosultak még mindig magas aránya olyan generációk, melyeknél a rendszerváltás előtti időszak hatásainak társadalmi beágyazottsága figyelhető meg, melynek példája a pártpreferenciák eltitkolása az esetleges szankcióktól, büntetésektől való félelem miatt. Jellemző továbbá az ún. shame-faktor, amikor az adatfelvétel során a válaszadó magánvéleményét szégyelli, ezért nevez meg más – az ő megítélése szerint népszerűbb, vagy elfogadottabb – pártot, mint saját preferenciáját. [Angelusz (2002), Karácsony-Lakatos (2006)]. Emellett hasonló torzító tényezőnek bizonyulhat a választásnál a potenciális szavazó megítélése az adott kutatóintézet politikai beállítottságát illetően [Kolosi-Tóth (2002)]. Amennyiben úgy véli a válaszadó, hogy az intézet politikailag nem független, előfordulhat, hogy az intézettől eltérő pártpreferencia esetén inkább megtagadja a választást, vagy az udvariasabb „nem tudom” kategóriát választja, legrosszabb esetben pedig azt a pártot jelöli meg, amelyet úgy vél, hogy az intézet is preferál.

A tanulmányok nagy része főként ezt a látens közvéleményt jelölte meg a téves előrejelzések okozójának. Ez a társadalomtudományokban általánosan elfogadott jelenség azonban, mint korábban is említettük, nehezen bizonyítható és mérhető. Létezése ugyanakkor mégis kevésbé cáfolható, hiszen a különböző kutatóintézetek szisztematikus, egyirányú téves előrejelzését egy ilyen jelenség - a releváns érvekkel alátámasztva - jól magyarázza. Kutatásunkban ennek ellenére mégis megpróbálkozunk annak bizonyításával, hogy nem „csupán” ez a tényező okozta a torzulást, hanem a 2002-es választásokon minden korábbinál magasabb nemválaszolási arány figyelmen kívül hagyása, melynek kezelése esetén pontosabb előrejelzés készülhetett volna.

3. HIÁNYZÓ ADATOK STATISZTIKAI KEZELÉSE (FZ)

3.1. Hiányzó adatok típusai

A statisztikai adatbázis úgy épül fel, hogy az adatok egy $n \times k$ – s mátrixba vannak rendezve ($\mathbf{X}$), mégpedig úgy, hogy az n darab sor különböző megfigyelési egységeket (sokasági- vagy mintaelemeket), a k különböző oszlop pedig különböző változókat jelent. Vagyis a mátrix egy eleme, X_{ij} jelenti azt az értéket, amit az i -edik megfigyelés a j -edik változóban felvesz. Ha nincs érték hozzárendelve valamelyik X_{ij} elemhez, akkor adathiány lépett fel.

Rudas (2006) alapján a hiányzó adatok csoportosíthatók aszerint, hogy mennyire kiterjedt a hiány:

- Coverage error (lefedési hiba): a sokaságról rendelkezésünkre álló lista nem teljes, tehát egyes elemek mintába kerülési valószínűsége nulla¹.
- No contact error: a kiválasztott megfigyelési egységgel nem sikerült kapcsolatba lépni (nem nyitnak ajtót, nem veszik fel a telefont, stb.).
- Unit non-response (egységszintű nemválaszolás): a kiválasztott megfigyelési egységgel sikerült kapcsolatba lépni, de minden kérdésre kiterjedően megtagadja a válaszadást.
- Item non-response (elemszintű nemválaszolás): a kiválasztott megfigyelési egységgel sikerült kapcsolatba lépni, és a válaszmegtagadás a kérdéseknek csak egy körére korlátozódik.

¹ Bár a dolgozatban általában részleges, mintavételes megfigyelésre utalunk, megjegyezzük, hogy ez a fajta hiba teljes körű adatfelvételek, például népszámlálások esetén is felmerül: ún. post-enumeration survey-k során azt a kérdést tették fel, hogy „Önt lekérdezték-e a népszámláláson?”, és az eredményekből az derült ki, hogy 2-3%-a a népességnek nem lett megszámmlálva.

Az adathiány mintázata alapján történő csoportosítás azzal foglalkozik, hogy mely adatok megfigyeltek és melyek nem. Eszerint lehet:

- egyváltozós: csak egyetlen változó esetében fordul elő adathiány
- többváltozós: több változó is adathiányos, de úgy, hogy az egyes megfigyelési egységekhez vagy minden adathiányos változónál hiányzik az adat, vagy egyiknél sem
- általános: bizonyos kérdésekre bizonyos egységek esetében hiányzik az adat, de szabályszerűség nem mutatkozik
- monoton: ha az egyik egységnél egy adat hiányzik, akkor az utána következő minden változóhoz tartozó érték hiányzik
- látens változó: a változó nem megfigyelhető, tehát minden megfigyelési egységnél hiányzik az értéke

Little – Rubin (2002) az alapján csoportosítja az adathiányokat, hogy miért lettek hiányzóak: ezt adathiány-mechanizmusnak nevezik. A hiányzó és a meglévő adatok közti kapcsolat szerint tehát lehet:

- teljesen véletlenszerű adathiány (MCAR: missing completely at random): a hiányzás ténye független minden (megfigyelt és hiányos) változóértéktől.
- véletlenszerű adathiány (MAR: missing at random): a hiányzás ténye független a hiányos változó szerint felvett értéktől, de nem független valamelyik megfigyelt változó szerinti értéktől.
- nem véletlenszerű adathiány (NMAR: not missing at random): a hiányzás ténye nem független a hiányos változótól, viszont független a megfigyelt változóktól.

Mivel adathiány a legkülönbözőbb szándékos és véletlenszerű okok miatt felmerülhet (a mintakeret nem naprakész, a válaszadó nincs otthon a nap egy bizonyos szakában, bizalmatlan a kérdezőbiztossal szemben, nem akar, vagy nem tud válaszolni), általában elő is fordul. A jelenség nem egyedi, így kezelésre szorul. A hiányzó adatok kezelésére szolgáló módszereket a teljesség igénye nélkül mutatjuk be [Little – Rubin (2002) alapján].

3.2. Hiányzó adatok kezelése

3.2.1. Teljesen megfigyelt, vagy elérhető egységek elemzésén alapuló eljárások

(a) adathiányt tartalmazó esetek törlése (listwise adattörlés)

A hiányzó adatokat tartalmazó sorokat töröljük az egész adatbázisból, tehát nem használjuk fel azokat az információkat, amelyek olyan megfigyelési egységektől származnak, akik legalább egyszer nem produkáltak adatot. Előnye egyszerűsége és az adatok összehasonlíthatósága, hátránya viszont, hogy figyelmen kívül hagy meglévő információkat. Ha MCAR-típusú adathiány lépett fel, akkor jól használható, más esetekben azonban torzít. Viszont még ekkor is érdemes lehet kihasználni az egyszerűségét, ha alacsony a nemválaszolási arány.

(b) elérhető adatok elemzése (available case analysis)

A hiányzó adatokat tartalmazó megfigyeléseket nem töröljük: minden elemzéshez az összes választ felhasználjuk. Ekkor az adathiányról gyakorlatilag nem veszünk tudomást, de kiküszöböljük az információk elhagyását. Hátránya továbbá, hogy így minden elemzés tulajdonképpen más-más adatbázison készül, tehát nem összehasonlíthatóak az eredmények.

3.2.2. Átsúlyozás

A módszer alkalmazásához először a megfigyelési egységeket valamely változó(k) szerint utólag rétegezzük. Feltéve, hogy minden rétegben van olyan elem, amelynél nincs adathiány, a rétegeken belül a válaszolók válaszaiból következtetünk a nemválaszolók válaszaira. Ezt úgy tesszük meg, hogy a válaszolók válaszhoz egy súlyt rendelünk, melyet a rétegen belüli válaszolási arány reciprokaként kaphatunk. Minthogy az így kapott súly minimuma 1, és annál nagyobb, minél magasabb volt a nemválaszolási arány a rétegen belül, ezzel a módszerrel azt fejezzük ki, hogy nem minden válasz egyformán fontos: annak a rétegnek az elemei, amelyben relatíve kevesen válaszoltak, nagyobb szerepet kapnak a válaszadóktól származó adatok, mint egy olyan réteg esetén, ahol sokan válaszoltak.

Ezzel persze azt is feltételezzük, hogy a rétegen belül az adathiány MCAR-típusú volt, tehát a nemválaszolók is ugyanazt mondták volna, mint a válaszolók.

3.2.3. Imputáció alapú eljárások

Az imputáció elve az, hogy kiegészítjük az adatbázist, mesterségesen, utólag pótoljuk a hiányzó adatokat valamilyen megfelelő adattal. Az imputáció utáni elemzések közben azonban mindig figyelembe kell venni, hogy az imputált értékek befolyásolják a becslések tulajdonságait.

(a) logikai imputáció

Az imputáció legegyszerűbb módja, ha logikai alapon nagy biztonsággal tudunk következtetni a hiányzó adatra egy korábbi adatfelvétel alapján. Ez tehát leginkább paneladatok esetén használható.

(b) pótlás középértékekkel

A hiányzó adatokat vagy az egész adatbázis, vagy egy rétegzés utáni al minta meglévő adataiból számított középértékkel (számítási átlag, módusz, medián) pótoljuk. Ez a módszer egyszerű, de az adatok változatosságát alulbecsli².

(c) regressziós módszerek

Azt a változót, amelyben adathiány lépett fel, egy regressziós modell eredményváltozójává tesszük, és az adathiány nélkülieken mint magyarázó változókon regresszáljuk. Az így kapott regressziós egyenlet segítségével becsült értékeket számítunk, és a hiányzó adatokat ezekkel pótoljuk.

(d) hot deck imputáció

² Például ha az átlaggal pótlunk, a várható érték torzítatlan marad, viszont a szórás valószínűleg kisebb lesz, mint a sokasági szórás.

A teljes egységek között keresünk olyat, amely nagyon hasonlít a hiányos egységhez, és az ő válaszát imputáljuk a hiányzó adat helyére. Maga az eljárás nem nehéz, viszont problémás, hogyan definiáljuk a hasonlóságot két megfigyelés között. Ebbe a körbe tartozik az is, ha nem egyetlen értéket, hanem több hasonló egység valamilyen középértékét, vagy közülük véletlenszerűen egyet helyettesítenek be.

A hot deck egyik népszerű típusa a közelítő bayesi bootstrap (Approximate Bayesian Bootstrap). Logisztikus regressziót építünk, melynek eredményváltozója az adathiány egy tetszőleges hiányos változóban, magyarázó változói pedig a többi változó. Ezzel megbecsüljük a válaszolás odds-át minden hiányos megfigyelésre, az odds alapján csoportokba rendezzük őket, végül az ily módon kialakított csoportokra valószínűsítjük meg az imputálást (véletlen mintavétellel).

(e) többszörös imputáció (multiple imputation)

A hiányzó adatok helyére nem egyetlen adatot imputálunk, hanem többet (jellemzően 3-10 különbözőt), így több teljes adatbázist készítünk, melyekre külön-külön elvégezzük az elemzéseket, majd összegezzük az eredményeket. Gyakran időigényes és bonyolult eljárás, de kevésbé esetleges eredményeket ad.

3.2.4. Modell alapú eljárások

(a) várakozásmaximalizáló (expectation maximization)

A meglévő adatok alapján maximum likelihood becslést készítünk, és az így becsült értékeket helyettesítjük a hiányzó adatok helyére. A paramétereket újrabecsljük, és addig iterálunk, amíg a paraméterek nem konvergálnak.

(b) modellek nem véletlenszerű adathiány kezelésére

A hiányzó adatokban mintázatokkal különítünk el a teljes változókból, és ezt mint magyarázó erőként építik be a modellbe.

4. EMPIRIKUS KUTATÁS (VK, FZ)

4.1. Adatbázis (VK)

Empirikus kutatásunk alapja a Tárki adatbank által rendelkezésünkre bocsátott, az első forduló előtti felmérések eredményét összefoglaló adatbázis, melynek pontos címe: *„A felnőtt lakosság választói preferenciái a 2002-es magyar országgyűlési választások első fordulójára előtt”*.

- Az adatgyűjtemény által lefedett időszak: 2002.04.01.-2002.04.30.
- Az adatgyűjtés ideje: 2002.04.04.-2002.04.06.
- A választások eredménye: 2002.04.07.

Az adatok földrajzi hatókörét tekintve Magyarországra terjednek ki, a gyűjtemény survey típusú, időbeli dimenzióját tekintve keresztmetszeti. Az adatgyűjtés kérdezőbiztosok által, kérdőíves adatfelvétel alapján történt. Az adatbázis tartalmazza az adatfelvétel évét, hónapját és napját, valamint azt, hogy melyik megyében történt a lekérdezés. Ezután következnek a kérdőíven szereplő kérdések és válaszlehetőségeinek kódolt eredményei.³

A mintavételi eljárást tekintve többlépcsős, rétegzett valószínűségi véletlen mintáról beszélhetünk. Először megye és település szerinti rétegzést végeztek, majd a kiválasztott településen a véletlen sétás kérdéshez kiindulópontot választottak és minden negyedik háztartásban végeztek adatgyűjtést. Olyan személy került kiválasztásra, aki általában a legkevesebbet tartózkodik otthon. A valamivel több mint 15.000 fős reprezentatív minta két (egy 10.000-es és egy 5.000-es) önmagában is reprezentatív mintából tevődik össze.⁴

Az adatbázisban található elemszám 15.256, a változók száma 38.

³ A kérdőív megtalálható az 1.számú mellékletben.

⁴ http://www.tarki.hu/cgi-bin/katalogus/tarkifo_hun.pl?sorszam=TDATA-F30

Érdekesség, hogy a 8 napos moratórium előtti előrejelzés kis minta (kb 1.500 fő) alapján készült, míg a választásokat megelőző két nap alatt 15.000 lekérdezés rögzítésével próbálták előrejelezni az eredményeket. Azt már korábban tárgyaltuk, hogy nagy minták esetében nagyobb lehet a nem mintavételi hibából eredő torzítás. A két időpontban készült felmérések összevetésével láthatjuk, hogy valóban nagyobb eltérés-négyzetösszeggel⁵ (SSE) becsülték a nagyobb mintával az eredményeket, bár ennek mértéke nem jelentős.

Tárki	8 napos	első forduló
MSZP	- 4,25	- 4,05
Fidesz-MDF	4,73	4,93
SZDSZ	0,33	- 1,07
MIÉP	- 0,67	0,13
Centrum	- 0,20	- 0,40
Minta	1 519	15 000
SSE	41,03	42,03

3. táblázat: Az első forduló előtti és a 8 napos moratórium előrejelzéseinek összehasonlítása

Az biztosan állítható, hogy két nap alatt 15.000 lekérdezést véghez vinni, majd egy nap alatt feldolgozni az adatokat, elvégezni az elemzéseket - mivel a választások előtti két napban végezték a felmérést – óriási kihívás és magas munkaerőigénnyel, költséggel jár. A fenti összesítő táblázat alapján felvetődik a kérdés, hogy megéri-e egyáltalán, hiszen nem tudták a nagyobb mintával jobban előrejelezni az eredményeket. A legnagyobb probléma az előrejelzések kapcsán azonban nyilvánvalóan nem a mintanagyságban vagy a mintavételi eljárásban rejlett, hanem a minden korábbinál magasabb nemválaszolási arányban, amivel nem foglalkoztak az elemzések során. Erről a Tárki a következőképp nyilatkozott [Kolosi-Tóth (2002)]:

„Miért nem tűnt fel, hogy a bizonytalanok és válaszmegtagadók csoportja igen nagy, és közvetlenül választások előtt növekszik? Erre a kérdésre a legkönnyebb válaszolni: feltűnt, de mivel ez korábban is így volt, nem tulajdonítottunk neki jelentőséget. Azt feltételeztük ugyanis, hogy a bizonytalanok és a válaszmegtagadók jelentős része úgyszemint fog elmenni a választásokra. Emellett szólt például a legutóbbi választás tapasztalata, hiszen akkor így történt.”

⁵ A becslés „jóságát” az SSE-vel jelölt, a becsült és a tényleges értékek eltéréseinek négyzetösszegét jelentő mutatóval mérjük, és egy becslés akkor jobb a másikonál, ha az SSE-je kisebb.

Mielőtt azonban rátérnénk a nemválaszolók - vagyis a hiányzó adatoknak - kezelésére, bemutatjuk milyen korrekciókat végeztünk el az adatbázison. Első lépésben kitöröltük azt a személyt, aki az adatbázisban szereplő születési év alapján a választások évében még csak 16 éves volt. Besoroltuk azokat a megkérdezetteket, akiknek szerepelt a születési dátuma az adatbázisban, de valamilyen oknál fogva nem lett átvezetve a megfelelő korcsoport kategóriába, megszüntetve ezáltal a rendszer által hiányzónak vélt adatok egy bizonyos részét. Ezen felül az iskolai végzettség tekintetében azzal az egyszerűsítéssel éltünk, hogy a 6 ismérvértéket (lásd 4. táblázat) összevontuk és 3 ismérvérték mentén kategorizáltuk be.

Tárki szerinti bontás	Új kategóriák
kevesebb, mint 8 általános 8 általános szakmunkásképző	alap
érettségi	közép
főiskolai diploma/technikus egyetemi diploma	felső

4. táblázat: Iskolai végzettség kategorizálása

4.2. Hiányzó adatok kezelése (VK)

Ahhoz, hogy megállapítsuk mekkora a hiányzó adatok mértéke az adatbázisban, lefuttatunk az SPSS-ben egy missing value analysis-t. Az elvégzett korrekciók után kialakított új adatbázisból első lépésként kiválasztottuk azt a változót, amelyik elemzésünk fókusza lesz. Ez a „partlist” változó, amely a kérdőív 2. pontjában szerepel (lásd 1.számú melléklet), vagyis az a kérdés, hogy a válaszadó melyik párt listájára fog szavazni. A válaszlehetőségeket és a kódolást a következő táblázat foglalja össze.

Kód	Válaszlehetőség	Kód	Válaszlehetőség
01	Centrum Párt	09	Munkáspárt
02	Fidesz-MDF	10	Reform Kisgazdapárt
03	FKgP	11	SZDSZ
04	Kisgazdapárt	12	Szociáldemokrata Párt
05	Magyar Vállalkozó Egységpárt	13	Új Baloldal
06	Magyarországi Roma Párt	77	nem megy el szavazni
07	MIÉP	88	megtagadja a választást
08	MSZP	99	nem tudja

5. táblázat: A pártok kódolása a kérdőív alapján

A hiányzó adatok mértékét a következő táblázatok mutatják:

Univariate Statistics			
	N	Missing	
		Count	Percent
<i>partlist</i>	9833	5422	35,5

6. táblázat: A hiányzó adatok mértéke a pártlistára adott válaszok alapján

Megállapíthatjuk ezek alapján, hogy a hiányzó adatok aránya jelentős, a *partlist* változó esetében 35,5%. Ez magában foglalja a fenti táblában található „nem megy el szavazni”, „megtagadja a választást” és a „nem tudja” kategóriákat.

Felvetődik természetesen a kérdés, hogy miért nem vettük ki azokat, akik a „nem megy el szavazni” lehetőséget jelölték meg. A válasz nem igényel hosszú tudományos fejtegetést, abból az egyszerű feltevésből indultunk ki, hogy az, aki azt válaszolta, hogy „nem megy el szavazni”, egyrészt egyáltalán nem biztos, hogy tényleg nem fog elmenni, másrészt előfordulhat, hogy ezt tekintette udvariasabb formának a válaszmegtagadás helyett.

Összességében tehát megállapítottuk, hogy a 35,5%-os „missing” nagyon jelentős, mindenképpen kezelni kell. A továbbiakban erre teszünk kísérletet a korábbiakban bemutatott módszerek közül azok segítségével, amelyek alkalmazhatók jelen adatbázis hiányzó adatainak korrigálására és új előrejelzések készítésére.

Fontos megemlíteni, hogy mivel minden változónk minőségi ismérv, vannak módszerek, melyek nem alkalmazhatók az adathiány kezelésre.

4.2.1. Adathiányt tartalmazó esetek törlése (listwise)(VK)

Magát a módszert már ismertettük a második fejezetben, így ennek tárgyasára most nem kerül sor, csupán az elvégzett műveleteket ismertetjük és az elemzés végeredményét. A listwise módszernek megfelelően töröltünk minden olyan sort az adatbázisból, amelyben legalább egy változó kapcsán adathiány lépett fel.

A művelet végrehajtása után a 7. táblázatban lévő gyakoriságokat kaptuk, pártonként leválogatva. Várható volt – a modell egyszerűségét tekintve -, hogy az eredmények nagy valószínűséggel még inkább torzítottak lesznek, ettől függetlenül meg kellett bizonyosodnunk róla, hiszen akár pontosabb előrejelzés is születhetett volna. Ezúttal nem vezetett használható eredményre ez a módszer.

		<i>partlist</i>			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Centrum Párt	213	3,3	3,3	3,3
	Fidesz-MDF	3127	48,1	48,1	51,4
	FKgP	56	,9	,9	52,3
	Kisgazdapárt	7	,1	,1	52,4
	Magyar Vállalkozó	4	,1	,1	52,4
	Egységpárt				
	Magyarországi Roma Párt	7	,1	,1	52,5
	MIÉP	255	3,9	3,9	56,5
	MSZP	2368	36,4	36,4	92,9
	Munkáspárt	114	1,8	1,8	94,7
	SZDSZ	343	5,3	5,3	100,0
	Szociáldemokrata Párt	2	,0	,0	100,0
	Új Baloldal	1	,0	,0	100,0
	Total	6497	100,0	100,0	

7. táblázat: Pártlistákra adott válaszok relatív gyakorisága a listwise módszerrel

4.2.2. Elérhető adatok elemzése (available case analysis) (VK)

Az available case analysis esetén nem töröltük semmit az adatbázisból, figyelmen kívül hagytuk a hiányzó adatokat. Relatív gyakoriságokat számoltunk, melyeket a 8. táblázat foglal össze. Jól látható, hogy figyelmen kívül hagyva a hiányzó adatokat szignifikánsan eltérő adatokat kapunk a választási eredményekhez képest, ugyanakkor a „valid percent” oszlop sok esetben jó közelítéssel, néhány esetben pedig szinte egy tizedesjegy pontossággal mutatja a Társi által publikált eredményeket. Ez az oszlop annyit jelent, hogy a relatív gyakoriságokat csak a tényleges válaszadók számához viszonyítva kalkuláltuk.

		<i>partlist</i>			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Centrum Párt	390	2,6	4,0	4,0
	Fidesz-MDF	4510	29,6	45,9	49,8
	FKgP	111	,7	1,1	51,0
	Kisgazdapárt	12	,1	,1	51,1
	Magyar Vállalkozó	6	,0	,1	51,1
	Egységpárt				
	Magyarországi Roma Párt	15	,1	,2	51,3
	MIÉP	369	2,4	3,8	55,0
	MSZP	3644	23,9	37,1	92,1
	Munkáspárt	199	1,3	2,0	94,1
	Reform Kisgazdapárt	1	,0	,0	94,1
	SZDSZ	570	3,7	5,8	99,9
	Szociáldemokrata Párt	4	,0	,0	100,0
	Új Baloldal	2	,0	,0	100,0
	Total	9833	64,5	100,0	
Missing	nem megy el szavazni	634	4,2		
	megtagadja a választ	2145	14,1		
	nem tudja	2611	17,1		
	System	32	,2		
	Total	5422	35,5		
Total		15255	100,0		

8. táblázat: Pártlisták relatív gyakorisága az elérhető adatok elemzése módszerrel

4.2.3. Átsúlyozás (FZ)

Az átsúlyozásnak akkor van értelme, ha bizonyíthatóan van kapcsolat a pártpreferencia és valamely vizsgált demográfiai ismerv között. A rendelkezésünkre álló demográfiai ismérvek a nem, a kor, a településtípus és az iskolai végzettség.

A válaszadó neme és a listás szavazás között szignifikáns, de nagyon gyenge kapcsolat áll fenn (lásd 9. táblázat).

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	,091	,000
	Cramer's V	,091	,000
	Contingency Coefficient	,091	,000
N of Valid Cases		15215	

9. táblázat: A nem és *partlist* változók közötti kapcsolat szorossága és szignifikanciája

A kapcsolat szignifikáns volta miatt az átsúlyozást elvégeztük, tehát először a nem ismérve szerint rétegeztük a mintát, majd megfigyeltük a válaszolási arányt⁶ a két így képzett csoportban. Ez az arány a férfiak esetében 66,0%, míg a nőknél 63,1% volt, így a férfiak válaszai 1,515-szörös, a nőké pedig 1,585-szörös szorzóval számítanak. Magyarán, mivel a nők között nagyobb volt a véleményhiány, egy nő válasza a becslésben nagyobb szerepet kap, mint egy férfié, akitől könnyebb volt megtudni a véleményét. Az ily módon kapott eredményeket a 10. táblázat tartalmazza.

A becslés hibája a Tárki becslésénél nagyobb lett, de a kis pártokat alapvetően jól jelezte előre a nem szerinti átsúlyozás módszere.

⁶ Az átsúlyozás szempontjából nemválaszolásnak minősítettük a „nem megy el szavazni”, a „megtagadja a választ”, a „nem tudja” eseteket, valamint az adatrögzítésből kimaradt eseteket.

<i>partlist</i>	Átsúlyozás (nem)	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (átsúlyozás)
Centrum	3,96%	3,50%	3,90%	0,40	-0,06
Fidesz-MDF	45,92%	46,00%	41,07%	-4,93	-4,85
MIÉP	3,72%	4,50%	4,37%	-0,13	0,65
MSZP	37,03%	38,00%	42,05%	4,05	5,02
SZDSZ	5,81%	4,50%	5,57%	1,07	-0,24
			SSE	42,03	49,21

10. táblázat: A nem szerinti átsúlyozással kapott eredmények összesítése páronként

Az életkor hatása szintén szignifikáns, és a kapcsolat szorosságát mérő mutató a nemnél erősebb kapcsolatot jelez. (11. táblázat)

Symmetric Measures

	Value	Approx. Sig.
Nominal by Nominal Phi	,166	,000
Cramer's V	,117	,000
Contingency Coefficient	,163	,000
N of Valid Cases	15188	

11. táblázat: Az életkor és *partlist* változók közötti kapcsolat szorossága és szignifikanciája

A válaszolási hajlandóság a fiatalok között volt a legmagasabb (67,75%), a középkorúak esetében valamivel kevesebb (65,16%), az idősek viszont határozottan kisebb arányban nyilatkoztak a pártpreferenciájukról (58,64%), tehát az ő válaszaik szerepelnek a legnagyobb, 1,7-szeres súllyal. Ennek eredménye (12. táblázat) egy kis korrigálódás, elmozdulás a Fidesz-szavazatokban lefelé, az MSZP-szavazatokban pedig felfelé, de érdemben nem változtatott az előrejelzésen.

<i>partlist</i>	Átsúlyozás (korkat)	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (átsúlyozás)
Centrum	3,93%	3,50%	3,90%	-0,40	0,03
Fidesz-MDF	45,69%	46,00%	41,07%	4,93	4,62
MIÉP	3,72%	4,50%	4,37%	0,13	-0,65
MSZP	37,31%	38,00%	42,05%	-4,05	-4,74
SZDSZ	5,72%	4,50%	5,57%	-1,07	0,15
			SSE	42,03	44,26

12. táblázat: Az életkor szerinti átsúlyozással kapott eredmények összesítése páronként

A településtípusnak is szignifikáns, de a korhoz hasonló mértékűen gyenge kapcsolata van az eredményváltozókkal (13. táblázat). A válaszolási arány Budapesten 70,2%, megyeszékhelyeken 60,8%, egyéb városokban 64,1%, községekben pedig 63,8% volt, vagyis a megyeszékhelyeken válaszolók kapták a legnagyobb súlyokat (1,65-szörös súllyal számítottak). A 14. táblázatból látható, hogy a becslés így sem javult.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,158	,000
	Cramer's V	,091	,000
	Contingency Coefficient	,157	,000
N of Valid Cases		15223	

13. táblázat: A településtípus és *partlist* változók közötti kapcsolat szorossága és szignifikanciája

<i>partlist</i>	Átsúlyozás (teltip)	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (átsúlyozás)
Centrum	3,93%	3,50%	3,90%	-0,40	0,03
Fidesz-MDF	46,03%	46,00%	41,07%	4,93	4,96
MIÉP	3,70%	4,50%	4,37%	0,13	-0,67
MSZP	37,04%	38,00%	42,05%	-4,05	-5,01
SZDSZ	5,72%	4,50%	5,57%	-1,07	0,15
			SSE	42,03	50,21

14. táblázat: A településtípus szerinti átsúlyozással kapott eredmények összesítése páronként

A negyedik ismerv a végzettség, mely az előzőeknél kissé erősebb kapcsolatban áll a listás szavazatra vonatkozó válasszal (15. táblázat). Az ezen ismerv szerinti átsúlyozás szintén nem javít a Tárki becslésén (16. táblázat).

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,173	,000
	Cramer's V	,123	,000
	Contingency Coefficient	,171	,000
N of Valid Cases		15184	

15. táblázat: A végzettség és *partlist* változók közötti kapcsolat szorossága és szignifikanciája

<i>partlist</i>	Átsúlyozás (végzettség)	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (átsúlyozás)
Centrum	3,85%	3,50%	3,90%	-0,40	-0,05
Fidesz-MDF	45,96%	46,00%	41,07%	4,93	4,89
MIÉP	3,68%	4,50%	4,37%	0,13	-0,69
MSZP	37,17%	38,00%	42,05%	-4,05	-4,88
SZDSZ	5,71%	4,50%	5,57%	-1,07	0,14
			SSE	42,03	48,25

16. táblázat: A végzettség szerinti átsúlyozással kapott eredmények összesítése páronként

4.2.4. Pótlás középértékkel (FZ)

Mint ahogy a négy demográfiai ismerv szignifikáns, de gyenge kapcsolatot mutatott a listás szavazással, az elemzés következő részében egyszerre vettük figyelembe mindegyiket, vagyis nem-kor-végzettség-településtípus dimenziók szerint rétegeztünk.

Ezt követően először a módusszal, mint jellemző (és nominális mérési skálájú eredményváltozónk esetén használható) középértékkel való pótlást végeztük el. Előzetes várakozásunk az volt, hogy ez a módszer a Fidesz-MDF javára fog további torzulást okozni, mert ez a párt lesz minden rétegben a módusz.

A módusszal való pótlás során a 17. táblázatban lévő eredményeket kaptuk.

<i>partlist</i>	Pótlás módusszal	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (pótlással)
Centrum	2,56%	3,50%	3,90%	-0,40	-1,34
Fidesz-MDF	54,89%	46,00%	41,07%	4,93	13,82
MIÉP	2,40%	4,50%	4,37%	0,13	-1,97
MSZP	34,10%	38,00%	42,05%	-4,05	-7,95
SZDSZ	3,75%	4,50%	5,57%	-1,07	-1,82
			SSE	42,03	263,31

17. táblázat: A középértékkel (módusz) való pótlás eredményeinek összesítése

Minden réteg hiányzó értékeit a módusszal pótoltuk, ahol pedig két módusz is volt, ott a hiányzó értékeket 50-50%-ban megosztottuk közöttük. A rétegenkénti móduszokat a 18. táblázat tartalmazza.

				Módusz					Módusz
férfi	fiatal	alapfokú	község	Fidesz-MDF	nő	fiatal	alapfokú	község	Fidesz-MDF
			város	Fidesz-MDF				város	Fidesz-MDF
			megyeszékhely	Fidesz-MDF				megyeszékhely	Fidesz-MDF
			főváros	Fidesz-MDF				főváros	Fidesz-MDF
		középfokú	község	Fidesz-MDF			középfokú	község	Fidesz-MDF
			város	Fidesz-MDF				város	Fidesz-MDF
			megyeszékhely	Fidesz-MDF				megyeszékhely	Fidesz-MDF
			főváros	Fidesz-MDF				főváros	Fidesz-MDF
		felsőfokú	község	Fidesz-MDF			felsőfokú	község	Fidesz-MDF
			város	Fidesz-MDF				város	Fidesz-MDF
			megyeszékhely	Fidesz-MDF				megyeszékhely	Fidesz-MDF
			főváros	Fidesz-MDF				főváros	Fidesz-MDF
	közép-korú	alapfokú	község	Fidesz-MDF		közép-korú	alapfokú	község	Fidesz-MDF
			város	MSZP				város	Fidesz-MDF
			megyeszékhely	Fidesz-MDF				megyeszékhely	Fidesz-MDF és MSZP
			főváros	MSZP				főváros	MSZP
		középfokú	község	Fidesz-MDF			középfokú	község	MSZP
			város	MSZP				város	Fidesz-MDF és MSZP
			megyeszékhely	MSZP				megyeszékhely	Fidesz-MDF és MSZP
			főváros	MSZP				főváros	MSZP
		felsőfokú	község	Fidesz-MDF			felsőfokú	község	Fidesz-MDF
			város	Fidesz-MDF és MSZP				város	Fidesz-MDF és MSZP
			megyeszékhely	Fidesz-MDF és MSZP				megyeszékhely	Fidesz-MDF
			főváros	MSZP				főváros	Fidesz-MDF
	idős	alapfokú	község	Fidesz-MDF		idős	alapfokú	község	Fidesz-MDF
			város	MSZP				város	MSZP
			megyeszékhely	MSZP				megyeszékhely	Fidesz-MDF
			főváros	MSZP				főváros	MSZP
		középfokú	község	MSZP			középfokú	község	MSZP
			város	MSZP				város	Fidesz-MDF és MSZP
			megyeszékhely	MSZP				megyeszékhely	MSZP
			főváros	MSZP				főváros	Fidesz-MDF
		felsőfokú	község	MSZP			felsőfokú	község	MSZP
			város	MSZP				város	MSZP
			megyeszékhely	MSZP				megyeszékhely	Fidesz-MDF
			főváros	Fidesz-MDF				főváros	Fidesz-MDF

18. táblázat: A móduszok bemutatása a négy dimenzió mentén

Korábbi felvetésünk – miszerint minden esetben a Fidesz-MDF lesz a leggyakrabban előforduló elem - nem igazolódott be, a 72 rétegre végzett összehasonlításból az derült ki, hogy 40-ben a Fidesz-MDF, 23-ban az MSZP volt a módusz, a maradék 7 rétegben pedig egyforma gyakorisággal szerepel ez a két párt. Másik feltevésünk, miszerint ez a módszer sokkal inkább félreviszi az előrejelzést, beigazolódott.

4.2.5. Hot deck imputáció (FZ)

Az eljárás során a teljes egységek között keresünk olyat, amely nagyon hasonlít a hiányos egységhez, és azt a választ imputáltuk a hiányzó adat helyére. A 19. táblázat foglalja össze a kapott eredményeket.

<i>partlist</i>	Hot deck	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (hot deck)
Centrum	3,81%	3,50%	3,90%	0,40	0,09
Fidesz-MDF	46,02%	46,00%	41,07%	-4,93	-4,95
MIÉP	3,61%	4,50%	4,37%	-0,13	0,76
MSZP	37,25%	38,00%	42,05%	4,05	4,80
SZDSZ	5,65%	4,50%	5,57%	1,07	-0,08
			SSE	42,03	48,12

19. táblázat: A hot deck imputációval kapott eredmények páronként

Látható, hogy továbbra sem sikerült alacsonyabb eltérésnégyzetösszeggel, páronként jobb becslést adni a tényleges eredmények vonatkozásában. A módszer alkalmazhatóságának nehézségét bizonyítja, hogy nem könnyű definiálni azt, hogy mi a hasonló.

Az imputáció során a korábbi 4 ismerv mentén (nem-kor-végzettség-településtípus) leválogattuk a kérdezés során az egyes pártlistákra leadott „voksokat”, majd ezeket százalékos formába öntve meghatároztuk a relatív gyakoriságokat, figyelmen kívül hagyva a „megtagadja a választ”, „nem tudja” és „nem megy el szavazni” kategóriákat. Ezután ezeket a hiányzó adatokat a relatív gyakoriságoknak megfelelően imputáltuk az egyes pártok viszonylatában.

4.2.6. Többszörös imputáció Monte-Carlo szimulációval (VK, FZ)

A hiányzó adatok helyére Monte-Carlo szimulációval több értéket is generálunk, majd az így kapott különböző adatbázisokat együttesen értékeljük. Mint Rubin (1987) megjegyzi, 3-10 imputálásnál többet nem érdemes elvégezni, mert a relatív hatékonyságot nagyobb számú adatbázis már nem növeli tovább. Mi ezeknek megfelelően 5 imputálást hajtottunk végre.

Az egyes pártokra adott becslésünk a különböző imputálásokkal kapott becslések átlaga lesz:

$$\bar{\Theta} = \frac{\sum_{j=1}^m \hat{\Theta}_j}{m}$$

<i>partlist</i>	Monte-Carlo	Tárki	Tényleges	Eltérés (Tárki)	Eltérés (Monte-Carlo)
Centrum	3,87%	3,50%	3,90%	0,40	0,03
Fidesz-MDF	45,26%	46,00%	41,07%	-4,93	-4,19
MIÉP	3,66%	4,50%	4,37%	-0,13	0,71
MSZP	36,59%	38,00%	42,05%	4,05	5,46
SZDSZ	5,72%	4,50%	5,57%	1,07	-0,15
			SSE	42,03	47,98

20. táblázat: A Monte-Carlo szimulációval kapott eredmények bemutatása pártonként

Az imputálás által elért relatív hatékonyság megállapításához kiszámoljuk az egyes imputációkon belüli varianciát,

$$\sigma_B^2 = \frac{\sum_{j=1}^m \sigma_j^2}{m} = 45,5$$

Majd a külső varianciát,

$$\sigma_K^2 = \frac{\sum_{j=1}^m (\hat{\Theta}_j - \bar{\Theta})^2}{m-1} = 260,4$$

A teljes variancia a kettő összege lesz, egy korrekciós tényezővel bővítve:

$$\sigma^2 = \sigma_B^2 + \left(1 + \frac{1}{m}\right) \sigma_K^2 = 357,9$$

Ezáltal a hiányzó információk aránya

$$\gamma = \frac{(1 + m^{-1}) \sigma_K^2}{\sigma^2} = 0,87$$

Az ötszörös imputálás relatív hatékonysága a végtelen számú imputáláshoz képest

$$Eff_r = \left(1 + \frac{\gamma}{m}\right)^{-1} = 0,8514$$

Tehát az ötszörös imputálással 85,14%-os relatív hatékonyságot értünk el a végtelen számú imputáláshoz képest.

4.2.7. Modell a rejtőzködő és impulzus szavazókra (VK)

A megkísérelt adathiány kezelési módszerek - ahogy a fenti összefoglaló táblázatok eredményei is mutatják – nem jártak túl nagy sikerrel. Elmondható, hogy gyakorlatilag ugyanazokat a számokat tudtuk előrejelezni, mint amelyeket a Társi is publikált, az elvégzett korrekciók hatása a végeredményre valamennyi esetben elenyésző volt. Ebből arra következtettünk, hogy alapfeltevésünk, - mely szerint az adathiány típusa ún. MCAR, vagyis teljesen véletlenszerű – nem helyénvaló, ettől fogva a korábban felsorakoztatott statisztikai módszerek sem alkalmasak az előrejelzési modell szignifikáns javítására.

Más módszerhez kell folyamodnunk, amivel a nem véletlenszerű adathiányt megpróbáljuk kiküszöbölni. A korábbiakban bemutatott társadalomtudományi elméletek számos jelenséget mutattak be, melyeket az előrejelzési torzítások okozójának véltek és amelyek kétségkívül befolyásolhatták az eredményeket, azt azonban nem tudjuk, hogy milyen mértékben. Tisztán statisztikai módszerekkel közelítve az eltéréseket beláttuk, hogy nem elégségesek a becslések javítására. Ezért úgy gondoltuk, hogy a kétféle ötvözet lehet az, amely közelebb visz minket a megoldáshoz. A következőkben megpróbáljuk ezeket a társadalmi jelenségeket, hatásokat számokban leképezni és megfigyelni a hatásukat.

Kolosi-Tóth (2002) művében részletesen foglalkozik az okokkal, amelyek befolyásolhatták a választások kimenetelét, ugyanakkor az adatgyűjtés és elemzések során nem merültek fel. Egy korábbi fejezetben, már idéztünk a műből, melyben arról nyilatkoztak, hogy tisztában voltak a nemválaszolási arány kimagasló mértékével, de nem foglalkoztak vele, mert a korábbi évek tapasztalta azt mutatta, hogy szükségtelen. Bebizonyosodott azonban, hogy valamilyen okból kifolyólag magatartásbeli különbségek jelentkeztek a társadalom egyes rétegeiben, melyet nem tudtak megfelelően leképezni az előrejelzések során. Úgy gondolták, hogy azok, akik megtagadják a válaszadást, vagy bizonytalanok, azok valószínűsíthetően ugyanolyan arányban fognak végül szavazni, mint a biztos szavazók körében kapott relatív gyakoriságok, vagyis az eloszlás ugyanolyan lesz mindkét csoport esetében. Ez végül nem így történt, a szerzőpáros az alábbi kérdéseket fogalmazta meg, melyek kapcsán ismertetjük majd az elemzésünk szempontjából releváns válaszokat.

„Miért nem tűnt fel, hogy ez a csoport igen nagy, és közvetlenül a választások előtt növekszik?”⁷

„Minek tulajdonítható, hogy a bizonytalanok és a válaszmegtagadók csoportja a várakozásokkal ellentétben teljes egészében elment szavazni?”

„Miért különbözik ennek a csoportnak a pártpreferenciája radikálisan a biztos szavazó pártválasztóktól? Azért, mert ők már régi elkötelezett szavazók, csak a választáskutatás számára nem akarják felfedni a preferenciáikat, vagy azért, mert még tényleg bizonytalanok és nem rendelkeznek válasszal a kampány néhány olyan hívó szavára, amelyek az utolsó napokban különösen erősen tematizálódtak.”

Az első kérdésre a választ már egy korábbi fejezetben bemutattuk, így azzal most nem foglalkozunk. Érdekesebb azonban az, hogy azt feltételezik, hogy a csoport teljes egészében elment szavazni. Erre az lehet a válasz, hogy sokkal inkább a baloldali szavazók nem nyilatkoztak pártpreferenciájukról, és érvényesülhetett a mobilizációs készség azon fajtája, amikor a vesztesnek látszó fél helyzete motiválja a potenciális választót a részvételre. Magyarán nagyobb mértékben aktivizálták magukat a baloldali szavazók a Fidesz-MDF előrejelzett előnyösebb pozíciója miatt. Ezzel egyetemben kizárhatjuk az utánfutó effektust, hiszen akkor a bizonytalan szavazók is a Fidesz-MDF-re tették volna le voksukat, ez pedig nagy valószínűséggel jobboldali győzelmet jelentett volna.

A harmadik kérdést a legnehezebb megválaszolni. Itt első lépésben alapvetően azt vizsgáljuk, hogy van-e különbség a válaszmegtagadók és a bizonytalan szavazók között. Kolosi-Tóth (2002) más elemzések alapján megállapította, hogy politikai attitűd szempontjából a két csoport gyakorlatilag nem tér el egymástól, társadalmi státusuk ugyanakkor különböző. Ez alapján nagyon fontos kiindulópont lesz az a 2002-es vizsgálat, amelyben pártonként elemezték a bizonytalanok és válaszmegtagadók társadalmi státusát. A vizsgálat eredményét tekintve arra a következtetésre jutottak, hogy az SZDSZ támogatói alkotják a legmagasabb társadalmi státusú csoportot, őt követik a Fidesz-MDF szavazók, végül pedig az MSZP támogatói következnek.

⁷ Csoport kapcsán a szerzőpáros a nem ismert pártpreferenciájú, biztos szavazókra gondol.

A rangsor megállapításánál a fogyasztási index⁸, pénzügyi index, és az iskolázottság voltak a legfőbb befolyásoló tényezők. (lásd 2. számú melléklet) Elemzésünk egyik legfőbb kiindulópontja ezáltal az a megállapítás lesz, mely szerint a bizonytalanok valamennyi szempontból az átlag alatti társadalmi státus csoportjába sorolhatóak. A válaszmegtagadók ugyanakkor jellemzően magasabb társadalmi státussal rendelkeznek.

Második lépésben két csoportra bontjuk a potenciális választókat. Az egyik csoportba azokat fogjuk sorolni, akik a részvételi szándék kapcsán azt válaszolták, hogy biztosan elmennek szavazni, viszont a pártlista esetében megtagadták a válaszadást. Ők lesznek az ún. rejtőzködő szavazók, vagyis a korábbi besorolás szerint a válaszmegtagadók, akik a magasabb státussal rendelkeznek. A másik csoportot pedig az impulzív szavazók fogják alkotni, akik bizonytalanok és a pártlisták viszonylatában a „nem tudom” kategóriát jelölték meg.

Esetükben feltételezhetjük, hogy nincs pártkötődésük, amennyiben elmennek szavazni, az utolsó pillanatban döntenek el, hogy kire teszik le a voksukat. Azt azonban nem tudjuk, hogy a döntéseket milyen impulzus tényezők befolyásolják.

Harmadik lépésben megpróbáljuk becsülni külön-külön mindkét csoport eloszlását a pártok tekintetében. A rejtőzködő szavazók kapcsán azzal a feltételezéssel élünk, hogy a biztos részvételt ígérő, konkrét pártot is megjelölő válaszadók eloszlásával azonos az eloszlásuk, így ezzel könnyedén tudunk korrigálni a rendelkezésünkre álló adatbázisból az előrejelzésekre vonatkozóan. Az impulzus szavazók preferenciáit azonban jóval nehezebb becsülni. Belátható, hogy ezen a csoporton belül magasabb kell, hogy legyen az MSZP-t támogatók aránya, hiszen minden korábbi elemzés előnyösebb Fidesz-MDF pozíciót mutatott. Ennek a problémának az áthidalására kötöttük össze a bizonytalanok alacsony státusát az eloszlásuk meghatározásával. Mivel a fogyasztási indexre és pénzügyi indexre vonatkozóan nem volt adatunk, így a korábban bemutatott vizsgálatban szignifikáns tényező, az iskolázottság alapján készítettünk becslést a következőképpen:

⁸ A fogyasztási index a tartós fogyasztási cikkekkel való ellátottság indexe, a pénzügyi index pedig a megtakarítási és biztosítási instrumentumokkal való ellátottságot mutatja

Az adatbázisból leválogattuk a rejtőzködőket, akik biztosan elmennek szavazni, de nem mondták meg, hogy kire. Ez összesen 1.929 főt jelent a mintában, akiket a biztos szavazó pártválasztók mintabeli arányában osztottunk fel a pártok között.⁹ Így a Fidesz-MDF 46%, az MSZP 38%, az egyéb pártok pedig 16%-os súllyal szerepelnek ezen potenciális szavazók „felosztásánál”.

A 2.382 impulzus szavazóhoz kapcsolódóan (biztosan- valószínűleg- vagy nem tudja, hogy elmegy-e szavazni és azt sem tudja még, hogy kire) megnéztük az adatbázisban a pártpreferenciával nem rendelkező bizonytalan szavazók arányait, az iskolai végzettség (alap-közép-felsőfok) mentén összesítve. A kapott relatív gyakoriságokkal pedig a korábbi tanulmányban szereplő rangsor szerint osztottuk szét a pártok között a csoport tagjait.

Ez azt jelenti, hogy az alapfokú végzettségnél kapott súllyal szoroztuk meg az impulzus szavazók számát az MSZP esetében, a középfokúval a Fidesz-MDF esetében, a felsőfokúval pedig az egyéb pártok esetében (az SZDSZ miatt). Az eredményeket a 21. táblázat összesíti.

Párt	Biztos szavazók	Impulzus szavazók	Rejtőzködő szavazók	Σ	Becsült eredmény	Tényleges eredmény	Eltérés
Fidesz-MDF	3849,00	481,21	883,73	5213,95	41,0%	41,1%	0,11
MSZP	3182,00	1764,44	729,04	5675,48	44,6%	42,1%	-2,53
Egyéb pártok	1388,00	136,34	316,23	1840,57	14,5%	16,9%	2,42
						SSE	12,28

21. táblázat: A rejtőzködők és impulzus szavazók becslésével kapott eredmények pártanként

Az MSZP és egyéb pártok esetében a legfőbb eltérést a MIÉP kiemelten nagy válaszmegtagadó aránya jelenti, melyet a Tárki nagyjából 50.000 főre becsült. Ez a választásra jogosultak megközelítőleg 1%-a.

⁹ Mivel a kis pártokat alapvetően jól jelezték előre a közvélemény-kutatók, továbbá hasonló eredményekre jutottunk az egyes adathiánykezelési módszereknél is, így azzal az egyszerűsítéssel éltünk, hogy csak a két nagy pártot különböztettük meg, a többi az „egyéb” kategóriában szerepel

Társadalmi státusukat tekintve a MIÉP szimpatizánsok jellemzően alacsonyabb végzettségűek és az idősebb korosztályból kerülnek ki, vagyis pont onnan, akiket teljes egészében MSZP szavazóknak feltételeztünk. Ez okozza az MSZP javára, a MIÉP kárára történő 1-2% pontos eltérést, ráadásul a felsőfokú végzettségűek, akiket az egyéb pártok közé soroltunk az SZDSZ miatt, jóval kisebb arányt képviselnek a társadalomban ugyanúgy, mint a mintában, ezért becsültük alul az egyéb pártok eredményét az MSZP-vel szemben.

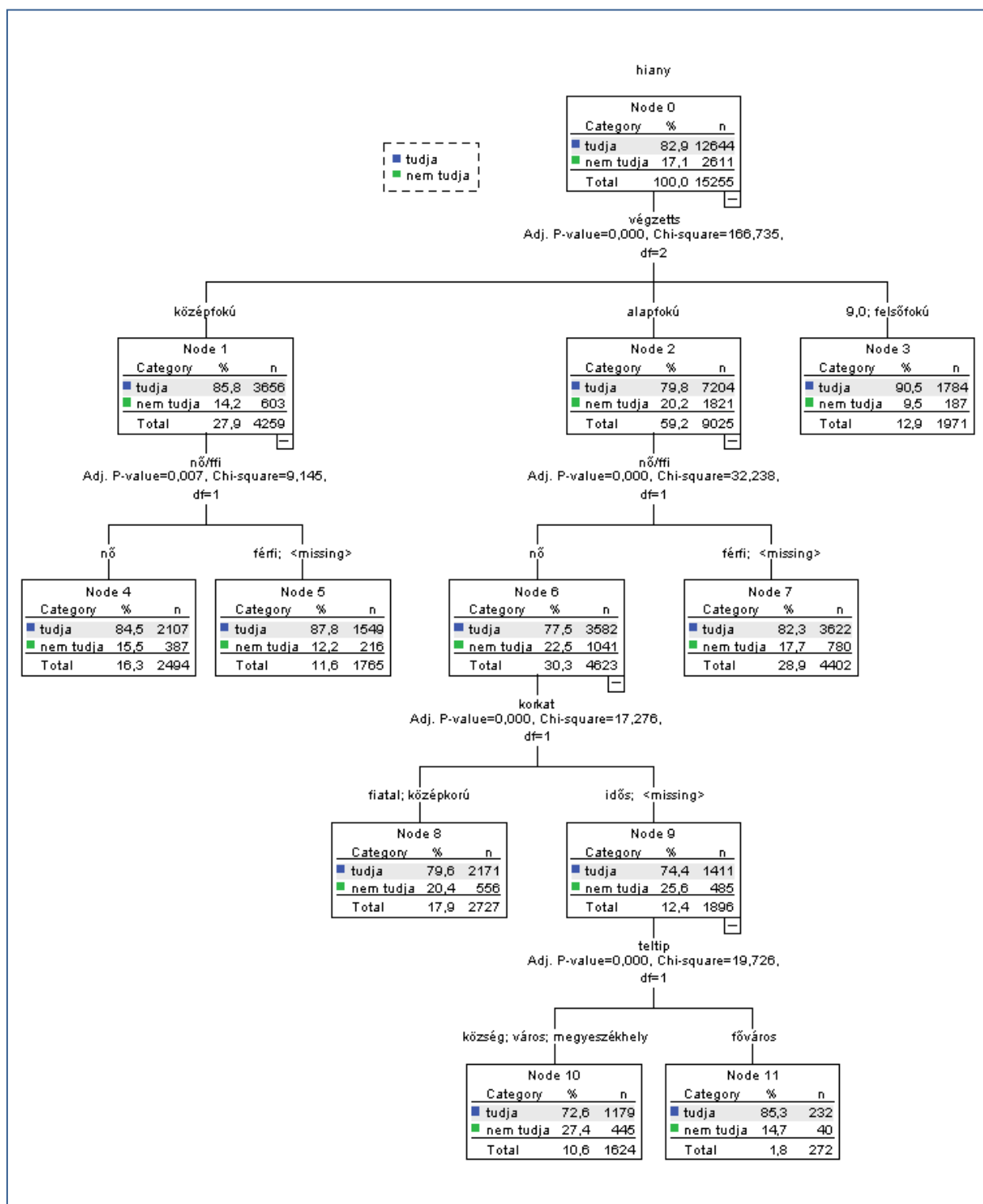
4.2.8. Döntési (klasszifikációs) fa (FZ)

A fenti gondolatmenet arra épített, hogy az általunk használt demográfiai ismérvek közül az iskolai végzettségnek van a leginkább szignifikáns hatása arra a tényre, hogy a válaszadó a „nem tudja” választ adja-e meg a pártistára vonatkozó kérdésnél. Ennek alátámasztására egy döntési fát készítettünk, melyben a *hiány* változó kerül elemzésre: ez egy olyan bináris (dummy) változó, melynek értéke 1, ha a *partlist* változóban a „nem tudja” értéket vette fel egy mintaelem, és 0 minden más esetben.

A magyarázó változók között a korábban is használt nem, kor, végzettség és településtípus szerepeltek. A fa¹⁰ (22. táblázat) alapján jól látható, hogy a végzettség mentén van a legnagyobb különbség ($\chi^2 = 166,735$) a véleményhiány alakulásában: az alacsony végzettségűek 20,2%-a, a középfokú végzettségűek 14,2%-a, a felsőfokú végzettségűek 9,5%-a nem tudta 1-3 nappal a választás előtt, hogy melyik pártra fog szavazni. Az ezt fontossági sorrendben követő szempont, a *nem* ehhez képest kevésbé szignifikáns, a χ^2 statisztika értéke 32,238 és 9,145, illetve a felsőfokú végzettségűeket más szempont mentén nem is érdemes tovább bontani.¹¹

¹⁰ A döntési fa CHAID algoritmussal készült, de 4 leveles mélységet engedélyeztünk, hogy a négy változó fontossági sorrendje jól kirajzolódjon. Az ágnak legalább 100 eleműnek kellett lennie, hogy további elágazások indulhassanak belőle, a leveleknek pedig legalább 50 eleműnek kell lenniük, hogy kiválhassanak az ágból.

¹¹ A fa CRT algoritmussal futtatva is a végzettséget, mint ismérvet mutatta a legszignifikánsabbnak



22. táblázat: Döntési fa a demográfiai ismérvek mentén

5. ÖSSZEFOGLALÁS (FZ)

A dolgozatban használt módszerek eredményeit a 23-24. táblázatok tartalmazzák.

Pártlista	Listwise	Available Case Analysis	Átsúlyozás (nem)	Átsúlyozás (kor)	Átsúlyozás (település)	Átsúlyozás (végzettség)
Centrum	3,30	4,00	3,96	3,93	3,93	3,85
Fidesz-MDF	48,10	45,90	45,92	45,69	46,03	45,96
MIÉP	3,90	3,80	3,72	3,72	3,70	3,68
MSZP	36,40	37,10	37,03	37,31	37,04	37,17
SZDSZ	5,30	5,80	5,81	5,72	5,72	5,71
SSE (5 párt)	82,00	48,22	49,21	44,26	50,17	48,22
SSE (2 párt)	81,34	47,83	48,72	43,81	49,70	47,73

23. táblázat: Az egyes módszerekkel kapott eredmények összegzése I.

Pártlista	Módusz	Hot Deck	Multiple Imp. (Monte-Carlo)	Rejtőzködők - Impulzus szavazók	Tárki	Tényleges
Centrum	2,56	3,81	3,87	n/a	3,50	3,90
Fidesz-MDF	54,89	46,02	45,26	41,00	46,00	41,07
MIÉP	2,40	3,61	3,66	n/a	4,50	4,37
MSZP	34,10	37,25	36,59	44,60	38,00	42,05
SZDSZ	3,75	5,65	5,72	n/a	4,50	5,57
SSE (5 párt)	263,18	48,13	47,90	n/a	42,03	
SSE (2 párt)	254,19	47,54	47,37	6,51	40,71	

24. táblázat: Az egyes módszerekkel kapott eredmények összegzése II.

Az eredeti statisztikai eszközök, melyekkel a fellépő jelentős adathiányt kezeltük, nem vezettek a Tárki előrejelzésénél jobb becslésekre. A listwise adattörlés nagyjából kétszer akkora hibát produkált, mint az elérhető adatok elemzése. Ráadásul esetünkben nem merül fel a korrelációs mátrixok definitségének problémája, lévén minden változónk nominális vagy ordinális mérési skálájú – tehát indokolatlannak tartjuk az adathiányos elemek törlését.

Az átsúlyozást négy demográfiai dimenzió mentén tudtuk megvalósítani. A nem, a településtípus és a végzettség szerinti átsúlyozás hasonló eredményeket adott, a kor szerinti súlyozás kissé sikeresebb volt. Mivel mind a négy ismerv ugyan szignifikáns, de gyenge kapcsolatot mutatott a pártlistás szavazattal, a kor szerinti átsúlyozás sikere inkább meglepő, mint hogy alátámaszthatónak tekinthetnénk az alakulását.

Az imputációs eljárások szintén nem adtak jelentős javulást, sőt a módusszal való becslés teljesen félrevitte az előrejelzéseket. Ennél némiképp eredményesebb volt, amikor a hiányzó adatokat nem ugyanazzal az egyébként jellemzőnek vélt értékkel, hanem a négy demográfiai ismerv szerint meghatározott eloszlással pótoltuk. Végül többszörös imputálást végeztünk Monte-Carlo szimulációval, amelynek segítségével az előzőekhez képest az eredmények kismértékben javultak.

Ezek a módszerek mind azzal a feltételezéssel éltek, hogy az adathiány MCAR-típusú, vagyis teljesen véletlenszerű: a nemválaszolók pártlistás szavazatainak megoszlása egyezik a válaszolók szavazatainak megoszlásával, csak annyi a különbség a két csoport között, hogy megmondták-e a preferenciáikat vagy sem. Azáltal, hogy egyik eljárással sem lehetett legalább a Társiét meg nem haladó hibát elérni, illetve egyszer sem tűnt fel győztesen az MSZP, az a tény vált bizonyítottá, hogy az adathiány nem volt teljesen véletlenszerű.

A nem teljesen véletlenszerű (MAR) adathiány hipotézisét már csak azért is el kellett vetnünk, mert ahol a pártlistára vonatkozó kérdés nem került megválaszolásra, ott a többi változóban is hiány volt – kivéve a demográfiai ismérveket, amelyekkel való kapcsolatra már utaltunk. Így az adathiány ez esetben NMAR, vagyis nem véletlenszerű volt.

Az elméleti megfontolások mentén korrigált modell – amely a rejtőzködő és impulzív szavazók eloszlását becsülte –, bár csak a Fidesz és az MSZP eredményeit volt hivatott előrejelezni, ebben minden más módszernél, valamint a Társi előrejelzésénél is pontosabb volt. Nem utolsósorban pedig a valódi győztest prognosztizálta.

6. IRODALOMJEGYZÉK

Angelusz, Róbert [2002]: Közvéleménykutatások és a pluralizmus ignoranciája - *Médiakutató* 3, 2002 ősz (55-68)

Angelusz, Róbert [2006]: A közönség elérhetősége, avagy a közvélemény-kutatások reprezentativitásának nehézségei – *Csepli* 60. 2006, elektronikus dokumentum http://www.csepli.com/kotet/csepli60_angelusz_robert.pdf (utolsó letöltés 2012.03.25.)

Karácsony, Gergely – Lakatos, Zsuzsa [2006]: Választási előrejelzések Magyarországon, 1990-2006. Sikerek, tévedések és tanulságok - *Angelusz Róbert és Tardos Róbert (szerk.): Mérésről mérésre. A választáskutatás módszertani kérdései. Budapest: Demokrácia Kutatások Magyar Központja Alapítvány, 2006. 185-220. oldal*

Kolosi, Tamás – Tóth, István György [2002]: Egy tévedés története - *Társadalmi riport 2002, Kolosi Tamás, Tóth István György, Vukovich György (szerk.). Budapest: TÁRKI, 339-367. oldal*

Lazarsfeld, P. & Berelson, B. & Gaudet [1944]: The People's Choice – *Columbia University Press, New York*

Little, Roderick J.A. – Rubin, Donald B. [2002]: Statistical Analysis with Missing Data - *John Wiley & Sons*

Magyar Gallup Intézet [2002]: A közvélemény-kutatás fekete napja (szerző nélkül) <http://www.gallup.hu/Gallup/release/ppref020408.htm> (utolsó letöltés 2012.03.25.)

Noelle-Neumann, E [1980]: Die Schweigespirale – *Öffentliche M. - unsere soziale Haut, München & Zürich, Piper-Verlag*

Oravecz, Beatrix [2008]: Szelekciós torzítás és csökkentése az adósmínősítése modelleknél (Ph.D értekezés) - *Budapesti Corvinus Egyetem, Gazdálkodástudományi Doktori Iskola, Budapest, 2008.*

Rubin, Donald B. [1987]: Multiple Imputation for Nonresponse in Surveys – *John Wiley & Sons*

Rudas, Tamás [2006]: Közvélemény-kutatás, Értelmezés és kritika – *Corvina Kiadó 2006*

7. MELLÉKLETEK

7.1. 1. számú melléklet – A Társi által használt kérdőív

ELSŐ FORDULÓ			
S O R S Z Á M			
Kérdezői ig. szám:			

8. Véleménye szerint melyik párt nyeri a választásokat?

- 1 – FIDESZ Magyar Polgári Párt – Magyar Demokrata Fórum
 2 – MIÉP, Magyar Igazság és Élet Pártja
 3 – MSZP, Magyar Szocialista Párt
 4 – SZDSZ, Szabad Demokraták Szövetsége
 5 – egyéb párt
 8 – M 9 – NT X –

9. Ön mit szeretne inkább? Azt, hogy ...

- 1 – a Fidesz-MDF vezetésével alakuljon kormány, vagy azt, hogy
 2 – az MSZP vezetésével alakuljon kormány?
 3 – egyéb válasz
 8 – M 9 – NT X –

10. A választások után megalakuló kormányban kit látna legszívesebben miniszterelnökként?

- 1 – Medgyessy Pétert
 2 – Orbán Viktort
 3 – másvalakit
 4 – egyik politikust sem
 8 – M 9 – NT X –

11. Hol helyezné el saját politikai nézeteit ezen a baloldaltól jobboldaliig terjedő skálán? **2.VÁLASZLAP**

01	02	03	04	05	06	07	08	09	10
bal			88 – M		99 – NT		X		jobb

12. Mire számít, a következő 12 hónapban az ország gazdasági helyzete...

- 5 – sokkal jobb lesz,
 4 – valamivel jobb lesz,
 3 – nem változik,
 2 – valamivel rosszabb lesz, vagy
 1 – sokkal rosszabb lesz?
 8 – M 9 – NT X –

13. Részt vett-e Ön a legutóbbi, 1998. májusi parlamenti választások első fordulóján?

- 1 – igen 2 – nem
 8 – M 9 – NT, nem emlékszik X –

14. **HA IGEN**

Melyik párt listájára szavazott az első fordulóban?

- 1 – FIDESZ– Magyar Polgári Párt
 2 – MSZP, Magyar Szocialista Párt
 3 – FKgP, Független Kisgazda Párt
 4 – SZDSZ, Szabad Demokraták Szövetsége
 5 – MIÉP, Magyar Igazság és Élet Pártja
 6 – MDF, Magyar Demokrata Fórum
 8 – M 9 – NT X –

15. A kérdezett neme:

- 1 – férfi
 2 – nő
 X –

16. Ön melyik évben született?

19 évben X –

17. Mi az Ön legmagasabb iskolai végzettsége?

- 1 – kevesebb, mint 8 általános
 2 – 8 általános
 3 – szakmunkásképző
 4 – középiskolai érettségi
 5 – főiskolai diploma, felsőfokú technikum
 6 – egyetemi diploma
 X –

18. Ön jelenleg dolgozik /nyugdíjas?

- 01 – alkalmazott
 02 – saját vállalkozásában dolgozik
 03 – alkalmi munkából, megbízásokból él
 04 – munkanélküli
 05 – nyugdíjas
 06 – GYES, GYED
 07 – sorkatoná
 08 – háztartásbeli
 09 – eltartott, szociális segélyből él, családja támogatja
 10 – jövedelmeiből él
 11 – tanuló
 12 – segítő családtag
 13 – egyéb
 X –

19. Mi (volt) az Ön (legutolsó) foglalkozása, beosztása?

ÖNÁLLÓK, VÁLLALKOZÓK

- (tulajdonos vagy résztulajdonos)
 01 – gazdálkodó (mezőgazdaságban)
 02 – iparos (termelő tevékenységet folytat)
 03 – kereskedő
 04 – szolgáltató
 05 – szellemi szabadfoglalkozású

ALKALMAZOTTAK

VEZETŐK:

- 06 – felső vezető (osztályvezető felett)
 07 – középszintű vezető (osztályvezető.)
 08 – alsó vezető (osztályvezető alatt)
 09 – közvetlen termelésirányító (pl. művezető)

SZELLEMI

- 10 – diplomához kötött
 11 – egyéb szellemi (diploma nélkül)

FIZIKAI

- 12 – szakmunkás (nem mezőgazdasági)
 13 – betanított munkás
 14 – segédmunkás
 15 – mezőgazdasági fizikai
 88 – M X –

KÖSZÖNJÜK VÁLASZAIT!

7.2. 2. számú melléklet – Kolosi-Tóth (2002) 354. oldal

	Fogyasz- tási/vagy oni index*	Pénzügyi index**	Iskolai osztályok száma	Átlag életkor (év)	Száz budapes- tíre jutó falusiak aránya (%)	Száz diplo- másra jutó alapfokú végzett- ségű (fő)	Száz foglal- kozta- tottra jutó nyugdí- jas (%)	Egykori MSZMP- tagok aránya (%)
SZDSZ	5,2	2,7	12,9	40,7	37	93	12	0
Fidesz– MDF	4,1	1,9	10,9	43,7	299	526	54	2
MSZP	3,9	1,7	10,7	48,1	167	522	95	12,1
Válasz- megta- gadó	3,9	1,7	11,1	49,8	227	321	77	3,1
Egyéb párt	3,8	1,6	10,5	49,5	89	683	78	7,7
Nem tudja	3	1	9,6	51,2	269	1786	123	3,3
Átlag	3,9	1,7	10,8	46,8	197	507	74	6,4

Megjegyzések: * A tartós fogyasztási cikkekkel való ellátottság indexe, 1–7 skálán.

** A megtakarítási és biztosítási instrumentumokkal való ellátottság, 1–6 skálán.