

MISKOLCI EGYETEM
GAZDASÁGTUDOMÁNYI KAR

SZILÁGYI ROLAND

**MINTAVÉTELEN ALAPULÓ
BECSLÉSEK HIBÁINAK KEZELÉSE,
KÜLÖNÖS TEKINTETTEL A NEMVÁLASZOLÁS
OKOZTA PROBLÉMÁKRA**

PH.D. ÉRTEKEZÉS TÉZISEI

MISKOLC

2011

MISKOLCI EGYETEM
GAZDASÁGTUDOMÁNYI KAR

SZILÁGYI ROLAND

**MINTAVÉTELEN ALAPULÓ
BECSLÉSEK HIBÁINAK KEZELÉSE,
KÜLÖNÖS TEKINTETTEL A NEMVÁLASZOLÁS
OKOZTA PROBLÉMÁKRA**

PH.D. ÉRTEKEZÉS TÉZISEI

A DOKTORI ISKOLA NEVE: **VÁLLALKOZÁSELMÉLET- ÉS GYAKORLAT
DOKTORI ISKOLA**

A DOKTORI ISKOLA VEZETŐJE: **PROF. DR. SZINTAY ISTVÁN**
egyetemi tanár

TUDOMÁNYOS VEZETŐ: **PROF. DR. BESENYEI LAJOS**
egyetemi tanár

MISKOLC

2011

TARTALOM

1. A kutatás célja, az értekezés felépítése	3
2. A kutatás módszere	6
3. A kutatás új és újszerű megállapításai.....	7
4. A kutatás eredményeinek hasznosítási területei	19
Irodalomjegyzék.....	20
A szerző témához kapcsolódó publikációi.....	22
Melléklet	25

1. A KUTATÁS CÉLJA, AZ ÉRTEKEZÉS FELÉPÍTÉSE

A piaci szerkezet módosulása, a gyors környezeti változások, az információs technikák fejlődése eredményeképpen a döntések előkészítéséhez szükséges időtáv jelentősen lerövidült. Az üzleti gazdasági élet döntéshozói csak úgy vehetik fel hatékonyan a versenyt – az idővel és a versenytársakkal egyaránt – ha folyamatosan fejlesztik a döntés-előkészítés során alkalmazott technikákat. A tömegesen előforduló jelenségek jellemzésére irányuló kvantitatív elemzési módszerek széles tárháza nyújt lehetőséget az egyedek jellemzőinek vizsgálatára. Egy sikeres kutatásnak azonban csak egyik kulcspontja a megfelelő kvantitatív módszer megválasztása és tudományos alkalmazása. Másik kulcspontja mindenképp a felhasznált adatok információtartalmának megbízhatósága kell, hogy legyen.

A mintára épülő vizsgálatok és következtetések egyre nagyobb szerepet kapnak a gazdasági döntések meghozatalában és az információ képzésben egyaránt. A mintavételek terjedését elsősorban a költségek és a vizsgálatához szükséges idő csökkentése indukálja. A mintán alapuló felmérések nem csak mikro szinten, hanem makrogazdasági vizsgálatoknál is egyre népszerűbbek, de a mintavételek terjedésének azonban nagy veszélye is van, pontosan a minta minősége miatt. A mintán alapuló adatokból nyert információk nagyon sok és sokféle hibát hordozhatnak magukban. Ezeknek a hibáknak a feltárása és matematikai-statisztikai módszerekkel történő értékelése, valamint az eredményre gyakorolt negatív hatásuk csökkentése képezi kutatásom tárgyát. Céлом olyan elemzési módszer, szempontrendszer kidolgozása, mely alkalmazkodik a modern információs technikák és technológiák nyújtotta lehetőségekhez, és jól kiegészíti az eddig alkalmazott módszertant.

Kutatási munkám elsődleges céljának a mintavételen alapuló kutatások lehetséges hibáinak, és azok eredményekre gyakorolt negatív hatásainak feltárását jelöltem meg. Ezt követően megoldási változatokat kerestem a hibák, majd elsődlegesen a nem véletlen jellegű hibák kezelésére. A hibák kezelésének leghatékonyabb módja, ha megelőzzük a keletkezésüket. Azonban ha a hiba már bekövetkezett, akkor a kezelés első fázisaként fel kell térképezni a hiba okát. Milyen okokra vezethetők vissza a társadalomtudományi kutatások hibái? Melyek lehetnek azok az eredő tényezők, amelyek alapján egy kutatás, vagy annak eredménye megalapozatlannak minősül? Ezekre a kérdésekre a szakirodalom a következő válaszokat adja: megbízhatatlan, megalapozatlan, kis mintára épül, magas a relatív hiba, nem szignifikáns az eredő problémára irányuló hatása, stb.

További kritikai tényezők merülhetnek fel abban az esetben, amikor a kutatás egy mintavétel során nyert primer adathalmazt dolgoz fel, elemez, és von le különböző következtetéseket. A mintavétellel ugyanis a hibaforrások száma fokozódik. A mintavétel során elkövethető hibák már az előállított forrásadatok mennyiségében és minőségében is torzulást eredményezhetnek, ami azért veszélyes, mert a rossz alapadatokból a legjobb módszertan alapján, professzionális elemzési technikák és eszközök segítségével végzett alapos számítások is téves következtetésekhez vezetnek.

A fenti problémákra és kérdésekre a megoldásokat és a válaszokat a KSH által rendelkezésemre bocsátott 2005. évi Háztartási Költségvetési Felvétel (HKF) adatbázisának vizsgálatával kerestem. Az adatok hitelességének alátámasztásához bemutattam a HKF adatainak forrását, a felvétel jellemzőit. Különböző minták elemzésén keresztül törekedtem a hibák hatásának bemutatására, és a negatív hatások csökkentésének tudományos magyarázatára.

Doktori munkám első részében a következtetések alapját képező minta kialakításának körülményeit vizsgáltam. Első lépésben áttekintettem a mintavételi eljárásokra vonatkozó elméleti alapvetéseket, majd a mintavételi- és hibaszámítási módszerek hazai és nemzetközi szakirodalmát tanulmányoztam át, ami segített az eddigi kutatási eredmények felhasználásában.

Tudományos munkám további részében a potenciális hibaforrások azonosításával és rendszerezésével foglalkoztam, melyhez a statisztikai-matematikai megközelítéssel kívül számos adalékkal szolgáltak más diszciplínák, úgymint marketing, szociológia területén végzett kutatások eredményei is. A kérdőíves felmérések, közvélemény-kutatások tapasztalati segítettek megismerni a hibák kezelése érdekében tett lépéseket, és a hibák méretének csökkentésére alkalmazott eszközöket, módszertani elgondolásokat.

A különböző módszerek közül nagyobb részletességgel mutattam be a kalibrációs eljárás alkalmazásának elméleti hátterét. Tettem ezt egyrészt azért, mert az alkalmazott statisztikai kutatások egyre nagyobb jelentőséget tulajdonítanak a módszer sikeres alkalmazásának. Másrészt a kalibráció olyan módszer, melynek algoritmusait nem tartalmazzák a széles körben alkalmazott statisztikai elemző szoftverek. A módszer alkalmazására speciális számítógépes program szükséges, amely nehezen hozzáférhető a vállalkozások számára. A dolgozat empirikus kutatásokat tartalmazó részeiben pedig kizárólag olyan eljárásokat alkalmaztam, amelyek nemcsak a hivatalos statisztikák készítésénél, hanem a vállalkozások gyakorlatában is „egyszerűen” alkalmazhatók.

Munkám során 53 különböző mintavételi eljárás alapján generáltam mintákat a sokaságból – a HKF adatbázisából –, annak érdekében, hogy minél részletesebben vizsgálhassam a mintavételi tervek eredményre gyakorolt hatását. Következtetéseim, becslési eredményeim ellenőrzésére lehetőséget biztosított az, – a gyakorlatban nem teljesülő feltétel – hogy a vizsgált jelenség sokasági információi a birtokomban voltak. A minták egyszerű véletlen, rétegzett, illetve több ismérv szerint rétegzett eljárással készültek, mivel tapasztalataim szerint a vállalati kutatásokban jellemzően nem használnak bonyolultabb mintavételi eljárásokat.

Ezt követően olyan szempontok kidolgozására vállalkoztam, amelyek alapján – ha nem is a minőség összes értelmezhető kritériuma tekintetében, de néhány fontosabb jellemző alapján – minősíteni, rangsorolni lehet a különböző mintákból nyert adatokat, becslési eredményeket. Természetesen mindezt számos elméleti feltétel fennállása mellett kíséreltem meg, mely feltételek az alkalmazott gyakorlatban meglehetősen ritkák, olykor egyáltalán nem teljesülnek. A mintavételi tervek rang-

sorolásánál átlag- és hányadosbecslés eredményeinek javítása szempontjából rangsoroltam a mintavételi terveket.

A hazai és nemzetközi kutatások tapasztalatai alapján egyaránt elmondható, hogy a válaszadás hiányossága talán az egyik legnagyobb probléma, ami a felmérések készítésénél felmerül. Manapság nem ritkák az 50 %-on aluli válaszadási aránnyal rendelkező kérdőívek. A probléma súlyosságát jól érzékeltetik Ay-Vita (1998) kísérleti jövedelem felvételének tapasztalatai, melyek szerint az előzetes felkérések során közel 90%-os nemválaszolással szembesültek.

Nyilvánvaló, hogy a szelektív válaszadás nemcsak a mintanagyságot csökkenti, hanem növeli a becslések varianciáját is, valamint a torzítás mértékét. Éppen ezért a dolgozat további részeiben a nem mintavételi hibák egyik legfontosabb típusának, a nemválaszolási hibának a vizsgálatával foglalkoztam. Ezen belül is a részleges, vagy item szintű nemválaszolással. Különböző elemzési módszerek segítségével azt kutattam, vajon a megtagadások pótlása mekkora hatást gyakorol a leíró modellek eredményeire. A kérdés megválaszolásához a háztartások összes fogyasztási kiadásának becslését végezve különböző mértékű nemválaszolásokat generáltam, figyelembe véve azt a feltételezést, (melyet már több tapasztalati kutatás is igazolt pl.: Keszthelyiné (2006.), Havasi (1997.), Havasi-Schnell (1996.)), miszerint a jövedelemmel, fogyasztással kapcsolatos kérdésekre a magasabb jövedelemmel rendelkezők válaszolnak inkább vonakodva, belőlük kerül ki a nemválaszoló jelentős hányada. Ezt követően azt vizsgáltam, hogy a mintavételi terv hatásának van-e szerepe a nemválaszolás okozta torzítás kialakulásában.

Kutatásom utolsó fázisában a modell alapú eljárások szerepét teszteltem a nemválaszolás kezelésében. A mintában fellelhető információk alapján három klasszifikációs algoritmus: diszkriminancia-analízis, döntési fa és logisztikus regresszió alapján azonosítottam a nemválaszoló háztartásokat, majd figyelembe véve azok jellemzőit, a választ megtagadó háztartások nemválaszolási valószínűségeit térképeztem fel. A nemválaszolási valószínűségek súlyként való felhasználásával sikerült csökkenteni a becslési eredményekben rejlő torzítást. Azonban nem felejtendő el, hogy a gyakorlati kutatásoknak nem csupán az a célja, hogy enyhítsék a torzítás negatív hatásait, hanem a sokasági paraméter minél pontosabb és megbízhatóbb becslése.

A nemválaszolás okozta torzítás kiküszöbölésére tett lépések között fontos szerepe van a tendenciák azonosításának. Érdemes megvizsgálni a válaszadók és nemválaszoló vizsgált ismérvbeli tendenciái közötti különbséget. Az eljárás sikere érdekében megfelelő csoportokat kell kialakítani a mintában a vizsgált ismérvvvel sztochasztikus kapcsolatban álló és a nemválaszolást generáló ismérv(ek) alapján. A tendenciákat ezen csoportok mentén kell vizsgálni és modellezni. Tanulmányomban a fogyasztási kiadás becsléséhez a háztartások jövedelme alapján képeztem csoportokat és a különböző jövedelmi tizedekbe eső háztartások fogyasztási kiadásaiban tapasztalható exponenciális tendenciákat azonosítottam, eltérő válaszadási arányok mellett. Ezek felhasználásával megalkottam a súlyozott tendenciák becslési modelljét, mely a torzítás mértékét megfelelő keretek közé szorítja.

2. A KUTATÁS MÓDSZERE

Dolgozatomban nemcsak a mintavételen alapuló felmérések potenciális hibáinak meghatározásával, az egyes hibatípusok egyszerű bemutatásával foglalkozom, hanem olyan eljárásokat dolgozok ki, melyek a hiba mértékének minimálisra csökkentését, a torzítás mérséklését teszik lehetővé. Mindezeket olyan általánosított formában teszem, hogy hasznos segítségül szolgáljanak a társadalomtudomány területén tevékenykedő egyéni kutatók és kutató szervezetek számára egyaránt.

A mintabeli adatok alapján készített elemzések hibáinak vizsgálatára nagyméretű adatbázisok a legmegfelelőbbek. Anyagi források hiányában az elemzést megbízható szekunder adatokon végzem. Azonban szükséges foglalkoznom az adatgyűjtés, megfigyelés különböző módszereivel, hiszen azok más-más típusú hibákat generálhatnak.

Az adatelemzés során a HKF (Háztartási Költségvetési Felvétel) adatbázisán végzek kutatásokat. Ennek keretében a meglévő eredményeket egy újszerű logikai struktúra alapján kívánom csoportosítani. Olyan módszereket kívánok alkalmazni, melyek nemcsak makroszinten, hanem akár kis vállalkozások szintjén is hasznosíthatók. Hiszen a vállalkozások általában nincsenek olyan hibaszámítási szoftverek és algoritmusok birtokában, mint a hivatalos statisztika képviselői. Megfelelő módszerek hiányában pedig nem képesek jó minőségű információk előállítására. Ennek az úrnek a kitöltését célozzák az empirikus kutatásom módszertani eredményei.

A feldolgozás során az alkalmazott bonyolult és időigényes módszerek, matematikai számítások elvégzése, illetve a grafikus ábrák, táblázatok szemléletesebb formában történő megjelenítése érdekében a Windows alapú SPSS 17.0 statisztikai szoftvert, valamint a Microsoft Excel táblázatkezelő szoftvert használom.

3. A KUTATÁS ÚJ ÉS ÚJSZERŰ MEGÁLLAPÍTÁSAI

A mintavétel tervezésekor, a mintavételi eljárás kiválasztásakor nem csak rétegzett minták esetében megfogalmazott elvárás a reprezentativitás. A gyakorlatban sokszor számos jellemző alapján várnak reprezentativitást a mintán alapuló felvételektől. Ezzel kapcsolatban két megállapítást tehetők:

Az elvárás valós előnyöket realizál a becslési végeredményekben, abban az esetben, ha megfelelő változók mentén törekszünk a reprezentativitásra.

Vannak ugyan olyan esetek is, amikor bizonyos változók alapján indukált reprezentativitás kifejezetten a költségek csökkentésére, az adatgyűjtési munka megkönnyítésére, ésszerűbbé tételére irányul. Ezekben az esetekben a becslési eredményekre a reprezentativitás nincs bizonyítottan közvetlen hatással.

A több szempont szerinti reprezentativitásnak az a hátránya, hogy több ismérv alapján rengeteg réteg, illetve keresztosztály képződik, melyek mérete indokolatlanul kicsivé válhat. A következő jellemzők alapján vizsgáltam az általam használt mintákat annak eldöntésére, hogy a reprezentativitási követelmények megfogalmazásakor érdemes-e több szempontot érvényesíteni:

- Mintavételi terv hatása/ Design Effect – Deff,
- Paraméter relatív szórása / Coefficient of Variation – CV (variációs együttható):

$$\frac{SE\hat{\theta}}{\hat{\theta}}, \text{ átlag esetében: } \frac{S_{\bar{x}}}{\bar{x}}$$

- Effektív mintanagyság: n/Deff .

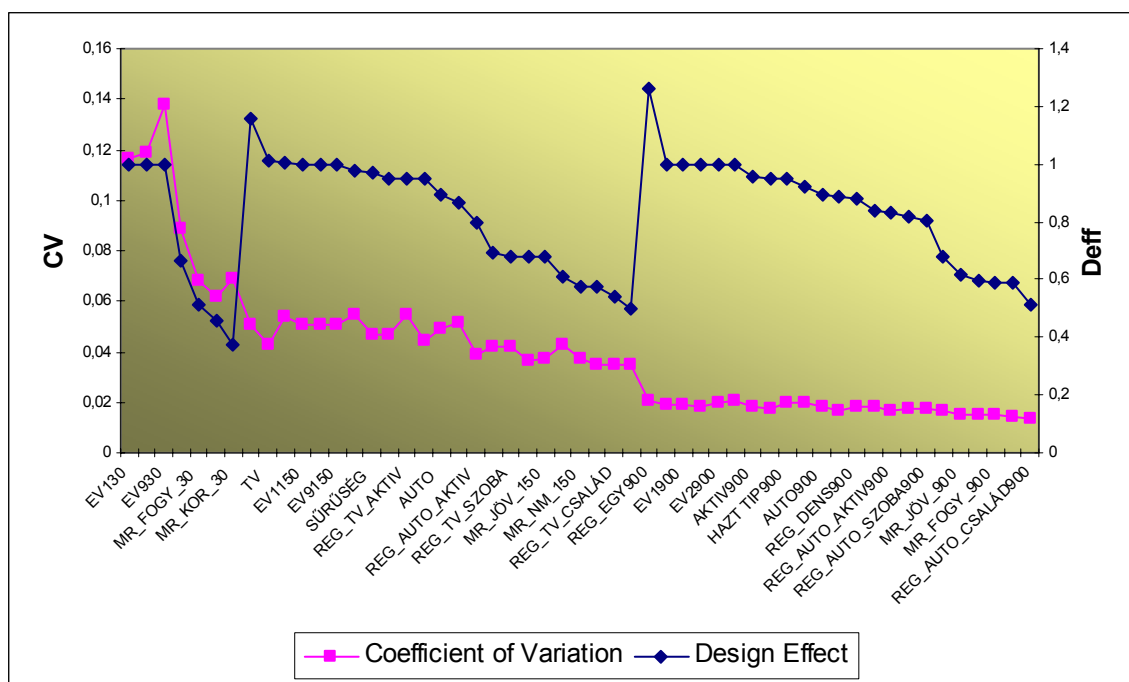
A fenti jellemzőket több változó mentén határoztam meg a várható érték becslése során azt vizsgálva, hogy a különböző eljárással választott minták között érvényesül-e valamilyen sorrendiség.

A háztartások egy főre eső fogyasztásának hányadosbecslése során kapott eredmények felhasználásával a minták vizsgált jellemzőit az 1. ábra mutatja be, amely alapján feltételezhető, hogy mind hatásosság, mind pontosság szempontjából relevánsabb eredményeket biztosítanak azok a részletesebb információk alapján készített minták, melyeknél a rétegzéshez több, a vizsgált változóval sztochasztikus viszonyban álló változó kerül bevonásra a mintavételi terv kidolgozásában. Az eredmények validálása után a következő alapozó tézis fogalmazható meg:

1. tézis

A szakirodalomból ismeretes, hogy a mintavételen alapuló kutatásokból származó megalapozott eredmények, valamint relatíve alacsony hibák előállítása és publikálása csak megfelelő adatbázisok alapján garantálható. Kutatásaim során megállapítottam, hogy Magyarországon a lakossági bázisú felmérések mintavételen alapuló kutatási eredményeinek javításához szükséges adatok (nem minden esetben hozzáférhetően) diszpergálva megtalálha-

tók a különböző hivatalok és szervezetek adatgyűjtéseinek köszönhetően. A kutatási eredmények hozadékanak javítása érdekében azonban ezek integrációjára van szükség.



1. ábra: A háztartások átlagos egy főre jutó fogyasztásának becslési jellemzői az effektív mintanagyság szerinti rangsorolásban

Az 1. tézis helytállóságát támasztja alá, hogy a mintavételen alapuló kutatásokban nem szokatlan jelenség a kiegészítő információk használata, melyeket elsősorban a mintavételi hiba meghatározására, a válaszadási arány növelésére, a torzítások feltárására alkalmaznak több-kevesebb sikerrel. Lásd Estevao – Särndal (2002.), Roy – Safiuzzaman (2006.) kétfázisú minták alkalmazására vonatkozó tanulmányiban.

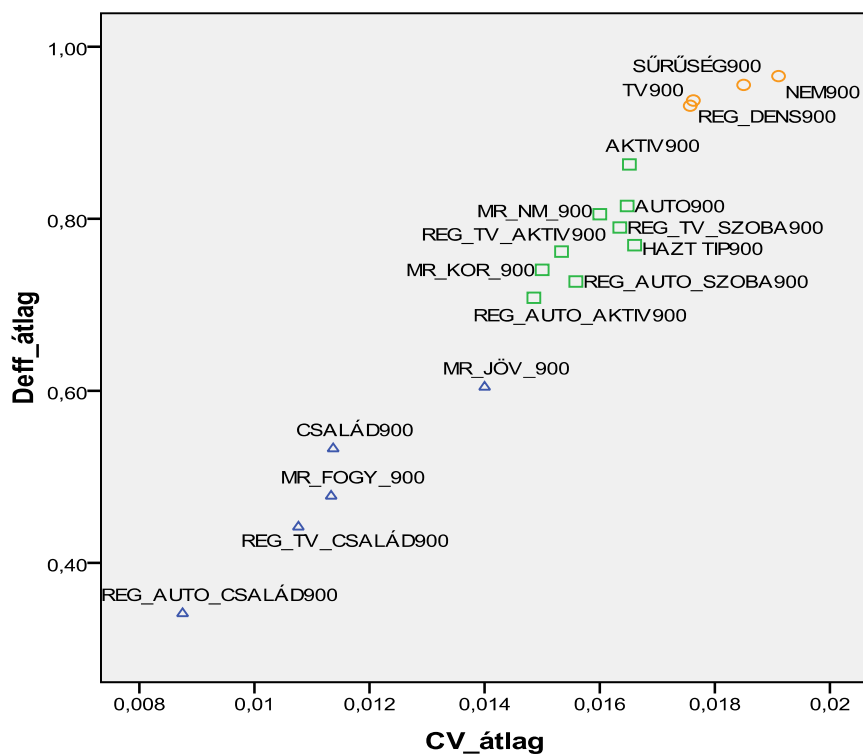
A nemválaszolás okozta torzítás kezelésében elengedhetetlen szerepet tulajdonít Särndal és Lundström (2008.) a kiegészítő információknak, kiemelve, hogy egyáltalán nem mindegy, milyen minőségű kiegészítő információk kerülnek felhasználásra.

Valamint a részosztályok, csoportok becslésére alkalmazott módszerekben is nagy segítséget nyújtottak a kiegészítő információk, Estevao – Särndal (2004.) alapján.

Az 1. ábra rávilágít arra a közismert feltevésre, mely a nagyobb méretű minták előnyös tulajdonságait hangoztatja. Kitűnik, hogy az effektív mintanagyság tekinthető vezérelvnek a minták rangsorolása során, ami biztosítja a másik két szempont szerinti rangsor alakulását is.

Annak érdekében, hogy a rangsorolás eredményeinek összehasonlíthatóságát, értelmezhetőségét és szemléltethetőségét javítsam, a minta nagyságának figyelmen kívül hagyásával folytattam az osztályozást. Bizonyítandó, hogy a fenti jellemzők azonos méretű minták esetében is képesek rangsort felállítani a különbö-

zó mintavételi tervek között. Az osztályozás során hierarchikus klaszter-analízist végeztem. Az elemzés dimenziót különböző változók Deff mutatóinak átlaga, valamint CV mutatóinak átlaga adta, melyben a 900 elemű minták a következő csoportokat képezték.



2. ábra: A 900 elemű minták klaszterei

A 2. ábrán látható, hogy amennyiben megszűnik a különböző méretű minták okozta szélsőséges ingadozás, a minták a két dimenzió között húzódó átló mentén helyezkednek el. Ami azt jelenti, hogy mindkét dimenzió arányosan jelentős szerepet játszik a klaszterek kialakításában. A Deff és a CV azért is tűnik jó párosításnak, mert, a relatív hiba és a mintaméret –a standard hibának és a minta elemszámának a sajátos fordított négyzetes viszonya okán – összefüggésben vannak egymással. A Deff viszont – éppen ellenkezőleg – a minta méretétől független eredményeket nyújt.

2. tétel

A szakirodalom alapján a matematikai, statisztikai módszerek demonstrálják a Deff mutató azon tulajdonságát, hogy képes megmutatni, mennyivel rosszabb, vagy jobb az adott mintavételi terv, egy ugyanolyan méretű egyszerű véletlen mintánál. Empirikus kutatásaim során ezen túlmenően azt is feltártam, hogy a Deff mutató más mutatókkal együtt képes az egyszerű véletlen mintánál jobbnak bizonyuló mintavételi tervek esetében hatékonysági rangsor elkészítésére.

Több vizsgált változón (egy főre jutó fogyasztási kiadás, háztartás mérete, háztartás összes kiadása) külön-külön végzett elemzések arra is lehetőséget adtak, hogy megvizsgáljuk, hogy a mintavételi tervben szereplő rétegeképző ismérvek vagy ismérveknek a vizsgált változóhoz fűződő sztochasztikus viszonya mutat-e valamilyen összefüggést a Deff és a CV által állított rangsorral. Megfogalmazható, hogy a rétegeképző ismérv vagy ismérvek korrelációs együttthatója – többszörös rétegzés esetén többszörös korrelációs együttthatója – determinisztikus viszonyban van a Deff és CV által kialakított rangsorral. Elmondható, hogy minden esetben a kevésbé hatékony minták klaszterébe kerültek a 0,4-nél kisebb korrelációs együttthatójú rétegeképző ismérvvel rendelkező minták.

A válaszadás hiányossága talán az egyik legnagyobb probléma, ami a felmérések készítésénél felmerül. A nemválaszolás hatásának vizsgálatára az egy háztartásra eső átlagos teljes fogyasztási kiadás értékére végeztem becsléseket több különböző mintavételi eljárás és mintavételi terv alapján. A számos minta közül az MR_FOGY-900 elnevezésű mesterséges információk alapján rétegzett minta bizonyult a leghatásosabbnak, melyben a teljes sokaság közel 10%-a került kiválasztásra a fogyasztási kiadások deciliseinek megfelelő arányos rétegekben. A minták közötti választásnál a hatásosság és pontosság szempontjainak érvényesítése volt az elsődleges, ennek megfelelően a Deff és CV mutatók értékei alapján történt a szelekció. A mintavételi terv hatásossága 0,27, a relatív standard hiba pedig 0,5% értéket mutatott.

Kísérleteim során vizsgáltam a válaszadási arány többféle mértékét. Egyértelműen bizonyítást nyert, hogy a nemválaszolás mértékének növekedése rontja a becslési eredményeket. Itt azonban nem kívántam összehasonlítani a különböző mértékű nemválaszolások esetén kapott eredményeket. Csupán egyetlen esetet emelek ki – egy, a gyakorlatban igen kedvezőnek számító 10% nemválaszolást tartalmazó mintát –, melyen bemutatom az imputálás „jótékony” hatását a különböző módszerek alapján.

Megvizsgáltam a hasonlóságon alapuló klasszifikációs módszerek alkalmazási lehetőségét a nemválaszoló egyedek (háztartások) azonosítására. A vizsgálatot klaszter-analízisre építettem, melyben első feladat volt olyan változókat találni, amelyek a hasonlóságot eredményezik, valamint sztochasztikus összefüggést mutatnak a háztartások fogyasztásával, és nem korrelálnak egymással.

Az elemzés alapját a következő változók biztosították:

- a háztartás mérete (fő),
- a lakóingatlan hasznos alapterülete (m^2),
- a háztartás tulajdonában levő autók száma (db),
- a háztartás tulajdonában levő televíziók száma (db),
- a háztartásfő jelenlegi foglalkoztatási státusza.

Az 1. táblázatból jól látható, hogy a fenti változók segítségével sikeresen azonosítottam a nemválaszolók táborát, hiszen 93,3%-uk abba a klaszterbe sorolható, amelyik jobb életkörülményeket mutat.

1. táblázat: Kontingencia tábla a klaszterhez tartozás és a nemválaszolás csoportosításához

<i>Klasztertagság</i>	<i>Mértékegységek</i>	<i>Válaszolók</i>	<i>Nemválaszolók</i>	<i>Összesen</i>
1. klaszter (jobb körülmények)	(fő)	531	84	615
	Megoszlás a klaszterek között (%)	65,6	93,3	68,3
2. klaszter (rosszabb körülmények)	(fő)	279	6	285
	Megoszlás a klaszterek között (%)	34,4	6,7	31,7
Összesen	(fő)	810	90	900
	Megoszlás a klaszterek között (%)	100,0	100,0	100,0

3. tézis

Abban az esetben, ha a nemválaszolók a vizsgált anyagi jellegű változók tekintetében (fogyasztás, jövedelem, stb.) hasonlóságot mutatnak, akkor egyéb, a vizsgált változóhoz kapcsolódó jellemzők, valamint további demográfiai, társadalmi, gazdasági szempontok szerinti hasonlóságok is detektálhatók a mintaegyedekben, ami lehetővé teszi a klasszifikációs módszerek alkalmazását a nem mintavételi hibák csökkentésére.

A klaszter-analízis eredményeinek felhasználása mellett regressziós összefüggésen alapuló, valamint hot deck imputációt is alkalmaztam. Utóbbiban a legközelebbi szomszéd módszere alapján történt a donorok kijelölése, ahol a hasonlóságot Mahalanobis távolság alapján mértem. A teljes fogyasztás átlagbecslését elvégeztem a teljes mintán, az imputálás nélküli adatokon, valamint a különböző imputált adatbázisokon.

Az 2. táblázat adatainak elemzéséhez tudni kell, hogy a sokasági paraméter, melynek becslésére vállalkoztam, jelen esetben ismert: 1.744.633,- Forint. Jól látható, hogy a teljes mintából megfelelően és viszonylag alacsony standard hibával becsülhető a sokasági paraméter. Abban az esetben viszont, amikor adathiány lépett fel, a becslés értéke jelentős mértékben több mint 9%-kal alulmúlta a sokasági értéket.

Az imputációk eredményeként kapott mintákból származó becslések javultak az adathiányos mintához képest, hiszen többségében a sokasági paramétertől kevesebb, mint 4,5%-kal kisebb értéket becsülnek.

A torzítás azonban továbbra is jelen van a becslésekben, hiszen a 95%-os megbízhatósági szint mellett számított konfidencia intervallumok nem fedik a sokasági paraméter értékét. Ilyen értelemben tehát nem tekinthetők sikeresnek az imputálási módszerek, hiszen jelentősen alulbecsülik a paramétert. Itt emlékez-

zünk arra, hogy mindössze 10%-os nemválaszolásról beszélünk, ami kedvezőnek számító körülmény.

2. táblázat: A háztartások átlagos fogyasztási kiadásainak becslése

<i>Módszer</i>	<i>Becslés</i>	<i>Standard hiba</i>	<i>Coefficient of Variation</i>	<i>Relatív eltérés a sokasági paramétertől</i>	<i>Az eltérés mértéke (%)</i>	<i>Konf: (0:igen), (1:nem)</i>	<i>Relatív standard hiba</i>
Teljes minta	1728151,47	8721,276	0,005	99,055%	0,945%	0	0,4999%
Nemválaszolt 10% a felső 30%-ból	1584815,61	7964,054	0,005	90,839%	9,161%	1	0,4565%
Imputált 5. iteráció	1668815,81	12595,015	0,008	95,654%	4,346%	1	0,7219%
Klaszterből imputált R becsléssel	1665379,74	10121,469	0,006	95,457%	4,543%	1	0,5801%
Imputált EV becsléssel	1668897,02	31730,657	0,019	95,659%	4,341%	1	1,8188%
Klaszterből imputált EV becsléssel	1665488,15	31896,867	0,019	95,464%	4,536%	1	1,8283%
Klaszterből a felső 10% imputálásával EV becsléssel	1716799,67	33390,132	0,019	98,405%	1,595%	0	1,9139%
Klaszterből a felső 10% imputálásával R becsléssel	1716685,51	8618,436	0,005	98,398%	1,602%	1	0,4940%
Mahalanobis távolság alapján imputált	1620586,09	10633,907	0,007	92,890%	7,11%	1	0,6095%

4. tézis

Nemválaszolás esetén az egyszerű¹ imputáció nem képes kellő pontossággal visszaadni az eredeti, teljes minta tulajdonságait, így a megbízható becslésre igen kicsi az esély, különösen, ha a vizsgált sokaság eloszlása nem szimmetrikus. Megállapítható, hogy egy baloldali aszimmetriát mutató sokaság lineáris statisztikáit becsülve az egyszerű imputált becslések torzítottan alulbecslik a sokasági paramétert.

A 4. tézis megállapításai alapján az imputáció alkalmazásán túl más eszközökre is szükség van a nemválaszolás okozta torzítás csökkentésére. Mivel a mintavételi terv minősége hatással van a becslés pontosságára, hatásosságára, megbízhatóságára,

¹ Egyszerű alatt itt azt értem, amikor egy imputációs módszert alkalmazunk egy változó mentén történő pótlásra nélkülözve a különböző módszerek kombinációját és egyéb összetett eljárásokat.

gára, ezért feltételezhető, hogy az alaposan, előrelátóan megtervezett mintavétel csökkenti a becslési eredmények nemválaszolás okozta torzításának mértékét.

A hipotézis ellenőrzéséhez a fogyasztási kiadások becslésére vonatkozó legjobb mintavételiterv-hatást és a legkisebb relatív hibát mutató MR_FOGY_900 mintát vizsgáltam. Alapvető célom volt, hogy a rétegeképző ismerv minél szorosabb sztochasztikus kapcsolatot mutasson a becslés tárgyát képező ismervvel. Esetemben ez az elvárás természetesen teljesül, mivel ($r=0,903^{**}$), a két változó szignifikáns kapcsolatát jelzi. Az elemzés feltételeinek ellenőrzése után a nemválaszolások generálását végeztem. Élve azzal a feltételezéssel, hogy a jövedelemmel, fogyasztással kapcsolatos kérdésekre jellemzően a magas ismervértékkel rendelkező egyedek nem válaszolnak. Az adathiányos egyedeket a fogyasztási kiadás-változó szerint csökkenő sorrendbe rendezett mintából választottam ki, kezdve egy igen kedvezőnek számító 10%-os nemválaszolási aránnyal. Ezt követően, lépésenként 5-5%-kal növeltem a nemválaszoló arányát egészen az 50%-os mértékű nemválaszolásig. Megállapításaim érvényességének biztosításához további, hasonló méretű mintákban is – többek között egyszerű véletlen kiválasztást és többszörös rétegzést alkalmazva – generáltam adathiányokat.

A különböző válaszadási szinteken, a mesterségesen rétegzett minta adta a legkisebb becsléseket, másként fogalmazva ez tartalmazza a legnagyobb torzítást. Az eredményekből egyértelműen következik az a tény is, hogy a nemválaszolás mértékének növekedésével az alulbecslés egyre drasztikusabb méreteket ölt, 50% válaszadás mellett akár 46%-kal is alábecsülhetjük a sokasági paramétert.

5/a. tézis

Abban az esetben, ha a nemválaszolás vélt vagy valós oka sztochasztikus összefüggést mutat a rétegeképző ismervvel, a rétegzés növeli a nemválaszolás okozta torzítás mértékét.

Az általam várttól eltérő eredmények magyarázatát a szórásnégyzet felbontása során találtam meg. Megállapítottam, hogy a különböző mintavételi tervek esetében a legnagyobb különbségek a külső eltérés négyzetösszegekben mutatkoztak. A belső eltérés négyzetösszegek viszonylagos stabilitása azzal magyarázható, hogy a fogyasztási decilisekben a háztartások fogyasztási kiadásai eleve jelentős különbségeket mutatnak. Mivel az MR_FOGY_900 minta esetében a rétegeképző és a nemválaszolást előidéző ismerv azonos, így a csoportok közötti eltérés négyzetösszeg (és annak aránya a teljes eltérés négyzetösszeghez viszonyítva) itt a legmagasabb. Ez eredményezi az erőteljesebb alulbecslést, hiszen előfordul, hogy teljes csoportok tartoznak a nemválaszoló táborába. Abban az esetben, ha teljes csoportok adatai hiányoznak, a többi csoport nagyobb szóródása mérsékli a negatív torzító hatást. A kontrollminták egy részében nincs rétegzés, a másik részük pedig kisebb sztochasztikus összefüggést mutat a fogyasztással (egyúttal a nemválaszolással), mint az MR_FOGY_900 minta.

A kutatónak a mintavétel megtervezésekor figyelembe kell vennie a vizsgált ismértvel kapcsolatos megtagadási várakozásokat, és ha azok potenciálisan teljes rétegeket érintenek, akkor érdemes lazítani a vizsgált ismértvel sztochasztikus kapcsolatban levő rétegeképző ismértvhez fűződő reprezentativitási követelményeken.

A nemválaszolás becslésére alkalmazható módszerek közül olyat szerettem volna választani, amely a lehető legkevesebb külső információ bevonását igényli, hiszen a kutatók számára ezek igen szűkösen állnak rendelkezésre. Többféle módszer alkalmazására tett kísérletről számol be Foster (1996.), melyekben a legsikereesebb eredményeket egyértelműen akkor érték el, ha a jelenséghez fűződő olyan változókat alkalmaztak, melyek egy censusból vagy mikrocensusból származnak. Dolgozatomban azokat a lehetőségeket kívánom feltárni, melyek elsősorban a vállalkozások, illetve a vállalkozások által alkalmazott kutatók számára biztosítanak megfelelő alapot a helyes becslések készítéséhez. Így a mintában fellelhető információkat felhasználva három klasszifikációs algoritmussal kísérleteztem:

- diszkriminancia-analízis,
- CHAID döntési fa,
- logisztikus regresszió.

Minden létrehozott modell kiértékelését a pontosság, az elsőfajú hiba és a másodfajú hiba mentén végeztem, így az eredmények könnyen összehasonlíthatóak. Az algoritmusok outputjai első megítélésre hasonlóknak tűntek, a részletesebb vizsgálat után azonban a döntési fa bizonyult kevésbé megbízhatónak, valamint a keresztérvényesség-vizsgálat eredményei is a CHAID módszer esetében voltak a legrosszabbak.

A diszkriminancia-analízis során a kovariancia mátrixok azonosságára vonatkozó hipotézist tesztelő Box's M teszt eredménye nem bizonyult szignifikánsnak. Így a logisztikus regresszió alkalmazása mellett döntöttem.

A logisztikus regresszió-függvény paramétereinek becslésekor különböző változókombinációkkal dolgozva, a magyarázó változók optimális összetétele a következő volt:

- HD14_02: autók száma a háztartásban,
- HA09: lakóhely népsűrűsége,
- HC08: iskolai végzettség,
- Jöv: jövedelemkategória.

A paraméterek szignifikancia vizsgálatából megállapítható, hogy a HC08: iskolai végzettség változó hatása nem minősül jelentősnek, ami egyértelműen annak tudható be, hogy az iskolai végzettség sem a fogyasztással sem a jövedelemmel nem mutat determinisztikus kapcsolatot. Emellett a népsűrűség szignifikancia szintje is meghaladja a társadalomtudományokban általánosan alkalmazott 5%-ot. Ennek ellenére szerepeltetem a modellben, mert hatását számos

nemválaszolást vizsgáló elemzésben kimutatták Varga (1999.), György (2004.), Johansson – Klevmarcken (2008.).

A logisztikus regresszió-függvény paramétereinek ismeretében a becsült feltételes valószínűségeket Varga (1999.) munkája alapján mintasúlyokká alakítom annak érdekében, hogy a potenciálisan nemválaszoló háztartások nagyobb súlyt kapjanak a fogyasztási kiadások becslése során.

Ezzel a súlyozási rendszerrel csökkenthető a fogyasztási kiadások alulbecslésének mértéke különböző nemválaszolási szinteken, ezt mutatja a következő táblázat.

3. táblázat: A fogyasztási kiadások átlagának becsült értéke (Ft) különböző nemválaszolási szinteken súlyozott és súlyozatlan adatokkal számolva

<i>Nemválaszolás mértéke</i>	<i>Súlyozatlan átlagos fogyasztási kiadás</i>		<i>Súlyozott átlagos fogyasztási kiadás</i>		<i>Az átlagok relatív eltérése súlyozatlan=100%</i>
	<i>Ft</i>	<i>A várható érték %-ában</i>	<i>Ft</i>	<i>A várható érték %-ában</i>	
TC a felső 10% nem válaszolt	1.475.398	84,57%	1.554.136	89,08%	105,3%
TC a felső 15% nem válaszolt	1.389.274	79,63%	1.494.852	85,68%	107,6%
TC a felső 20% nem válaszolt	1.316.725	75,47%	1.428.680	81,89%	108,5%
TC a felső 25% nem válaszolt	1.253.037	71,82%	1.371.626	78,62%	109,5%
TC a felső 30% nem válaszolt	1.193.976	68,44%	1.316.734	75,47%	110,3%
TC a felső 35% nem válaszolt	1.138.382	65,25%	1.266.237	72,58%	111,2%
TC a felső 40% nem válaszolt	1.084.923	62,19%	1.214.319	69,60%	111,9%
TC a felső 45% nem válaszolt	1.032.088	59,16%	1.172.525	67,21%	113,6%
TC a felső 50% nem válaszolt	979.840	56,16%	1.105.332	63,36%	112,8%

A 3. táblázat utolsó oszlopából látható, hogy a súlyozás hatására legalább 5%-kal sikerült javítani a nemválaszolás torzító hatásán. A modell és a súlyozás helyességét mutatja, hogy a nemválaszolási szintek növekedésével a súlyozási módszer torzítást csökkentő hatása egyre javul. Azonban nem felejthető el, hogy a gyakorlati kutatásoknak nem csupán az a célja, hogy enyhítsék a torzítás negatív hatásait, hanem a sokasági paraméter minél pontosabb és megbízhatóbb becslése. Ezt a célt azonban láthatóan csak részben sikerült teljesíteni. A sokasági érték ismeretében ugyanis látható, hogy a nemválaszolás mértékének szisztematikus növelésével a torzítás a súlyozás ellenére is drasztikus méreteket ölt.

6. tézis

A nemválaszolás valószínűségének becslésén alapuló átsúlyozás képes csökkenteni a nemválaszolás okozta torzítást, de a nemválaszolás szisztematikus növekedése esetében a torzítást csökkentő hatás lényegesen elmarad a tényleges torzítás mértékéhez képest.

Nyilvánvaló, hogy a kiegészítő információk alkalmazása segítséget nyújt a hibák mértékének csökkentésében. A kutatóknak azonban nem mindig van lehetősége

külső információk beépítésére, ezért a belső (mintabeli) információkat kell a lehető legnagyobb mértékben kiaknázni. A mintaegyedek megfelelő részletességű csoportosításával a válaszadói csoportokban megfigyelhető tendencia kivetíthető a teljes mintára, ezáltal a nemválaszoló egyedekre. A tendenciák modellezésével a nemválaszolás torzító hatása pedig csökkenthető. Érdeemes megvizsgálni a válaszadók és nemválaszolók vizsgált ismérvbeli tendenciái közötti különbséget. Az eljárás sikere érdekében megfelelő csoportokat kell kialakítani a mintában a vizsgált ismérvvvel sztochasztikus kapcsolatban álló és a nemválaszolást generáló ismérv(ek) alapján. A tendenciákat ezen csoportok mentén kell vizsgálni és modellezni.

Dolgozatomban a fogyasztási kiadás becsléséhez a háztartások jövedelme alapján képeztem csoportokat és a különböző jövedelmi tizedekbe eső háztartások fogyasztási kiadásaiban tapasztalható exponenciális tendenciákat azonosítottam, eltérő válaszadási arányok mellett. Tapasztalataim szerint a nemválaszolás alacsonyabb szintjein a függvények magyarázó ereje jobb volt, viszont a magasabb tizedekben jelentősen alulbecsülték a fogyasztási kiadások átlagát. Ezért elkészítettem a súlyozott tendenciák becslési modelljét melyben a fenti tendenciák becsült értékeit a függvények magyarázóerejének súlyozásával számított átlagos becsült értéként határoztam meg. Ezáltal az alacsonyabb nemválaszolási szinteken generált függvények (melyek egyre több réteg figyelembevételével lettek meghatározva, így pontosabbak is) nagyobb súllyal szerepelnek a végeredmények kialakításában.

Meghatározva az átlagos becsült értékek növekedésének mértékét a jövedelmi kategória növekedése mellett, adott felső réteg nemválaszolása esetén a válaszadó adatait a növekedés mértékével kiegészítve az átlagos fogyasztási kiadás minimálisan torzított becslését kapjuk.

A modell különböző nemválaszolási szintek mellett alkalmazható. Ezért a megvalósult válaszadási arány mellett a kutatónak mesterségesen kell további nemválaszolásokat generálni a csoportokban. A nemválaszolás generált mértékei közötti lépték természetesen változtatható azzal a kitéttel, hogy a lépték mértéke a nemválaszoló csoportok méretével arányos legyen.

A modell alkalmazását, a becslés menetét a mellékletben szemléltetem 30%-os nemválaszolás esetében. A táblázat 3-6. oszlopaiban található az exponenciális függvények által becsült átlagos fogyasztási kiadások a különböző mesterséges nemválaszolási mértékek mellett. A táblázat utolsó előtti sorában a függvények korrigált R^2 értékei találhatóak, melyek alapján az utolsó sor tartalmazza a kiszámított függvénysúlyokat. Minden jövedelmi tizedben a 3-6. oszlopban található becsült értékek és a megfelelő függvénysúlyok felhasználásával határozhatók meg a 7. oszlop adatai, az átlagos becsült fogyasztási értékek. Ezekből a tizedek közötti növekedés mértéke egyszerűen számítható.

7. tétel

Mesterséges nemválaszolási szintek generálásával, amennyiben a válaszadókra nézve jól kitapintható tendencia érvényesül, a súlyozott tendenciák becslési modellje segítségével hatásosabb becslés adható a sokasági paraméterre.

A súlyozott tendenciák becslési modellje az alábbi feltételezésekkel alkalmazható:

- léteznek olyan – a vizsgált tulajdonsággal összefüggő ismerv vagy ismérvek, melyek a nemválaszolást determinálják,
- ezen ismerv/ismérvek mentén a sokaság (lehetőleg egyforma méretű) csoportokba rendezhető,
- létezik a csoportok tendenciáit szignifikánsan leíró, megbízható matematikai függvény,
- a válaszadási arány nagyobb, mint 50%.

A feltételek teljesülésével a modell a háztartások átlagos fogyasztási kiadásainak relatíve jó közelítő értékét adja. Rendkívül előnyös tulajdonsága, hogy a jelentős mértékű alulbecslést, mely az imputálási, illetve átsúlyozási módszerek esetében tapasztalható képes ellensúlyozni. A modell sajnos a nemválaszolás magas szintjénél látszólag zavaró mértékű felülbecslést eredményezhet.

A modell nemválaszolás okozta torzításra gyakorolt csökkentő hatását mutatja be, illetve viszonyítja a súlyozásos eredményekhez a 4. táblázat.

4. táblázat: A nemválaszolást kezelő módszerek eredményeinek viszonyítása a sokasági paraméterhez

Nemválaszolás mértéke	Súlyozatlan átlagos fogyasztási kiadás		Súlyozott átlagos fogyasztási kiadás		Súlyozott tendenciák becslési modellje	
	Ft	A várható érték %-ában	Ft	A várható érték %-ában	Ft	A várható érték %-ában
TC a felső 10% nem válaszolt	1.475.398	84,57%	1.554.136	89,08%	1.684.073	96,53%
TC a felső 15% nem válaszolt	1.389.274	79,63%	1.494.852	85,68%	1.719.168	98,54%
TC a felső 20% nem válaszolt	1.316.725	75,47%	1.428.680	81,89%	1.721.414	98,67%
TC a felső 25% nem válaszolt	1.253.037	71,82%	1.371.626	78,62%	1.762.698	101,04%
TC a felső 30% nem válaszolt	1.193.976	68,44%	1.316.734	75,47%	1.767.559	101,31%
TC a felső 35% nem válaszolt	1.138.382	65,25%	1.266.237	72,58%	1.850.261	106,05%
TC a felső 40% nem válaszolt	1.084.923	62,19%	1.214.319	69,60%	1.858.925	106,55%
TC a felső 45% nem válaszolt	1.032.088	59,16%	1.172.525	67,21%	1.806.398	103,54%
TC a felső 50% nem válaszolt	979.840	56,16%	1.105.332	63,36%	1.826.144	104,67%

Míg a súlyozásos, illetve a nemválaszolások elhagyásával számított átlagos fogyasztási kiadások a magasabb nemválaszolási szinteken 40%-körüli torzítást is eredményezhetnek, addig a súlyozott tendenciák becslési modellje csupán 10%-körüli torzítást mutat.

A modell működésének ellenőrzésére azokat a lehetőségeket is elemeztem, ahol a nemválaszolás két oldalról csonkolja a mintát. Ebben az esetben a modell eredményeit a következő táblázat tartalmazza.

5. táblázat: A súlyozott tendenciák becslési modelljének eredményei a teljes fogyasztás becslésében, különböző nemválaszolási realizációkban

Nemválaszolás mértéke	Becslés nemválaszolások mellett		Súlyozott tendenciák becslési modellje	
	Becsült teljes fogyasztás (Ft)	A várható érték %-ában	Becsült teljes fogyasztás (Ft)	A várható érték %-ában
nv_f10	1.544.832	88,55%	1.675.538	96,04%
nv_f20	1.428.545	81,88%	1.693.421	97,06%
nv_f30	1.316.267	75,45%	1.706.211	97,80%
nv_a10	1.838.262	105,37%	1.734.666	99,43%
nv_a20	1.953.493	111,97%	1.756.714	100,69%
nv_f10a10	1.656.649	94,96%	1.687.891	96,75%
nv_f10a20	1.762.396	101,02%	1.708.690	97,94%
nv_f20a10	1.539.723	88,25%	1.698.880	97,38%
nv_f20a20	1.643.608	94,21%	1.715.116	98,31%
nv_f30a10	1.427.262	81,81%	1.699.990	97,44%
nv_f30a20	1.529.432	87,66%	1.708.190	97,91%

A táblázat rövidített jelölései némi magyarázatra szorulnak, miszerint „f” jelzi a felső jövedelmi tizedek nemválaszolását, „a” pedig az alsó jövedelmi tizedek nemválaszolását. Ez alapján nv_f20a10 azt jelenti, hogy a minta felső 20%-a és az alsó 10%-a nem válaszolt, tehát összességében 30%-os nemválaszolással realizálódott.

Amennyiben a nemválaszolás hatását nem veszem figyelembe, úgy 75%-tól 112%-ig változó mértékű alul és felülbecslés között ingadoznak a végeredmények.

Az utolsó oszlopból látható, hogy a súlyozott tendenciák becslési modelljének alkalmazásával a becslés értéket többnyire csupán néhány százalékos alulbecsléssel, stabilan közelítik a sokasági várható értéket. Azzal, hogy az alsó és felső értékekből is realizálódik nemválaszolás a modell kiegyensúlyozottabb lesz, aminek köszönhetően eltűnik az a szimmetrikus torzítás, ami a 4.táblázat utolsó oszlopában látható a nemválaszolási szint növekedésével. Ez azt igazolja, hogy a modell plasztikusan alkalmazható különböző nemválaszolási szinteken, és mértékek mellett.

Mindemellett a különböző módszerek együttes alkalmazása ajánlott, hiszen nem felejthető el, hogy adott minta csupán egy lehetséges realizációja a mintavételi tervnek, a vizsgált tulajdonság pedig valószínűségi változó, melyet a véletlenül kívül számos más tényező befolyásolhat, melyekre a fenti modellek külön-külön nem képesek megoldást nyújtani.

4. A KUTATÁS EREDMÉNYEINEK HASZNOSÍTÁSI TERÜLETEI

A mintára épülő vizsgálatok és következtetések egyre nagyobb szerepet kapnak a gazdasági döntések meghozatalában és az információ képzésben egyaránt. A mintavételek terjedését elsősorban a költségek és a vizsgálatához szükséges idő csökkentése indukálja. A mintán alapuló felmérések nem csak mikro szinten, hanem makrogazdasági vizsgálatoknál is egyre népszerűbbek, de a mintavételek terjedésének azonban nagy veszélye is van, pontosan a minta minősége miatt.

Éppen ezért úgy vélem, hogy az általam feltárt megállapítások és kidolgozott módszerek igen széles körben alkalmazhatók a mintán alapuló elemzések és következtetések hibáinak kiküszöbölésére, akár a tudomány elméleti, akár a vállalkozások gyakorlati területén.

A kutatási munkák elengedhetetlen része az adatgyűjtés, értékelés, s a megfelelő következtetések levonása. Az adatokból történő helytálló megállapítások megfogalmazásához nagyban hozzájárulhat, ha a kutatók figyelembe veszik, illetve felhasználják az általam kidolgozott módszereket.

Az oktatás területén fontosnak tartom, hogy ne csupán az egyes statisztikai számítási eljárásokkal ismertessük meg a hallgatókat, hanem – a fentebb vázolt módszerek figyelembe vételével – tanítsuk meg őket arra is, hogyan tudják szélesebb körben kiaknázni a minta adataiban rejlő információkat. Elért eredményeim eredményeim szerint arra ösztönzik a hallgatókat, ifjú kutatókat, hogy ne pusztán elfogadják a hiba mértékét a következtetéseikben, hanem törekedjenek annak csökkentésére is.

Tapasztalataim azt mutatják, hogy a közvélemény-kutatást végző szervezetek sincsenek mindig birtokában azon információknak, módszereknek, melyekkel a megszerzett adatokat helyesen tudnák feldolgozni, majd értelmezni. Sőt, gyakran már az adatfelvétel területén is problémák adódhatnak, ami elkerülhetetlenné teszi az egész felmérés téves, vagy túlzó megállapításait. Éppen ezért úgy vélem, hogy az ilyen jellegű tevékenységet folytató intézetek, vállalatok munkájának eredményessége jobbá válna, ha a kérdezőbiztosok felkészítésekor, a kérdőívek összeállításakor, s azok kiértékelésekor figyelembe vennék a kutatási munkámban megfogalmazottakat.

A súlyozott tendenciák becslési modellje nem csak megvalósult nemválaszolás esetén hasznosítható, hanem tervezett mintaelhagyás megvalósításával jelentős költségmegtakarításra nyújt lehetőséget.

Az eredmények széles körű felhasználását biztosítja, hogy olyan eljárásokat alkalmaztam, amelyek nemcsak a hivatalos statisztikák készítésénél, hanem a vállalkozások gyakorlatában is alkalmazhatók, valamint nagyobb anyagi ráfordítás nélkül biztosítanak megfelelő alapot a helyes következtetésekhez.

IRODALOMJEGYZÉK

Ay János – Vita László: Egy kísérleti jövedelmi felvétel főbb tapasztalatai; Statisztikai Szemle, 1998. 76. évf. 6. szám, pp. 515-532.

Estevao V. M. – Särndal C. E.: Borrowing strength is not the best technique within a wide class of design-consistent domain estimators; Journal of Official Statistics, Vol. 20, No. 4, 2004, pp. 645–669

Estevao V. M. – Särndal C. E.: The ten cases of auxiliary information for calibration in two-phase sampling; Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 233–255.

Foster, K.: Weighting the Family Expenditure Survey in Great Britain to compensate for non-response: an investigation using census-linked data. Helsinki. 1996.

György Erika: A nemválaszolás elemzése a munkaerő-felvételben; Statisztikai Szemle, 82. évf. 2004. 8. sz. pp. 747-772

Hajdu Ottó: Többváltozós statisztikai számítások; KSH, Budapest, 2003.

Havasi Éva – Schnell Lászlóné: Az 1996-os jövedelmi felvételre nem válaszoló háztartások – A megtagadások természete, a megtagadók sajátosságai; Központi Statisztikai Hivatal. Budapest. 1996.

Havasi Éva: Válaszmegtagadó háztartások; Statisztikai Szemle 1997. 10. sz. pp. 831-843.

Johansson F. – Klevmarcken A.: Explaining the size and nature of response in a survey on health status and economic standard; Journal of Official Statistics, Vol. 24, No. 3, 2008. pp. 431-449.

Keszthelyiné Rédei Mária: A lakossági jövedelmek mérésének megbízhatóbb módszere; Statisztikai Szemle, 2006. 84. évf. 5-6. szám pp. 518-551.

Kish, L: Kutatások statisztikai tervezése; Budapest, 1989.

KSH (1997): A háztartási költségvetési felvétel módszertana; Módszertani Füzetek 37. sz. KSH, Budapest.

KSH: A háztartások fogyasztása 2006, KSH, 2007.

Marton Ádám: A reprezentatív felvételek megbízhatósága 1991.

Roy D. – Safiuzzaman Md.: Variance estimation by Jackknife method under two-phase complex survey design; Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 35–51.

Sajtos László – Mitev Ariel: SPSS kutatási és adatelemzési kézikönyv; Alinea Kiadó, Budapest, 2007.

Särndal C. E. – Lundström S.: Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator; Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 167–191.

Varga Sára: A jövedelemfelvétel hiányzó adatainak pótlása; Statisztikai Szemle 1999. 77. évf. 2-3. sz. pp. 112-130.

A SZERZŐ TÉMÁHOZ KAPCSOLÓDÓ PUBLIKÁCIÓI

Szakcikk, könyvrészlet

Faktoranalízis In: Kvantitatív információképzési technikák Miskolci Egyetem, Elektronikus tananyag, 2011. . (megjelenés alatt)

Klaszteranalízis In: Kvantitatív információképzési technikák Miskolci Egyetem, Elektronikus tananyag, 2011. (megjelenés alatt)

A minta jellemzői; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 26-33. ISBN 978-963-06-7100-2

Szilágyi R. – Domán Cs.: Az adathiány kezelése mintavételes felmérésekben; Erdei Ferenc V. Tudományos konferencia – „Globális kihívások, lokális megoldások„ Kecskeméti Főiskola Kertészeti Főiskolai Kar Kecskemét, 2009. pp. 75-80. ISBN978-963-7294-74-7

Hipotézisvizsgálat; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 53-80. ISBN 978-963-06-7100-2

Szilágyi R. –Domán Cs.: Kalibráció a statisztikai becslésekben; „Gazdaság és társadalom” Nemzetközi tudományos konferencia Nyugat-magyarországi Egyetem Közgazdaságtudományi Kar Sopron, 2009. november 3. ISBN 978-963-9871-30-4

Mintavételi eljárások; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 9-25. ISBN 978-963-06-7100-2

Statisztikai becslés; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai II. Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2009. pp. 33-52. ISBN 978-963-06-7100-2

Szilágyi R.: The infiltration of the unfounded statistical information in the forming mechanism of competitiveness In.: XXII. microCAD International Scientific Conference 2009. Miskolc pp. 233-238. ISBN 978-963-661-881-0

Besenyey L. – Domán Cs. – Szilágyi R. – Varga B.: Faktoranalízis alkalmazásának lehetősége az innovációs potenciál mérése során; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 45-52, ISBN: 978-963-661-798-1

Grafikus ábrázolás; In.: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 58-73. ISBN 978-963-06-3135-8

Ismérvek közötti sztochasztikus kapcsolatok elemzése; In.: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 140-153. ISBN 978-963-06-3135-8

Besenyey L. – Domán Cs. – Szilágyi R. – Varga B.: Klaszteranalízis alkalmazásának lehetősége az innovációs potenciál mérése során; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 53-64, ISBN: 978-963-661-798-1

Szilágyi R.: Mintavételes eljárások; Oktatási segédlet Miskolci Egyetem, 2007.

Szilágyi R.: Pénzbeli ellátások beilleszkedési kölcsönhatásai; In: „Globális és hazai problémák tegnapról holnapig”, VI. Magyar (Jubileumi) Jövőkutatási Konferencia, 30 éves az MTA IX. Osztály Jövőkutatási Bizottsága, Konferenciakötet 2., Budapest, Arisztotelész Stúdium Bt., 2007, pp. 91-97, ISBN: 978-963-86670-8-3

Besenyey L. – Domán Cs. – Szilágyi R. – Varga B.: „Statisztikai mintaillesztés” program tervezése és megvalósítása; In.: Innovációmenedzsment kutatás és gyakorlat; Miskolc, Miskolci Egyetem Innovációmenedzsment Kooperációs Kutatási Központ, 2007, pp. 8-16, ISBN: 978-963-661-798-1

Szilágyi R.: Statisztika az üzleti életben In: Informatikai statisztikus és gazdasági tervező felsőfokú képzés II. kötet 6. fejezet HEFOP-3.2.2-P.-2004-10-0011-/1.0 sz. projekt, Miskolc, 2007.

Viszonyszámok; In: Domán Cs. – Szilágyi R. – Varga B.: Statisztikai elemzések alapjai Közgazdasági-módszertani képzés fejlesztéséért Alapítvány, 2007. pp. 42-57. ISBN 978-963-06-3135-8

Szilágyi R.: Kontár statisztikák; In: Doktoranduszok Fóruma Gazdaságtudományi Kar Szekciókiadványa, Miskolc, Miskolci Egyetem Innovációs és Technológia Transzfer Centrum, 2006, pp. 168-172.

Szilágyi R.: Minőségügyi statisztika; Oktatási segédlet Miskolci Egyetem, 2006.

Konferencia előadás

Szilágyi R.: A nemválaszolás torzításának becslése a mintavételes felmérésekben; „HITEL, VILÁG, STÁDIUM” Tudományos konferencia, Sopron 2010. november 3.

Szilágyi R.: Analysis of nonresponse; International Conference “Economic & Social Challenges and Problems, at The time of Crisis 2009” Faculty of Economy, University of Tirana, Albania, 2009.

Besenyei L. – Varga B. – Domán Cs. – Szilágyi R.: Az elemezhetőséget biztosító mintaillesztés megvalósítása, Innovációmenedzsment, Tudásteremtés – Tudás-transzfer Konferencia, Miskolc, 2006. november 15-16.

MELLÉKLET

*A súlyozott tendenciák becslési modellje
30%-os ténylegesen megvalósult nemválaszolás esetében*

<i>Jövedelmi tizedek</i>	<i>Fogyasztás a teljes válaszadás mellett</i>	<i>30%NV</i>	<i>35%NV</i>	<i>40%NV</i>	<i>45%NV</i>	<i>50%NV</i>	<i>Átlagos becsült függvény- érték</i>	<i>A tizedek becsült átlagai</i>
1	650.298	749.355	742.179	734.640	731.450	722.461	736.179	650.298
2	916.414	887.189	883.780	880.178	879.221	876.507	881.437	916.414
3	1.170.972	1.050.374	1.052.398	1.054.549	1.056.846	1.063.400	1.055.428	1.170.972
4	1.418.208	1.243.575	1.253.186	1.263.464	1.270.355	1.290.142	1.263.851	1.418.208
5	1.374.019	1.472.313	1.492.283	1.513.767	1.526.999	1.565.231	1.513.536	1.374.019
6	1.739.427	1.743.123	1.776.998	1.813.658	1.835.492	1.898.976	1.812.674	1.739.427
7	1.944.533	2.063.746	2.116.034	2.172.959	2.206.308	2.303.883	2.171.083	1.944.533
8	2.214.489	2.443.342	2.519.755	2.603.441	2.652.038	2.795.126	2.600.538	2.329.175
9	2.475.128	2.892.759	3.000.503	3.119.205	3.187.817	3.391.114	3.115.157	2.790.094
10	3.291.167	3.424.840	3.572.973	3.737.147	3.831.836	4.114.181	3.731.872	3.342.456
átlag	1.719.465	1.797.062	1.841.009	1.889.301	1.917.836	2.002.102	1.888.176	1.767.559
$\overline{R^2}$		0,9317	0,9408	0,9167	0,9218	0,8856		
függvény súlyok	1	0,20269	0,20467	0,19943	0,20054	0,19266		