6. AZ ÖKOLÓGIAI TÉVKÖVETKEZTETÉS

6.1. Egyéni viselkedés és csoportosított adatok: a probléma lényege

Az ökológiai tévkövetkeztetés problémája az aggregációs információveszteség miatt bekövetkező és ezért elvileg megoldhatatlan statisztikai módszertani kérdés, melynek alapvető volta indokolja külön fejezetben történő tárgyalását. A kérdéskör elsősorban a szociológusok és a választásokkal foglalkozó politológusok érdeklődését vívta ki, mert ezeken a területeken gyakran fordul elő, hogy csak területileg csoportosított adatok állnak rendelkezésre valamilyen, emberek tulajdonságára vagy viselkedésére vonatkozó információkból. Egyéni szintről csoportosított, de egyéni szinten különböző okok (mint az adatvédelem, adatkezelés bonyolultsága, egymástól független adatforrások, adathalmaz nagysága, adatvesztés, történeti információk) miatt elérhetetlen adatokkal azonban a társadalomtudomány minden területén találkozunk.

Robinson alapvető, 1950-ben megjelent cikkének definíciója szerint *az ökológiai korreláció tárgyát egyének csoportjai, az egyéni korreláció tárgyát egyének képezik (Robinson, 1950).* Az ökológiai korreláció elnevezés a rendszereket, sokaságokat vizsgáló humán ökológiára vezethető vissza. Bár az elnevezés megtévesztő lehet – hasonlóan például a regressziószámítás szó szerinti értelmezéséhez – ez a terminus azonban mégis általánosan elterjedtté vált a csoportosított adatok kérdéseivel foglalkozó szakirodalomban. Ugyanakkor Robinson eredeti definíciójához képest általában szélesebb értelemben használják ezt a kifejezést, és bármilyen típusú csoportosított adatból számított korrelációt értenek alatta, függetlenül attól, hogy az illető korrelációnak logikailag létezhet-e individuális párja vagy sem. Ez a nehezen elkerülhető, kisebb értelmezési bizonytalanság többnyire nem okoz gondot, mert a szerzők konkrét példáiból általában kiderül, mire gondolnak. Szükség esetén én is mindig pontosítani fogom, hogy a megállapítások az ökológiai korreláció melyik válfajára vonatkoznak.

Az Egyesült Államokban tagállami szinten a fekete népesség[1] aránya és az írástudatlanság aránya között számított korreláció például ökológiai korreláció. Ekkor a változók százalékos arányok lesznek, a sokaságokat, nem pedig az egyéneket jellemző tulajdonságok *(Robinson, 1950).* Az egyének csoportjaira vonatkozó ökológiai korrelációkból nem lehet következtetni az ugyanazon alapadatok alapján számított egyéni korrelációkra, és fordítva, az egyéni korrelációkból nem lehet következtetni az ökológiai korrelációkra. Az ökológiai korrelá-

ció nem alkalmas az egyéni korrelációk helyettesítésére, a két érték különbözik egymástól.

Az ökológiai tévkövetkeztetés problémájával kapcsolatban az egyéni és csoportosított adatok elérhetőségének három alapesetét célszerű megkülönböztetni:

- Ismertek az egyéni adatok, amikből tetszőleges módon lehet csoportosított adatokat létrehozni.
- Nem ismertek az egyéni adatok, és csak egyetlen összesített csoportosított adat ismert.
- Nem ismertek az egyéni adatok, és az összesített csoportosított adat mellett több részcsoport adata is ismert.[2]

Az ökológiai tévkövetkeztetés problémája az első két esetben nem lép fel. Első esetben érdeklődési körünknek megfelelő szinten, egyéni és kollektív adatokkal is vizsgálhatjuk a minket érdeklő jelenséget. Második esetben nincs módunk a két jellemző közötti kapcsolatra vonatkozó semmilyen következtetés levonására a meglévő adatainkból. Harmadik esetben fennáll az ökológiai tévkövetkeztetés lehetősége, hogyha a csoportosított adatokból (egyéb információ hiányában) következtetünk az egyéni viselkedésre. Ekkor például helytelenül így fogalmazhatjuk meg Robinson példájában a jellemzők közötti kapcsolatot: „a négerek között nagyobb az írástudatlanok aránya, mint a fehérek között". A helyes megfogalmazás így szólna: „azokban az államokban, amelyekben az átlagosnál nagyobb a néger lakosság aránya, az írástudatlanság aránya is nagyobb az átlagosnál". Ebben az illusztratív célokat szolgáló példában az eredeti adatok ismertek, a tévkövetkeztetés lehetősége tehát nem állt fenn.

Az alábbiakban az alapprobléma bemutatásánál nem Robinson példáját használom fel, hanem a realisztikusabb választási eredményeket, mivel ezeknél a szavazatok titkossága miatt az eredeti adatokat nem ismerhetjük. A legegyszerűbb esetet a következőképpen fogalmazhatjuk meg. Az alapadatok három osztályát ismerjük: egynél több körzetet, valamilyen társadalmi megoszlást és a szavazatok megoszlását körzetenként (22. táblázat). Csak a legegyszerűbb lehetőséggel foglalkozom, amikor mindkét változó dichotóm, de ez nem korlátozza a leírás érvényességét erre a speciális esetre, hiszen a nem dichotóm változóknál a nehézségek úgyis hatványozódnak. A sorok és oszlopok összesen adatait, a peremgyakoriságokat ismerjük, tudjuk külön a nők és a férfiak számát, valamint a fehér pártra és zöld pártra leadott összes szavazatot. A 23. táblázatban két körzetre vonatkozóan hipotetikus adatokat tartalmazó eredményeket láthatunk. A cellákban feltüntettem az egyes celláknak a peremgyakoriságokból következő, lehetséges minimális és maximális gyakoriságait. A problémát az jelenti, hogy a belső cellagyakoriságok, amelyek érdekes információkkal szolgálhatnának, ezen tág határok között bármilyen értéket felvehetnek.

22. táblázat Az adatok kereszttáblája
(The crosstabulation of data)

Körzet	Nemek szerinti megoszlás		Szavazatok megoszlása	
	Nők (1)	Férfiak (2)	Fehér párt (1)	Zöld párt (2)
1	X_{11}	X_{21}	Y_{11}	Y_{21}
2	X_{12}	X_{22}	Y_{12}	Y_{22}
…	…	…	…	…
N	X_{1n}	X_{2n}	Y_{1n}	Y_{2n}
Összesen	X_1	X_2	Y_1	Y_2

23. táblázat Hipotetikus adatok kereszttáblája a peremgyakoriságokból következő minimális és maximális cellagyakoriságok feltüntetésével
(Hipothetical data with the minimum and maximum cell frequencies)

1. körzet	Fehér párt szavazatai	Zöld párt szavazatai	Összesen
Nők	Minimum: 0 Maximum: 1000	Minimum: 0 Maximum: 1000	1000
Férfiak	Minimum: 0 Maximum: 1000	Minimum: 0 Maximum: 1000	1000
Összesen	1000	1000	2000

2. körzet	Fehér párt szavazatai	Zöld párt szavazatai	Összesen
Nők	Minimum: 0 Maximum: 700	Minimum: 500 Maximum: 1200	1200
Férfiak	Minimum: 0 Maximum: 700	Minimum: 100 Maximum: 800	800
Összesen	700	1300	2000

A csoportosított adatok használatára nemcsak az adatvédelem miatt kényszerülnek rá gyakran a kutatók, hanem azért is, mert külön adatforrásból származó eredményekkel tudnak csak dolgozni, amelyeknek az összekapcsolása egyéni szinten lehetetlen. Az emberekre vonatkozó egészségügyi, jövedelmi, képzettségi, demográfiai, igazságszolgáltatási, idegenforgalmi, munkaerőpiaci, kulturális, lakáshelyzetre vonatkozó adatokhoz többnyire külön-külön, területileg csoportosított formában jutnak hozzá, és nem egyéni szinten. A piackutatás során például gyakran előfordul, hogy ismertek egy termék körzetenkénti eladási forgalmai, és ettől függetlenül ismertek a körzetek demográfiai jellemzői. Ezek az ismeretek önmagukban is fontos szerepet játszanak a piaci stratégia alakításában. Az öko-

lógiai tévkövetkeztetés lehetősége azonban fennáll, hogyha ezekből a területi adatokból a termék vásárlóinak egyéni jellemzőire következtetünk. Ha lehetőség nyílik a vásárlók egyéni megfigyelésére, akkor módunk nyílik ilyen ismeretek szerzésére is.

A területi egységek szintjén rendelkezésre álló többi adat szintén fontos információkat szolgáltat a körzetek lakosságának együttes jellemzőiről, a körzeteknek a gazdasági, szociológiai, politikai, kulturális életéről. Különféle háttérinformációk, egyéb, egyedi felmérések ismeretében korlátozottan felhasználhatók az aggregált adatok is az egyedi viselkedésre vonatkozó érveléshez. A szavazati eredményeknek például egyes társadalmi csoportok, nemek, életkor, végzettség szerinti megoszlására az exit poll felmérésekből és a kérdőíves megkérdezésekből lehet következtetni, az ilyen eljárások során ismert módszertani korlátok tekintetbe vételével. Ezeket az eredményeket össze lehet vetni az ökológiai korrelációk eredményeivel.

Az ökológiai korrelációval kapcsolatos egyes kérdésekre már 1934-ben rámutatott Nephrash a népszámlálási adatok elemzése során. Megállapítása szerint a területileg csoportosított jellemzők közötti korrelációk nem a változók közötti kapcsolat erősségét mérik, hanem a földrajzi eloszlásuk közötti hasonlóságot vagy különbözőséget mutatják ki. További megszorításként a korrelációs együtthatókat csak akkor értelmezhetjük helyesen, hogyha a területi egységek egyenlő méretűek, vagyis egyenlő a területi kiterjedésük, amennyiben a területi eloszlásuk képezi a vizsgálat tárgyát, illetve egyenlő népességszámúak, amennyiben az emberek közötti gyakoriság a kérdéses. A körzetek méretének a csökkenésével növekszik a korrelációs együtthatók magyarázó ereje, egészen addig a szintig, ami alatt már nem csökkenthető a körzetek mérete. Külön problémát jelent az a kérdés, hogy a nagy méretű heterogén területeken számított korrelációs együtthatók mekkora jelentőséggel bírnak. A heterogén területek egy részén ugyanis jelentős korrelációt mérhetünk, más részén ugyanakkor pedig teljes korrelálatlanságot vagy ellentétes előjelű korrelációt lehet kimutatni[3] *(Neprash, 1934)*.

A területegységek összevonásának hatását Gehlke és Biehl a fiatalkorú fiúk által elkövetett bűntettek száma és az átlagos havi jövedelem között mért korreláción keresztül illusztrálta. Cleveland 252 népszámlálási körzetét a területi folytonosság figyelembe vételével egyre nagyobb területegységekbe összevonva a 24. táblázatban látható eredményeket kapták. A körzetek véletlenszerű összevonásakor a 25 körzetes együttható -0,544 lesz. Az eredmények Gehlke és Biehl szerint kérdésessé teszik, hogy ezeknek a korrelációknak bármiféle szerepük is lehetne az oksági magyarázatok során, és hogy a jellemzők eredeti tulajdonosaira, az egyénekre és családokra is igazak legyenek *(Gehlke–Biehl, 1934)*.

24. táblázat A fiatalkorú fiúk által elkövetett bűntettek és az átlagos havi jövedelem közötti korreláció
(Correlation between male juvenile delinquency and the median equivalent rental)

Területegységek száma	Korreláció az abszolút adatok között	Korreláció a fajlagos adatok között
252	-0,502	-0,516
200	-0,569	-0,504
175	-0,580	-0,480
150	-0,606	-0,475
125	-0,662	-0,563
100	-0,667	-0,524
50	-0,685	-0,579
25	-0,763	-0,621

Forrás: Gehlke–Biehl (1934)

A csoportosított adatokból az egyéni adatokra való következtetés veszélyeit a választási eredmények értékelése kapcsán is felfedezték. Ogburn és Goltra 1919-ben az újonnan választójogot nyert nők szavazási viselkedését próbálták meg kideríteni egy népszavazáson. A nők szavazataránya és a nem szavazatok aránya közötti pozitív korreláció alapján úgy gondolták, hogy a nők körében magasabb volt a nemmel szavazók aránya. Ezzel a következtetésükkel szemben azonban fenntartásaik voltak, mivel felismerték, hogy a választókörzetek határainak átszabásával negatív korrelációt is kaphattak volna *(King, 1997).*

6.2. Robinson hozzájárulása az ökológiai korreláció témaköréhez

Az ökológiai korrelációval kapcsolatban írt leghatásosabb, legbefolyásosabb, legtöbbet idézett tanulmány William Robinson tollából született meg 1950-ben. A cikk sikeréhez valószínűleg hozzájárultak egyértelmű megfogalmazásai, a probléma általános kezelése, a hatásos illusztrációk és a rendkívül világos okfejtések. Robinson rámutatott arra, hogy számos, akár klasszikusnak számító tanulmányban használnak teljesen tévesen ökológiai korrelációt az egyéni viselkedés leírására. Ökológiai korrelációt ekkor nem azért alkalmaznak, mert a csoportok tulajdonságai iránt érdeklődnek, hanem mert az egyéni korrelációk számítására adatok hiányában nincsen lehetőség *(Robinson, 1950).*

Ugyanannak az ökológiai korrelációnak számos teljesen eltérő egyedi korreláció felelhet meg, és viszont, az egyéni korrelációk is számos különböző ökológiai korrelációt eredményezhetnek az eltérő csoportosításoknak köszönhetően.

Robinson rámutat arra is, hogy az ökológiai korrelációt lehet számítani súlyozott és súlyozatlan adatokkal is, a kettő közül a súlyozott számítás tekinthető korrektebbnek.

Robinson két példán keresztül illusztrálja a problémát. A feketék és az írástudatlanok aránya közötti ökológiai korreláció mértéke 1930-ban az USA tagállamok szintjén 0,773, a népszámlálás kilenc országrészének szintjén 0,946, az egyéni korreláció pedig 0,203, ami töredéke az ökológiai korrelációknak. De ebben az esetben legalább az előjel változatlan maradt, nem úgy, mint a külföldön született amerikaiak és az írástudatlanság közötti korreláció esetében, amikor az ökológiai korreláció -0,526 és -0,619 (állami és országrész szinten), az egyéni korreláció viszont 0,118. Vagyis, mint az az adatokra vonatkozó kereszttáblákból is látszik, a külföldön született amerikaiak között magasabb arányban találunk írástudatlanokat, mint az USA-ban születetteknél, az ökológiai korreláció viszont pont ennek az ellenkezőjét sugallaná.

Összefoglalásként Robinson megállapítja, hogy az ökológiai korreláció nem alkalmas az egyéni korrelációk helyettesítésére. Bár elméletileg egybeeshet a két érték, ennek feltételei gyakorlatilag nem szoktak előfordulni. Robinson célja az volt, hogy az egyéni viselkedés vizsgálata kapcsán a jelentés nélküli ökológia korrelációk használatát kerüljék el, és használjanak egyéni korrelációkat *(Robinson, 1950).*

Robinson tanulmánya számos hatást gyakorolt a további kutatásokra. A Robinsont követően a témával foglalkozók szinte egyöntetűen egyetértenek abban, hogy az ökológiai korrelációk nem helyettesíthetik az egyéni korrelációkat, bár néhány kutató sikertelenül próbálta meg az ökológiai tévkövetkeztetés lehetőségét cáfolni *(King, 1997).* Az egyéni viselkedés leírására szolgáló ökológiai korrelációk használata visszaszorult, de nem tűnt el teljesen. A politológusok és szociológusok jelentős része, ha módja nyílt rá, inkább egyéni kérdőíves felméréseket kezdett el használni olyan kérdések vizsgálatára, amelyekhez korábban területileg aggregált adatokat használt fel *(King, 1997).*

A kutatások olyan irányokban kezdődtek meg, amely kérdésekre Robinsonnak egy tanulmány keretein belül nem terjedhetett ki a figyelme. Az egyik fő irányt az ökológiai korrelációk értelmezése, interpretálása, individuális korrelációkkal való tartalmi kapcsolatának elemzése jelentette. A másik kutatási terület az aggregált adatokból az egyéni adatokra való következtetés statisztikai módszertani, matematikai lehetőségének a vizsgálata. Végül empirikus felmérések az ökológiai tévkövetkeztetés jelentőségét ismert adatbázisokon alapuló számítások segítségével próbálták meg bemutatni. Egyes tanulmányokon belül a fenti megközelítések gyakran együtt fordultak elő, de különbözőségük miatt három külön alfejezetben foglalkozom velük.

6.3. Az ökológiai korrelációk értelmezése

Az ökológiai korrelációk értelmezésének és hasznosításának kérdésével először Menzel foglalkozott Robinson tanulmányához fűzött megjegyzésében. Ebben nem értett egyet Robinsonnak azzal a megállapításával, miszerint „minden olyan munkában, amely ökológiai korrelációra épül (…) csak azért használnak ökológiai korrelációkat, mert individuális tulajdonságokra vonatkozó korrelációk nem állnak rendelkezésre" *(Robinson, 1950; idézi Menzel, 1980, 466. o.)*. Menzel egyik példája szerint a börtönbüntetések és a válások száma között fennálló, szoros ökológiai korrelációból tévesen le lehet vonni azt a következtetést, hogy a börtönbüntetésre ítélt egyének különösen hajlamosak a válásra. De gyakoribb az az eset, amikor az ilyen jellegű ökológiai korrelációkat úgy használják fel az érvelésben, hogy „a börtönbüntetések és a válások száma is olyan más, a két jelenség mögött meghúzódó, közös ok függvénye, amelynek forrását nem az egyének, mint olyanok, hanem egyének közti különbségek és kapcsolatok – „a területi egységek, mint olyanok tulajdonságai" – alkotják. Ez utóbbiakat többnyire kulturális konfliktusnak, társadalmi anómiának, vagy valami hasonlónak nevezik. Aligha kell külön kiemelni, hogy az a kutató, aki a nők bírósági eseteinek száma és a fiatalkorú férfiak bírósági eseteinek száma között állapít meg korrelációt, ezt nem azért teszi, hogy a szóban forgó korrelációból arra következtessen, hogy a női bíróságok előtt megjelenő nőknek különösen nagy esélyük van arra, hogy a fiatalkorú férfiak ügyeit tárgyaló bíróságok előtt is megjelenjenek" *(Menzel, 1980, 466. o.)*.

Menzel harmadik példája a zsidó népesség százalékos részaránya és az antiszemitizmus mértékére vonatkozó skálán magas értékkel bírók részaránya közötti korrelációra vonatkozik. A két jellemző közötti pozitív kapcsolatból hibás lenne arra a következtetésre jutnunk, hogy a zsidók többsége antiszemita. Menzel utolsó példája, az orvosok száma és a csecsemőhalandóság közötti összefüggés sem ekvivalens Robinson példáival, mert azoknál az ökológiai korrelációnak létezett egyértelmű, „tiszta" individuális korrelációs megfelelője. Menzel példáinál – a büntetett előélet és az elvált családi állapot esetét leszámítva – azért nem beszélhetünk individuális korrelációkról, mert a két vizsgált tulajdonság egyszerre nem jellemezheti ugyanazt az egyént *(Boudon, 1980)*. Ezért csak csoportokra vonatkoztatott részarányuk közötti korreláció kiszámítása lehetséges.

Az ökológiai korreláció értelmének kibővülése így már Menzel értékes észrevételei kapcsán elkezdődött. Boudon a csoportokra jellemző változók három típusát különböztette meg Lazarsfeld és Menzel nyomán: analitikus, strukturális és globális változókat.[4] Az analitikus változók az egyéni tulajdonságok átlagai, a strukturális változók megoszlási viszonyszámok, a globális változók két eltérő ismérvből képzett intenzitási viszonyszámok. Az ökológiai tévkövetkeztetés

lehetősége Boudon szerint az analitikus változók között léphet fel, mert ezeknek az adatoknak vannak olyan egyéni megfelelőik, amelyek között lehet individuális korrelációt számolni. De a Menzel példái kapcsán említett okok miatt még ezek között sem mindig lehet egyéni korrelációt számolni. Az ökológiai korrelációk elemzésben betöltött szerepe azonban ekkor is jelentős lehet *(Boudon, 1980).*

Az írástudatlanság és a bőrszín közötti gyenge individuális, és szoros kollektív korreláció tényét például többféleképpen lehet értelmezni. Először is, lehet a csoportosítási hatásból fakadó látszólagos korrelációnak tekinteni, amelynek nincs saját, külön jelentése. Másodszor értelmezhetjük úgy, hogy a négerek iskolázottsága attól függ, hogy mekkora a számarányuk az adott államban. Az ökológiai korreláció ekkor tartalmaz egy individuális hatást, amely abból áll, hogy a négereknek kisebb az esélyük írni-olvasni megtanulni, és egy kollektív hatást, amely szerint minél nagyobb a négerek számaránya, annál inkább elhanyagolják beiskolázásukat. A harmadik értelmezés szerint „bizonyos államokat azonos történelmi körülmények és a gazdasági fejlődésből fakadó azonos problémák kényszeríttettek mind arra, hogy (a) jelentős számú néger munkaerőt importáljanak és tartsanak meg az állam határain belül, mind arra, hogy (b) elhanyagolják iskolarendszerüket" *(Boudon, 1980, 468. o.).* A második értelmezésben az individuális és kollektív hatás együttesen eredményezi a magasabb ökológiai korrelációt, a harmadik értelmezésben az ökológiai korrelációt közvetlenül magyarázzuk egy globális változóval. Boudon szerint számos esetben lehetőség van az értelmezések közötti választásra. Hogy a kutatók melyiket fogadják el, azt a változók közötti kapcsolat alapos vizsgálatával lehet eldönteni *(Boudon, 1980).*

Boudon a kontextuális hatások – más elnevezésekkel strukturális hatás, összetétel hatás, a szociológia klasszikusai közül Durkheimnél kollektív tudat, Tardenál utánzás, Tönniesnél közösség – többféle típusát különbözteti meg az alapján, hogy a kontextuális hatás milyen kapcsolatban áll az individuális hatással. A 25. táblázatban látható hipotetikus példájában egy individuális és egy kontextuális hatás is befolyásolja a szavazatok százalékarányának alakulását *(Boudon, 1987).* A kontextuális hatások elemzése a szociológiában Durkheimig vezethető vissza, aki azzal a nehézséggel nézett szembe, hogy az öngyilkosságokra és a vallási hovatartozásra vonatkozó statisztikai adatok csak területi csoportosításban álltak rendelkezésére. Azzal a körülménnyel, hogy a katolikus országokban kisebb az öngyilkossági arány, négyféle következtetés egyeztethető össze[5] *(Babbie, 1996; Davis–Spaeth–Houson, 1987; Moksony, 2002).*

25. táblázat Hogyan változik a kommunistákra leadott szavazatok százalékaránya egy individuális és egy kontextuális változó függvényében?
(The percentage of votes depending on an individual and a contextual variable)

	Túlnyomóan munkások lakta választókörzetek	Túlnyomóan nem munkások lakta választókörzetek
Munkások	50%	30%
Nem munkások	20%	10%

Forrás: Boudon (1987)

Sawicki a témáról adott áttekintésében úgy véli, hogy Robinson problémája egy általánosabb kérdés, az aggregációs probléma speciális részét alkotja, mégpedig azt az esetet, amikor az aggregálás kritériumaként a területi folytonosság jelenik meg. A földrajzi csoportosítás Sawicki szerint hasonlít a véletlenszerű csoportosításhoz, amikor a csoportosítási eljárás korrelációs együtthatóra gyakorolt hatását szinte lehetetlen előre jelezni. Az aggregált adatok vizsgálata különböző szinteken történhet, a konkrét szituációtól és a felmérés céljától függetlenül lehetetlen megállapítani a legjobb, optimális szintet. Mindegyik szintnek megvan a maga létjogosultsága. A szomszédsági, városrészi vagy városok közötti különbségeket vizsgáló tanulmányok más és más oksági magyarázatát tárhatják fel ugyanannak a jelenségnek.

Egy hipotézis vizsgálata különböző területi lehatárolásokkal történhet, amelyek eltérő eredményekre vezethetnek. Sawicki Syracuse városára vonatkozó társadalmi adatok vizsgálata során az eltérő aggregáltságú szinteken részben eltérő kapcsolatokat mutatott ki, például a népszámlálási kerületek szintjén a nagyobb tulajdon nagyobb társadalmi interakcióval párosult, míg az alacsonyabb szinteken, részletesebb térfelosztás mellett ilyen kapcsolat már nem volt kimutatható.

A csoportosított adatok vizsgálatát nem pusztán az a kényszer szüli, hogy adatok hiányában gyakran nincsen lehetőség egyéni adatok vizsgálatára, hanem az is, hogy bármiféle beavatkozás számára a csoportok jellemzői könnyebben elérhetőek, mint az egyéneké. Egyéni jellemzők gyakran változtathatatlan, objektív adottságok, például az egyén neme, kora, bőrszíne. A csoportok jellemzői ezzel szemben alakíthatóak. A számítások a cselekvés konkrét irányát nem jelölik ki, mivel ugyanaz az aggregált mutató az átlagok ismert tulajdonságai miatt számos módszerrel megváltoztatható *(Sawicki, 1973)*.

Az ökológiai korrelációk értelmezésének sokszínű szakirodalma nemcsak az adatok értelmezési csapdáinak feltárásához, hanem az egyéni cselekvés és a társadalmi környezet kapcsolatának elemzéséhez is nagymértékben hozzájárult. Adott kutatás keretén belül azonban többnyire az adatok többféle interpretálásá-

ra nyílik lehetőség, még a különféle mennyiségi és minőségi információk közötti összhang követelményének figyelembe vételével is.

6.4. A csoportosított adatokból az egyéni adatokra történő következtetés statisztikai módszerei

A peremgyakoriságok ismeretében az egyéni adatokra történő visszakövetkeztetésre három eltérő gondolatmeneten alapuló, matematikai értelemben egyszerű módszert javasoltak:

- ökológiai regresszió *(Goodman, 1953; Goodman, 1959),*
- határmódszer *(Duncan–Davis, 1953),*
- szomszédsági modell *(Freedman et al, 1991).*

Az ökológiai regressziónak elnevezett eljárásának ismertetésekor az egyszerűség kedvéért Goodman Robinson példájához hasonlóan a négerek/fehérek és írástudók/írástudatlanok dichotóm változók szerint különbözteti meg a népességet. Az ismertetett módszer bármilyen dichotóm változó esetén alkalmazható. A módszer alapgondolata szerint minden egyes körzetnél felírhatjuk a következő egyenletet:

$$Y=xp+(1-x)r,$$

ahol Y az írástudatlanok aránya, x a négerek aránya a népességben, p az írástudatlanok aránya a néger népességen belül, $1-x$ a fehérek aránya a népességben, r az írástudatlanok aránya a fehér népességen belül. Az egyenletben a p és az r paraméter az ismeretlen. Az egyenlet egyszerű átrendezésével a következő formát kapjuk:

$$Y=r+(p-r)x$$

A fenti egyenlőséget annyiszor tudjuk felírni, ahány körzetre vonatkozóan rendelkezünk adatokkal. Mivel az egyenletek számához képest kétszer annyi ismeretlenünk van (minden egyes egyenletben két ismeretlen), ezért az egyenletrendszer csak annak feltételezésével oldható meg, hogy a négerek és a fehérek közötti írástudatlanság aránya mindegyik körzetben azonos mértékű, vagyis p és r állandó. Így kapunk egy körzetek számával egyenlő egyenletrendszert, amelyben két ismeretlenünk van. Ekkor egyértelmű lineáris kapcsolat létezik y és x között, amelynek a meredekségét $p-r$, y tengellyel való metszéspontját r paraméter nagysága határozza meg *(Goodman, 1959).*

A gyakorlatban p és r értéke nem konstans, ez a megoldás egy átlagos eredményt szolgáltat. A hibatag értéke annál kisebb, minél inkább a bőrszín függvénye az írástudatlanság, és a négerek és fehérek közötti arányuk minél kevésbé

változik körzetről körzetre. Ha a terület függvényében nagy mértékben változik az írástudatlanság aránya, akkor Goodman nem ajánlja a módszer használatát *(Goodman, 1959)*.

Goodman módszerének további nehézsége, hogy a paraméterek nem mindig kerülnek a 0 és 1 közötti tartományba, vagyis előfordul, hogy például a fehér népesség -5%-a írástudatlan, vagy hogy egy csoport 110%-a szavazott egy adott pártra. Ezeknél a nyilvánvalóan lehetetlen adatoknál Goodman a lehetséges minimális és maximális értéket, a 0%-os és 100%-os arányt javasolta elfogadni *(Goodman, 1953)*. Sokkal lényegesebb problémát jelent azonban a területileg kiegyenlített arány feltételezése, amely minden tapasztalati ismeretünknek ellentmond *(Fotheringham, 1999; Anselin, 1999)*. Ezen nehézségek ellenére az ökológiai regressziót az ismeretlen paraméterek becslésére széles körben használták fel *(King, 1997)*.

Duncan és Davis tanulmányában egyetért Robinson megállapításaival, és rámutat arra, hogy a területileg csoportosított adatok vizsgálata során az ökológiai korrelációk számítása nem a legjobb elemzési módszer, pontosan azon tulajdonság miatt, hogy egy adott ökológiai korrelációhoz széles sávban mozgó egyéni korrelációk tartozhatnak. A peremgyakoriságok ismeretében számított lehetséges maximális és minimális individuális korreláció meghatározását a szakirodalomban határmódszernek (bounds method) nevezték el. Ennek alapgondolata a következő. Egy kétszer kettes kereszttábla peremgyakoriságai alapján fel lehet írni a cellagyakoriságok lehetséges minimális és maximális értékeit (hasonlóan a 6.1. alfejezet 20. táblázatához). Részletesebb térfelosztásban rendelkezésre álló adatoknál a kisebb területegységek kereszttáblái lehetséges minimális és maximális értékeit külön-külön is kiszámíthatjuk, majd végül ezeket összegezve a teljes területegységre is megkapjuk a kisebb területegységek alapján számított alsó és felső határokat. Ezek a határok az adatok területi eloszlásától és a térfelosztás részletességétől függő mértékben lesznek kisebbek a teljes területre vonatkozó kereszttáblából számítottnál. A kapott értékek segítségével meg lehet határozni az egyéni korreláció lehetséges minimális és maximális értékeit.

Duncan és Davis három népszámlálási adaton alapuló példán keresztül mutatta be a módszer alkalmazását. A 26. táblázatban látható eredményeik mutatják, hogy a térfelosztás részletességével szűkültek az individuális korrelációk számára elméletileg lehetséges határok. A tényleges egyéni korrelációk ebben a két esetben ismertek voltak *(Duncan–Davis, 1953)*.

26. táblázat A belső cellagyakoriságok és a változók közötti korrelációk lehetséges minimális és maximális értéke a határmódszerrel számítva
(The possible minimum and maximum correlation with bounds method)

Változók	Számítás alapja			
	Tényleges érték	Városi adatok	Kerületi adatok	Népszámlálási körzet adatok
Feketék aránya a szolgáltatásokban (%), 1940	37,8	0,0-100,0	21,1-44,5	25,1-40,7
Korreláció	0,289	-0,008-0,898	0,126-0,355	0,165-0,317
Feketék aránya a lakásfoglalók között (%)	92,5	0,0-100,0	75,4-98,0	84,1-95,8
Korreláció	0,116	-0,521-0,168	-0,002-0,153	0,058-0,138

Forrás: Duncan–Davis (1953)

A harmadik módszer, a szomszédsági modell kiszámítása a legegyszerűbb. Ez a modell azon a feltevésen alapul, hogy a körzetenként eltérő megoszlásokat kizárólag a területi hatás eredményezi *(Freedman et al., 1991).* Például a szavazási eredmények különbözőségét nem a bőrszín, a nyelvi, etnikai hovatartozás, végzettség vagy egyéb egyéni tulajdonság körzetenkénti eltérő megoszlása befolyásolja, hanem maguknak a körzeteknek, mint területi entitásoknak az egymástól való eltérése. Ez a kérdésnek az ökológiai regresszió módszerével ellentétes előjelű egyoldalú megközelítése, és nem érvényes például a nemzetiségi alapon történő pártválasztáskor. A cellagyakoriságokat ezen módszer szerint úgy kapjuk, hogy a megfelelő peremgyakoriságok szorzatát elosztjuk a teljes gyakorisággal.

Önmagában *az aggregált adatokból elvi, logikai okok miatt csak korlátozottan lehetséges az egyéni viselkedésre következtetni.* Goodman jól látta az ökológiai regresszió korlátjait, a határmódszer pedig eleve nem ígért többet, mint amit a csoportosított adatokból megtudhatunk az egyéni adatokról. A kutatók egy része azonban újabb és újabb javaslatokat tett ezeknek a módszereknek a továbbfejlesztésére. Az adatok becslésének javításában két fontosabb irányzat figyelhető meg. Az egyik a vizsgált jelenségre vonatkozó külső információk, más adatforrások igénybevételével, azok eredményeinek a felhasználásával próbálja meg pontosítani az ökológiai regresszió paraméterbecslését és a határmódszer által megadott intervallum szűkítését. Ilyen külső információ lehet például egy kérdőíves felmérés, amelynek adatai egyénekre vonatkozóan is ismertek a kutatók számára.

A kutatás másik iránya az egyre szofisztikáltabbá váló matematikai statisztikai eszköztár bevetésével próbálja meg kiküszöbölni az ökológiai regresszió két hiányosságát, vagyis a paramétereknek esetenként a logikailag lehetséges 0-1 tartományon kívülre kerülését és a területileg egységes paraméterek feltételezé-

sét. Ezek a módszerek azonban mind élnek a paraméterek eloszlására vonatkozó ellenőrizhetetlen feltevésekkel, vagyis amelyekről lehetetlen eldönteni, hogy adott esetben elfogadhatóak lennének-e vagy sem.

Számos ilyen módszer kidolgozása után King 1997-ben megjelent könyvének a címében egyenesen az ökológiai tévkövetkeztetés problémájának egy megoldását ígéri. Összességében a recenziók és reakciók alapján King „megoldása" vegyes fogadtatásban részesült. A több évtizedes probléma megoldását lelkesen üdvözlőktől *(Sui, 1999)*, a mérsékelt lelkesedésen *(Fotheringham, 1999)*, szkepszisen *(Raudenbusch, 1998)* és a visszafogott bírálaton *(Anselin, 1999)* át az egyértelmű elutasításig *(Freedman et al., 1998)* terjed a vélemények skálája. A módszert már több választási eredményt tartalmazó adatbázis elemzésére használták, és megjelentek a különféle változatai, melyek a paraméterek eltérő eloszlását feltételezik *(Withers, 2001)*.

A paraméterek feltételezett eloszlásfüggvényét változtatva rendre eltérő eredményeket kapunk, amelyeknek egyike sem lesz valóságosabb, jobb a másikhoz képest. Anselin álláspontjával tudok leginkább azonosulni, aki „bonyolult találgatós játéknak" nevezi a módszert. A csoportosított adatokból az egyéni adatokra való következtetésre javasolt bármilyen módszer tartalmaz igazolhatatlan, ellenőrizhetetlen feltevéseket, az ökológiai tévkövetkeztetés problémájának nincs megoldása *(Anselin, 1999)*. Vagyis ebben a „találgatós játékban" nem tudjuk kitalálni a helyes eredményeket, mert azok az adatfelvétel módja miatt ismeretlenek maradnak. Flanigan és Zindale az alkimistáknak az aranykészítés utáni kutatásához hasonlítja a cellagyakoriságok kitalálására kifejlesztett módszereket, azzal a különbséggel, hogy az alkimisták tudták, hogy kudarcot vallottak, míg az ökológiai következtetések esetén nincsen olyan kritérium, amellyel a hibát ki lehetne mutatni. A becslési eljárások pontosságát sem ismerhetjük meg, így arra sincs lehetőség, hogy közülük az alapján válasszunk, hogy melyik szolgáltat pontosabb eredményeket *(Flanigan–Zindale, 1985; Flanigan–Zindale, 1986)*. „Az egyre komplexebbé váló megközelítések nem jobb és jobb becslésekhez vezetnek, hanem az ökológiai tévkövetkeztetés bonyolultabb formáihoz" *(Flanigan–Zindale, 1986, 88. o.)*.

Egy rossz adat, illetve egy olyan adat, amelynek a megbízhatóságáról semmit sem tudunk, gyakran rosszabb, mint a semmilyen adat. A határmódszer az adatokban benne rejlő információkat tárja fel, az ökológiai regressziót Goodman a megfelelő óvatosság mellett javasolta. Ezek az egyszerűbb módszerek nem ígérnek többet a logikailag megvalósíthatónál, ezért alkalmazásuk nem jár veszéllyel. Az ökológiai következtetés megoldását ígérő módszerek eredményei a probléma lényegéből, az információk hiányából fakadóan bizonytalan adatokat képesek szolgáltatni, amelyeket bizonyosságként vagy akár csak erős valószínűségként kezelve az ökológiai tévkövetkeztetés új formáihoz jutunk el *(Flanigan–Zindale, 1985; Flanigan–Zindale, 1986)*.

6.5. Az ökológiai tévkövetkeztetés empirikus jelentősége

Az ökológiai tévkövetkeztetés lehetősége tisztán logikai úton is belátható és hipotetikus adatok segítségével is bemutatható. Az empirikus elemzéseket az egyedi adatokhoz való rendkívül nehéz hozzáférhetőség nagymértékben korlátozza. Az alábbiakban egyetlen ilyen elemzés eredményeit ismertetem. A módosítható egység problémája kapcsán további, az adatok csoportosításának hatására vonatkozó elméleti kérdések és empirikus tapasztalatok bemutatására nyílik majd lehetőségem.

27. táblázat Egyéni és ökológiai korrelációk kereszttáblája (a soronkénti összes esetszám százalékában)

(Crosstabulation of individual and ecological correlations (percentages of row totals)

Egyéni korrelációk: -tól/-ig	Ökológiai korrelációk: -tól/-ig										Összes esetszám
	-1.0/ -0.8	-0.8/ -0.6	-0.6/ -0.4	-0.4/ -0.2	-0.2/ 0.0	0.0/ 0.2	0.2/ 0.4	0.4/ 0.6	0.6/ 0.8	0.8/ 1.0	
Sunderland 1 kilométeres négyzetek (53 változó)											
-1.0/-0.8	100										1
-0.8/-0.6	50	50									4
-0.6/-0.4	12	44	32	12							25
-0.4/-0.2		9	36	34	15	4	1				180
-0.2/0.0			4	32	39	18	5	1			997
0.0/0.2			1	2	14	29	32	20	3		188
0.2/0.4						14	32	39	14		28
0.4/0.6							17	50	17	17	6
0.6/0.8								50	50		2
Összesen	6	32	117	387	444	248	117	66	13	1	1431
Firenze népszámlálási körzetek (40 változó)											
-1.0/-0.8	100										1
-0.8/-0.6		0									0
-0.6/-0.4			100								2
-0.4/-0.2	2	19	31	24	17	6					83
-0.2/0.0		1	7	21	32	23	14	2			603
0.0/0.2			1	6	10	28	28	22	3		78
0.2/0.4							18	27	55		11
0.4/0.6								100			1
0.6/0.8									100		1
0.8/1.0											0
Összesen	3	21	72	154	214	167	106	33	10	0	780

Forrás: Openshaw (1984a)

Openshaw szerint lényegében önkényes földrajzi körzetek szintjén állnak rendelkezésre a népszámlálási adatok, az aggregálás adatokra gyakorolt hatására ugyanakkor az eredmények értelmezése során meglepően kevés figyelmet fordítanak. A probléma viszont nagyon lényeges, és általánosságban is felveti a népszámlálási adatok használhatóságának kérdését.

Az ökológiai korreláció gyakorlati jelentőségét két elérhető adatbázison keresztül illusztrálja Openshaw. Az egyik a sunderlandi háztartások 10%-áról készült felmérés, a másik Firenze háztartásainak adatai. Az egyik esetben 53, a másikban 40 különböző változó közötti egyéni és ökológiai korrelációt számított ki és hasonlított össze Openshaw (27. táblázat). Az ökológiai korrelációkat Sunderlandban egy, valamint fél kilométeres négyzet alakú körzetekkel, Firenzében pedig a népszámlálási körzetek alapján csoportosítva számította. A területegységek méretének csökkenésével azok homogénebbé is válnak, így az egyéni és ökológiai korrelációk közötti különbségek csökkentek. Az aggregáció hatása nem előrejelezhető és matematikai eszközökkel nem korrigálható.

Openshaw szerint az ökológiai tévkövetkeztetés lehetősége függ a vizsgálat módszerétől és az eredmények interpretálásának módjától. Nincs ismert módszer, amely az aggregálás hatását képes lenne mérni. A 27. táblázat extrém értékei például rendkívül félrevezetők lehetnek, azok alapján téves következtetésekre lehet jutni. Ha individuális korreláció számítására az adatok hozzáférhetőségének hiányában nincs mód, akkor a torzító hatás nagyságáról nincs ismeretünk *(Openshaw, 1984a).*

Lábjegyzetek

[1] Robinson a néger szót használta, amit a mai példákban felváltott az afroamerikai kifejezés.

[2] Az adatok elérhetőségének negyedik típusának nevezhető az az eset, amikor az egyedi adatok között az adatok jellegéből következően nem számolhatunk korrelációt. Ilyen korrelációra a 6.3. és a 7.2. alfejezetben lesz példa.

[3] Ezt a kérdést a 9.7. alfejezetben illusztrálom a területi mozgó korreláció segítségével.

[4] Az adatok tipologizálásáról és az egyéni és csoportosított adatokat is felhasználó elemzések további kérdéseiről részletesebben lásd Moksony (1985), elsősorban 16-32. oldal.

[5] Ezek a következők: „(1) a katolikusok – az egyes országok vallási tagozódásától függetlenül – kevésbé hajlamosak az öngyilkosságra, mint a protestánsok; (2) a túlnyomóan katolikus országokban mind a katolikusok, mind a protestánsok kevésbé hajlamosak az öngyilkosságra, mint a nem katolikus országokban, ám egy adott országon belül nincs különbség a katolikusok és a protestánsok öngyilkossági arányszámai között; (3) a katolikusok inkább hajlamosak az öngyilkosságra, mint a protestánsok, de a túlnyomóan katolikus országokban mind a katolikusok, mind a protestánsok öngyilkossági arányszáma kisebb, mint a nem katolikus országokban; (4) a katolikusok öngyilkossági arányszáma – az egyes országok vallási tagozódásától függetlenül – állandó, ám a protestánsok öngyilkossági arányszáma annál kisebb, minél nagyobb az adott országban a katolikusok aránya” *(Davis–Spaeth–Houson, 1987, 271. o.).*