„I just ran two million regressions”, avagy módszertani paradigmák a kvantitatív társadalomkutatásban[1]

Németh Renáta, ELTE TáTK, Statisztika tanszék

Kulcskérdések a társadalomkutatásban – Kánon és apokrif c. konferencia, ELTE TáTK, 2012. május 25.

Kivonat

Kuhn óta elfogadott nézet, hogy a különböző korok paradigmatikus elméletei nem csak tartalmilag, hanem módszertanilag is különböznek egymástól. Ennek ellenére a módszert alkalmazó tudomány (előadásomban: a kvantitatív társadalomkutatás) vagy maguk a módszerek (előadásomban: a statisztikai eszközök) ritkán reflektálnak erre. Hasonlóan ritkán merül fel az a szempont, hogy magának a statisztikának is önálló, a társadalomtudományoktól részben független története van, saját paradigmákkal, melyek talán az általa szolgált tudományra, annak szemléletére, fogalmaira, kérdésfeltevéseire is hatást gyakorolnak. Előadásomban néhány példával igyekszem alátámasztani ezt az álláspontomat.

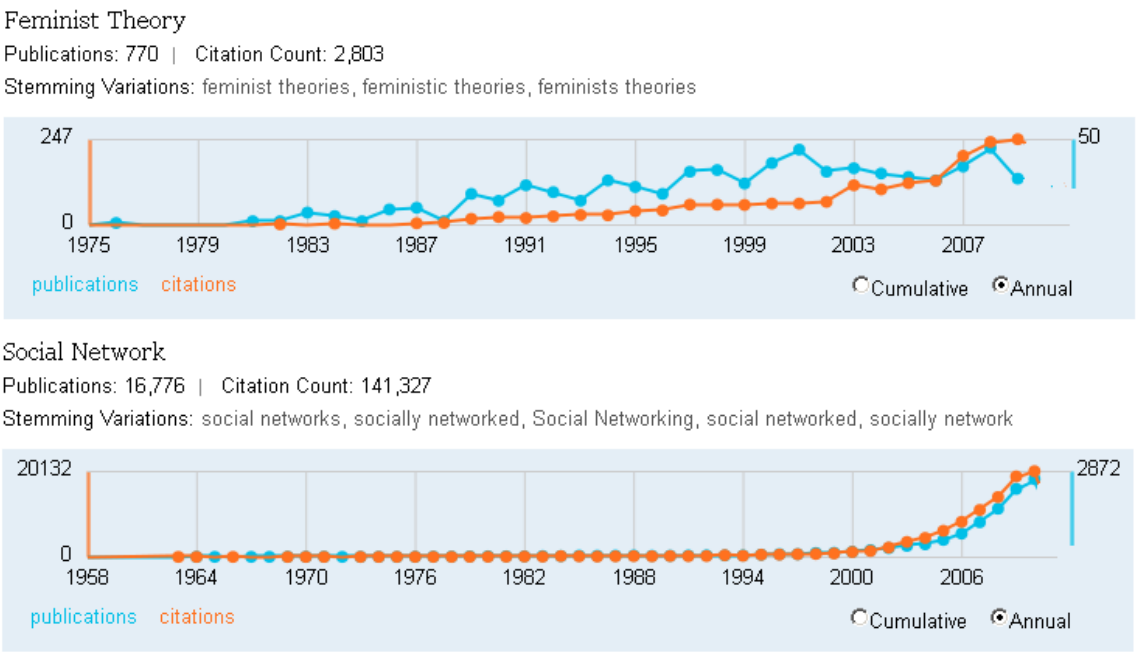
Bevezetés

Az ELTE Társadalomtudományi Kar Statisztika tanszékének munkatársaként a konferencia címéről („Kánon és apokrif”) a társadalomtudományok módszertani paradigmáira asszociáltam, ezzel kapcsolatban szeretném néhány gondolatomat megosztani Önökkel. Az előadásom címében szereplő büszke kijelentés („I just ran two million regressions”) jó példa e paradigmák létére. A mondat egy 1997-es, American Economic Review-beli cikk címe, melyben a gazdasági növekedés magyarázó faktorait határozza meg a szerző. Eljárása lényege, hogy a tudományos publikációkban a gazdasági növekedés potenciális magyarázó faktoraként feltüntetett 60 változó különböző kombinációival előálló kétmillió regressziót futtat le, majd megfigyeli, hogy az egyes változók hányszor voltak szignifikánsak, milyen eloszlása volt a regressziós együtthatóknak stb. A címadás azt jelzi, hogy a számítástechnikai kapacitások ezt addig nem tették lehetővé, tehát a kétmillió regresszió lefuttatása abban az időben meglepően nagy számnak számított - ma már nem az, a szimulációs technikák, mint a bootstrap, rutinszerűen végeznek ilyen nagyságrendű számítást, vagyis nem jelenhetne meg ilyen címmel cikk egy szaklapban. De e mögött a módszertani paradigma mögött nem csupán technikai újítások vannak. A szerzők tisztán adatvezérelt módon határozzák meg a növekedési faktorokat, azaz az összes lehetséges modell illesztésével, és nem a klasszikus elmélet-központú alapon, egyetlen, a hipotézisnek megfelelő regressziós modellt tesztelve. Az, hogy ez a cikk az egyik legnevesebb közgazdasági szaklapban jelent meg, azt is feltételezi, hogy ez a fajta bizonyítási eljárás elfogadott volt az aktuális tudományos közösség által.

[1] A dolgozat a konferencián elhagzott előadás kibővített változata. Elkészítése során az MTA Bolyai János kutatási ösztöndíj támogatását élveztem. Ezúton köszönöm Sik Domonkos baráti észrevételeit, ill. Keszei Ernő és Örkény Antal konferencia-előadásomat követő hozzászólását - a dolgozatba ezeket is beépítettem.

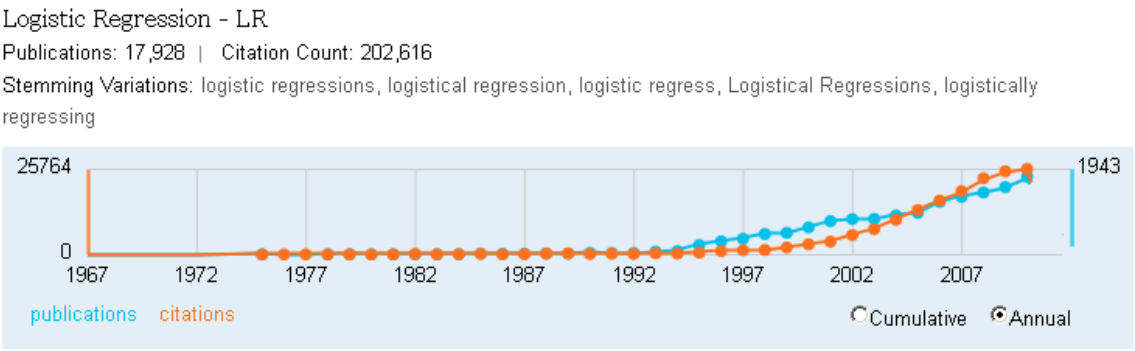
Kulcsszavak: tudománymetriai megközelítés

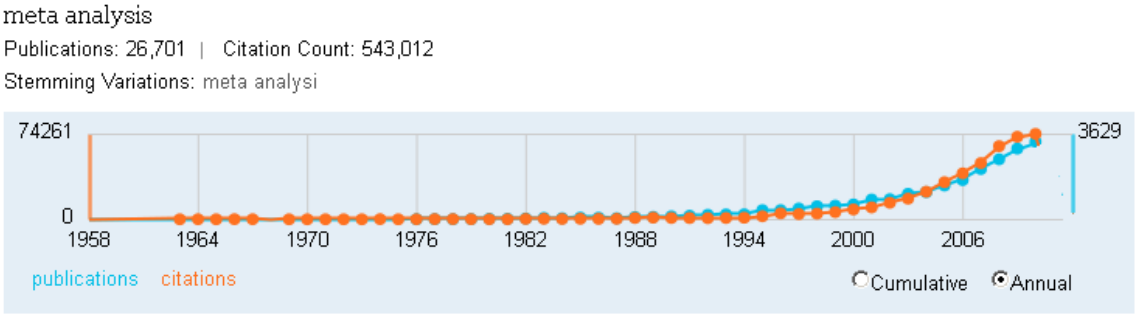
A tudományos paradigmákat legegyszerűbben (és persze kissé leegyszerűsítve) tudománymetriai úton közelíthetjük meg. Az alábbi két ábra a Microsoft Academic Search szolgáltatást felhasználva tudományos publikációk kulcsszavai alapján mutat egy-egy trendet. Az első ábra a *feminist theory*, a második a *social network*, mint kulcsszó előfordulását ábrázolja; a kék vonal a kulcsszót tartalmazó publikációk számát, a piros vonal az ilyen publikációkra mutató hivatkozásokét jelöli. Mindkét ábrán meredeken növekvő trendet láthatunk, amit nyilván az információs robbanás, a tudástranszfer innovatív új módjai, a publikációs felületek gyarapodása is generál, de a két ábra összevetésében a különbségek az érdekesek. A feminista elmélet 70-es évek elején jelenik meg kulcsszóként, míg a közösségi hálózat már 1962-ben, viszont az előbbi mai elterjedtségét már több évtizede megközelítette, szemben a közösségi hálózattal, ami az utóbbi tíz évben exponenciális ütemben növeli népszerűségét. Ha nem csak az arányokra figyelünk, hanem a konkrét számokra is: a feminista elmélet évi 50 publikációban szerepel népszerűsége csúcsán kulcsszóként, a közösség hálózat 2872-ben. Ennek oka az is, hogy az utóbbi nem csak a társadalomtudományokban szerepel; pl. a közösségi hálózatot kulcsszóként tartalmazó legidézettebb cikk többszáz hivatkozással számítástudományi.



A közösségi hálózatok kutatását nyilván az internet térnyerése, tehát egy új tárgy, egy újonnan jelentkező igény is generálja. De a kutatási témák születését és kihalását nem csak efféle objektív igények magyarázzák. Thomas Kuhn óta közhelyszerű, hogy a tudomány nem vizsgálható függetlenül a tudományt művelő közösségtől, ami esetünkben annak a szempontnak a bevonását jelenti, hogy a vizsgálatra kijelölt problémák körét a közös paradigmát képviselő tudományos közösség határozza meg. Részben ez, a kutatási témák egyezményessé válása magyarázza tehát a fenti ábrákon látható dinamikát.

Ugyanez igaz azonban nem csak a kutatási témákra, hanem a felhasznált módszerekre is: a tudományos közösség által hordozott paradigma jelöli ki a kutatási módszerek megválasztását. Nézzük most a kulcsszavak trendjét két statisztikai módszerre vonatkozóan.





A logisztikus regresszió módszerét David Cox 1970-ben publikálta elsőként, mégis, csak húsz év múlva lett igazán használt. A metaanalízis a 60-as évek eleje óta ismert módszer, de csak a '90-es évek óta van terjedőben. Az utóbbi évtizedben mindkét módszer népszerűsége nagy sebességgel növekszik, párhuzamosan azzal, ahogyan beágyazódnak a tudományos közösségek elismert eszközei közé[2]. A metaanalízis az adatsor alapján kétszer népszerűbb 2010-ben mint a logisztikus regresszió; ez társadalomtudománnyal foglalkozó hallgatóságom számára talán meglepő lehet, hiszen itt a metaanalízist szinte egyáltalán nem használják.

Más módszerekkel kapcsolatban is megfigyelhető ez a tudományterület-specifikusság. Vegyük például a modellilleszkedés jellemzésére használható BIC (Bayesian Information Criterion) és AIC (Akaike Information Criterion) mutatók közötti választás kérdését. Ezek a mutatók információelméleti alapúak, nem használnak klasszikus hipotézistesztet, és viszonylag frissek: az AIC-et Hirotugo Akaike, japán statisztikus vezette be 1973-ban, a BIC-et Adrian Raftery amerikai szociológus javasolta alkalmazásra 1986-ban. Nagyon eltérő a két mutató tudományterületenkénti ismertsége, s ez nem magyarázható matematikai tulajdonságaik eltérő voltával. A JSTOR statisztikája szerint a BIC inkább a szociológiai, az AIC a közgazdaságtudományi cikkekben használatos (szövegbeli előfordulások száma: szociológia 486 vs. 241, közgazdaságtudomány 609 vs. 1208), biztosan nem függetlenül attól, hogy a BIC-et a szociológus Raftery vezette be.

[2] Meg kell jegyezni, hogy statisztikai módszer nyilván nem szerepel minden olyan cikk kulcsszavaként, amelyben azt eszközként használják; inkább csak akkor, ha a cikk módszertani szempontból is új eredménnyel szolgál.

Toolbox helyett: módszertani paradigmák

Ha a módszerek kiválasztása tökéletesen racionális módon történne, valahogy úgy, mint amikor egy toolbox-ból mindig az adott problémára leginkább érvényes eszközt húzzuk elő, akkor **(1)** a BIC-AIC esetén megfigyelt tudományterület-specifikusság nem volna jelen. Hasonlóan ellentmond e toolbox jellegnek az, hogy **(2)** bizonyos kulcsszavak idősorain a módszerek hálózatokon keresztül történő terjedésére utaló exponenciális növekedést, vagyis járványszerű terjedést láttunk – ez bizonyíték arra, hogy vannak módszertani divatok, melyek tudományos kapcsolatokon keresztül terjednek[3].

A toolbox-jelleg ellen szól az is, hogy **(3)** tudományterületek érintkezésénél módszertani változások is beállnak. Ilyen pl. az analitikus szociológia vagy a szociofizika[4] a szociológia és a fizika (pontosabban a statisztikus fizika) határterületén. E megközelítések célja a kollektív viselkedés, komplex rendszerek, makrojelenségek tanulmányozása néhány változót tartalmazó óriási mintán, legtipikusabban a hálózatkutatásban. A fizikával történő érintkezés változást generált a szociológiai módszertanban is: a statisztikus fizika benyomulása tapasztalható a szociológia korábban statisztika által uralt módszertani mezejére. A különbség szembetűnő. A statisztikus fizika az egyéni viselkedés kisszámú jegyre egyszerűsített modelljéből indul, ebből vezeti le a nagy tömegre vonatkozó konklúzióját, akár konkrét empirikus adatok nélkül (analitikus módon matematikai levezetéssel vagy szimulációkkal). Ezzel szemben statisztika a tömeg (populáció) hipotetizált tulajdonságaiból indul ki és arra vonatkozóan tesz a konkrét adatokra alapozott valószínűségi következtetéseket. Az előrejelzés és a kauzalitás fogalmának megközelítése is más a két esetben: ha a modell „korrekt", akkor a fizikában automatikusan teljesül az előrejelzés, míg a statisztikai modellek ritkán állítják magukról, hogy korrektek lennének kauzálisan is.

Ugyancsak a toolboxszal szembeni érv az, hogy az eszközválasztásra nyilvánvaló módon hatnak a történelmi tradíciók is. Holmes (2007) pl. a többváltozós elemzések „francia módijáról" (*French way*) beszél: a francia statisztikusok a '60-as '70-es években a valószínűségi absztrakciókat praktikus szempontból haszontalannak ítélték, helyette adataik vizuálisan jól interpretálható geometriai reprezentációjával dolgoztak. Ez az amerikai statisztikus iskolától nagyon eltérő szemlélet. Maga a geometriai/vizuális gondolkodás Descartes óta kimondottan hangsúlyos a francia matematikai szemléletben, ami kihat a kutatásra (analízis, topológia[5]) és a matematikaoktatása is (ez is geometria-centrikus). Úgy tűnik tehát, hogy itt tágabb gondolkodási hagyományok is közrejátszhatnak, vagyis kicsit arról is szó van, miként gondolkozik egy francia/német/angolszász kutató.

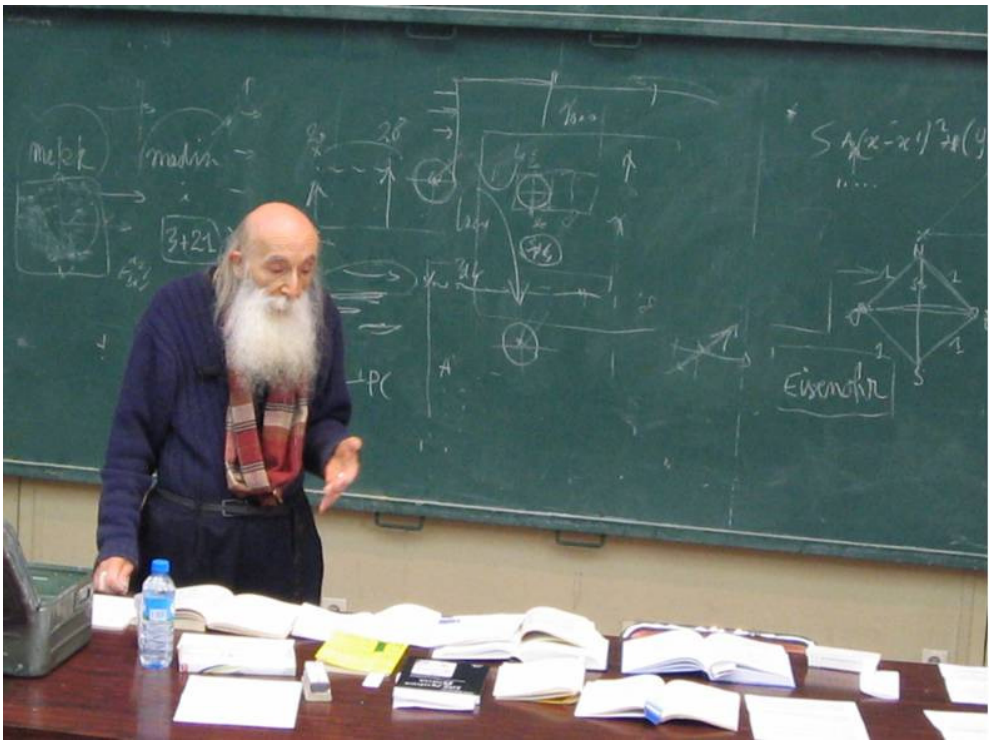
Ebből a kép-központú nézőpontból fejlesztette tovább és tette népszerűvé a '60-as években Franciaországban Jean-Paul Benzécri és iskolája a korrespondencia-analízist. Ez a módszer angolszász nyelvterületen sokkal kevésbé népszerű: a google találati listája a *factor analysis* vs. *L'analyse factorielle* keresőszavakra az angol kifejezés javára 16-szoros, míg a *correspondence analysis* vs. *L'Analyse des Correspondances* keresőszavakra csak hatszoros

[3] Ugyanis ha hálózaton terjed egy információ / betegség, akkor leggyakrabban exponenciális sebességű terjedés figyelhető meg. E sebesség matematikai feltétele az, hogy több, mint 1 legyen az informáltak/továbbfertőzöttek átlagos száma. Innét jön a vírusmarketingnek nevezett módszer neve is, ami szociális hálózatok mozgósításával valamely brand népszerűségének exponenciális ütemű növelését célozza.

[4] Az analitikus szociológia és a szociofizika terminusok angolul és magyarul is rögzültek már talán; nem teljesen ugyanazt jelölik, de tárgyunk szempontjból hasonló a megközelítésmódjuk.

[5] „Van a francia és van az orosz topológia. Lehet úgy is művelni a topológiát, mint az oroszok, de nem érdemes". – vezette be a matematikai szépséget mindig fontosnak tartó analízistanárom, Czách László az ELTE TTK-n előadását.

különbséget mutat. A magyar korrespondencia(-)analízis/elemzés keresőszó csak 250 találatot ad.

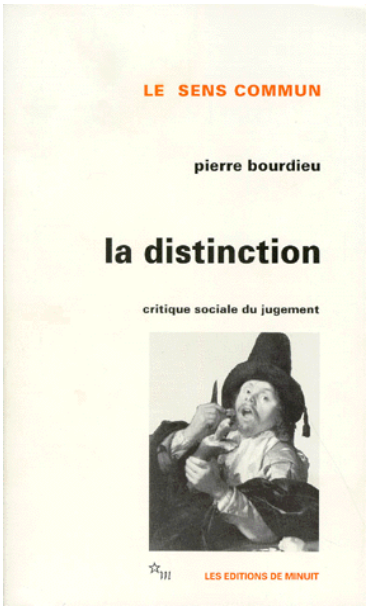


Jean-Paul Benzécri

A statisztikai elemzés francia módija: Bourdieu

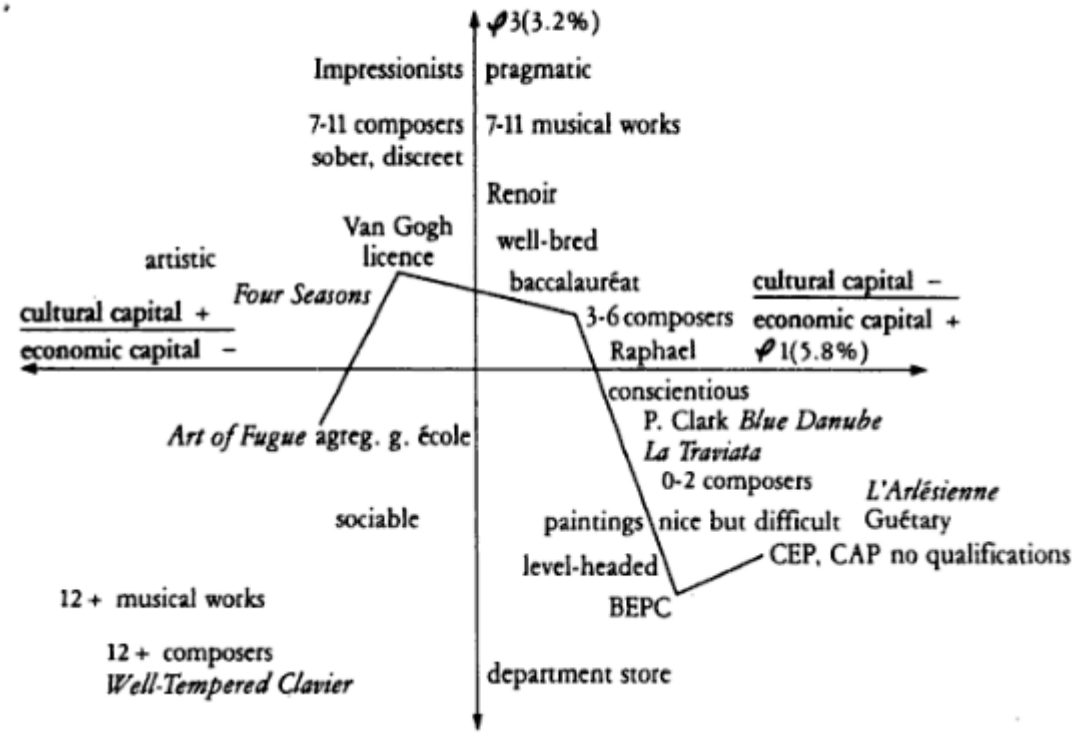
A korrespondenciaanalízis a legtöbb statisztikai módszerrel szemben nem numerikus, hanem kvalitatív változók közötti kapcsolat elemzésére alkalmas. Kreatív és gondos használatának legismertebb szociológiai példája Bourdieu La Distinction (1979) c. könyve[6]. Ez az empirikus vizsgálatokra épülő ízlésszociológiai munka annyira tipikus alkalmazása a korrespondancia-analízisnek, hogy azóta gyakran „La Distinction paradigma"-ként is utalják az eszköz ilyen használatát. Nem csak tudománytörténetileg, hanem témánk, a módszertani paradigmák szempontjából is fontos megemlíteni, hogy Bourdieu Benzécrihez és köréhez kapcsolódva ismerte meg a módszert, és azt is, hogy Bourdieu és Benzécri már az École Normale Supérieure-n ismerték egymást, sőt Benzécri nekrológja szerint Bourdieu haláláig leveleztek, tudományos és baráti kapcsolatban voltak.

[6] Bourdieu már korábban, 1976-ban az Az ízlés anatómiájában (*Anatomie du goût*) is használta ezt a módszert.



A korrespondencia-analízis lényege, hogy tulajdonságok együttjárását az őket reprezentáló síkpontok geometriai közelségébe fordítja. Alább a könyv egyik ábráját, egy korrespondencia-elemzési outputot láthatunk. Leolvasható pl., hogy a diszkrét, mértékletes bútorok preferálása az impresszionisták kedvelése mellett helyezkedik el, ahogyan a pragmatikus barátok birtoklása is. Az egyetemi végzettség (*agreg. g. école*) Bach *A fúga művészete* c. művének preferálása mellett van. De általánosabb strukturális következtetések is levonhatók a tengelyek mögött álló látens faktoroknak történő jelentéstulajdonítással, pl. az iskolai végzettség a síkon jobbról balra haladva növekszik, tehát a pontfelhő vízszintes tengelyének egyik fontos meghatározója a kulturális tőke lehet.

Figure 13 Variants of the dominant taste. Analysis of correspondences: simplified plane diagram of 1st and 3rd axes of inertia.



Ez a térbeli reprezentációt kínáló statisztikai eszköz párhuzamba állítható Bourdieunek a társadalomra vonatkozó térbeli víziójával, ahogyan azt már sokan észrevételezték (pl. Lebaron 2009, Rouanet és társai 2000). Bourdieu számára a társadalom térben létezik, számára a társadalmi kapcsolatok és távolságok elsősorban térbeli kapcsolatok és távolságok. (Térben, szó szerint: térképen. Lásd pl. a *Les Règles de l'art* elemzését az *Érzelmek iskolájá*ról: egy korabeli Párizs-térkép segítségével világítja meg, hogy Flaubert hőseinek lakóhelyei, költözései a szimbolikus társadalmi rendszerben történő elhelyezkedésnek, az abban való mozgásnak feleltethetők meg). Bourdieu komplex struktúrában gondolkodik, nem egyszerűen vertikálisan, egy-egy dimenzió mentén tagolja a társadalmat, amint azt pl. a lineáris regresszió igényelné/sugallná. Bourdieu maga is többször utal rá, hogy számára a regresszió nem megfelelő eszköz, mert túlságosan leegyszerűsít, és a különbségeket hangsúlyozza a hasonló csoportok keresése helyett. A bevett statisztikai eljárások (amik ebben az időben főként folytonos változókat használó többváltozós technikák, mint a lineáris regresszió) helyett ezért döntött a korrespondencia-analízis mellett. Így ír erről *A szociológus mestersége* (*Le Métier de Sociologue*) 1991-es német kiadásának előszavában:

„Gyakran használok korrespondenciaanalízist, mivel azt gondolom, hogy ez alapvetően egy olyan, relációkra épülő módszer, aminek a filozófiája tökéletesen kifejezi mindazt, ami véleményem szerint a társadalmi valóságot felépíti. Olyan eljárás, ami relációkban „gondolkodik", ahogyan arra én is kísérletet teszek a mező fogalmának használatakor."[7]

Módszertan és alkalmazó társadalomtudomány: hatás és visszahatás, koevolúció

Eddig a statisztikai módszer megválasztásának paradigmatikus jellegére hoztam példákat. Az utolsó példa a történelmi tradíciókról, a francia statisztikai hagyományról szólt, azonban egy lépéssel tovább is vezethet. Ha maga Bourdieu is ilyen erős párhuzamot érez szociológiai rendszere és statisztikai módszere között, és ha elemzői is tételről tételre azonosítják a korrespondenciaanalízis fogalmait Bourdieu szociológiai fogalmaival (pl. a disztinktív jegy nem más, mint az a kategória egy látens fogalom, azaz tengely szempontjából, ami az origótól messze van stb.), akkor talán érdemes felvetni azt a kérdést, hogy maga a módszer nem gyakorol-e hatást nem csak az eredményekre, hanem a társadalomszemléletre, a használt fogalmakra, kérdésfeltevésekre is. Gondoljunk csak arra például, hogy a korabeli technika számára a háromdimenziós ábrázolás még nehézkes volt, csak a kétdimenziós síkábrázolás volt elérhető az interpretációhoz (mint amilyent a fenti ábrán láttunk), még akkor is, ha a pontok struktúrája kettőnél több fontos dimenziót mutatott. Ez szükségszerű leegyszerűsítése a modellnek, és óhatatlanul egyszerűsíti az interpretációt, a társadalomképet is. Vagy gondoljunk arra, hogy a látens dimenziók fogalmát (ami a kulturális, gazdasági, társadalmi tőke fogalmához vezet el) a módszer a koordinátatengelyekkel maga implikálja, s ezek a tengelyek mint látens dimenziók fel sem merültek volna, ha Bourdieu pl. hálózatelemzési módszereket használt volna.

Utolsó példám azt mutatja, hogy a módszertan a mobilitáskutatás területén is paradigmatikus jellegű, és hogy a használt statisztikai eszköz itt is erősen formálja a kutatási koncepciót. Jonathan Kelley (aki a magyar olvasónak onnét is ismerős lehet, hogy Kolosi

[7] Saját fordításom.

Tamással publikált ’92-ben az American Sociological Review hasábjain egy magyar-ausztrál összehasonlító elemzést) alábbi, 1990-as írása is erre mutat:

„There has been a revolution in the study of social mobility: the once dominant Blau-Duncan paradigm has been overthrown by log-linear modeling. [...] The log-linear revolution was a noble experiment, and at first seemed to offer a bright new future beyond Blau and Duncan. [...] They have done impressive work with it; that is freely conceded even in this critical appraisal. Their Herculean labors in data preparation are ambitious in scope and commendable in quality. Their analytic methods are modern, employed with agreeable technical virtuosity, and presented with commendable clarity. [...] A decade and more has passed. Endless models have been fitted; legions of design matrices passed in review; chi-squares marshaled. [] When Goldthorpe began his work on stratification, the Blau-Duncan paradigm (Blau and Duncan 1967; Duncan and Hodge 1963; Duncan 1966) was in its heyday, having advanced study of social stratification from its speculative and non-cumulative beginning to the status of a normal science.”

Kelley ebben az írásában expliciten használja a paradigma kifejezést, mégpedig kuhniánus megközelítésben, említi a „normál tudomány” és a „tudományos forradalom” kifejezéseket is. Amit kiemelnék: az egyes paradigmákat az általuk használt módszerekről nevezi el, és tulajdonságaikban is technikai-módszertani jegyeket emel ki.

Ezeket a mobilitáskutatási paradigmákat alapvetően a társadalmi státusz mérésére használt változók típusa (folytonos vagy kategoriális) különbözteti meg. A Blau-Duncan paradima a foglalkozási státuszt mérő folytonos Duncan-féle társadalmi-gazdasági indexet (SEI) használta. Az index eredetijét az amerikai National Office of Vital Statistics munkatársai hozták létre az ötvenes években; azért, mert „They needed a way to reduce the minute detail of occupational categories into a small number of quantities, preferably one, *that could be used to calculate correlations*.” (Hout, 2007, kiemelés tőlem). Az indexnek ezek a tulajdonságai tették lehetővé később a foglalkozás útelemzésben szerepeltetését, a Blau-Duncan modell megszületését, a mobilitási folyamat hatásainak elkülönítését és számszerűsítését, a direkt és indirekt hatás fogalmának megjelenését. Ettől kezdve az Egyesült Államokban a társadalmi helyzet operacionalizálása inkább a (folytonos) foglalkozási presztízzsel történik, míg Európában elsősorban (kategoriális) osztályokat használnak (a legismertebb Erikson, Goldthorpe és Portocarero után a tizenegy kategóriás EGP osztályséma). Ez persze struktúraelméleti különbséget is jelez. Az osztályok kevés, nem feltétlenül rangsorolható, több dimenzió mentén (munkaerőpiaci helyzet, képzettség, szektor stb.) tagolódó, belül homogén csoportból állnak; míg a foglalkozási presztízs sok-sok, egyetlen dimenzió mentén rendezhető csoportot hoz létre. De a kiemelés Hout szövegéből - miszerint azért hozták létre az indexet, hogy korrelációt tudjanak számolni (!) - és talán Kelly írása is elképzelhetővé teszi az operacionalizálás módjának prózai módszertani indíttatását. A társadalomkutatási praxisban gyakran magunk is az aktuális módszer által megkívánt mérési szint szerint operacionalizáljuk változóinkat. Azt is meg lehet itt jegyezni, hogy (leginkább történeti okokból, a hagyományt követve) a társadalomtudományi képzés statisztika kurzusain a világ minden táján jelentősen nagyobb szerep jut a folytonos változókat használó többváltozós elemzéseknek, pedig a kutatói gyakorlatban előforduló változók nagyobb része kategoriális, ezért van szükség indexképzésre vagy „felskálázásukra”. A folytonos ill. kategoriális változók melletti döntésnek ugyanakkor messzemenő következményei vannak nem csak a szociológiai koncepcióra, hanem az adatokon végezhető elemzésre nézve is. A folytonos megközelítés a majdnem mindig automatikus normalitás-feltétellel egyszerűségre és lineáris kapcsolatokra törekszik, míg a kategoriális megközelítés finomabb modellezést tesz lehetővé (nemlineáris kapcsolatok, komplex interakciók), de csak kisszámú kategóriát képes kezelni.

Aktuálisan használt módszereink tehát óhatatlanul behatárolják a tudományos kérdések és társadalomreprezentációk körét, formálják szemléletünket. Előadásom ezzel kapcsolatos önreflexiónkhoz próbált hozzájárulni.

Hivatkozások

Holmes, S.P. (2007): Multivariate Analysis: The French Way. *Probability and Statistics*, Volume 2, pp 219-233.

Hout, M. (2007): Otis Dudley Duncan's major contributions to the study of social stratification. *Research in social stratification and mobility*, 26:109-118.

Kelley, J. (1990): The failure of a paradigm: Log-linear models of social mobility. In Clarke, Modgil and Modgil (szerk.): *John Goldthorpe: Consensus and Controversy*, London: Falmer Press. Pp. 319-346.

Lebaron, F. (2009): How Bourdieu „quantified" Bourdieu: the geometric modelling of data. In: Robson and Sanders (eds): *Quantifying theory: Pierre Bourdieu*. Springer.

Rouanet, H., Ackermann, W., Le Roux, B. (2000): The geometric analysis of questionnaires: The Lesson of Bourdieu's La Distinction. *Bulletin de Méthodologie Sociologique*, 65, 5-15.

Sala-I-Martin, X.X. (1997): I just ran two million regressions. *The American Economic Review*, 87(2)