STATISZTIKAI „EGYPERCESEK”

AZ ALACSONYAN AGGREGÁLT STATISZTIKAI ADATOK ELEMZÉSÉNEK NÉHÁNY KÉRDÉSE*

A statisztikai modellezésben az utóbbi időben szinte egyeduralkodóvá váltak a makromodellek, a magas aggregáltságú adatokra épülő elemzések. Ez a némileg egyoldalú fejlődés legalább két problémát vet fel. Az első probléma módszertani vonatkozású, és a továbbiakban elsősorban ezzel kívánunk foglalkozni. Ez abban áll, hogy tudományos kutatási eredmény, új vagy újszerű módszertani alkalmazás csak a tudományos közvélemény által ellenőrizhető adatbázison történhet meg. Ennek a feltételnek a statisztika közgazdaság-tudományi alkalmazása területén szinte kizárólagosan a publikált, elsősorban makroszintű, vagyis az egész nemzetgazdaságokra, illetve azoknak magas ágazati szintre vonatkoztatott adatbázisai felelnek meg. Ebből az következik, hogy a módszertani fejlesztések nem a módszeralkalmazó területek sokszínű teljességéből, hanem csak egy meglehetősen leszűkített csoportjából mérik fel a fejlesztési igényeket, és próbálják a módszertani fejlesztést megfeleltetni ezeknek. Ezt a problémát még súlyosbítja az a tény, hogy míg jellemző módon a természettudományokban a kísérletek, (így például az orvostudományi alkalmazások területén egy gyógyszerkísérlet) reprodukálhatók és megismételhetők, a közgazdaság- és társadalomtudományi területeken erre nincs lehetőség, ezért különös jelentősége van a különböző területeken folyó modellezési munkák eredményei összehasonlításának, az analógiák keresésének.

A másik probléma, amelyik nem kevésbé súlyos, abból adódik, hogy egy terület – nevezetesen a vállalati adatokon nyugvó mikromodellezés – sérül, háttérbe szorul. Ez persze elsősorban az illető vállalatok gondja, de nem függetleníthető a modellezés általános kérdéseitől, egyrészt azért, mert egy fontos terület visszaszorulásáról, másrészt azért, mert a makroelemzések alapjainak gyengüléséről szól.

A bevezető állítást, nevezetesen azt, hogy az alacsonyan aggregált adatbázison történő módszeralkalmazások a szakirodalmi publikációkban alulreprezentáltak, három kiválasztott szakfolyóirat részleges elemzésével kívánjuk bizonyítani. A vizsgálódás kissé esetleges, bizonyára vannak e témában átfogó elemzések, tárgyszavakra épülő többszempontú csoportosítások, de célunk pillanatnyilag csak az, hogy felhívjuk a figyelmet egy létező problémára, ezért a három folyóirat (egy amerikai, egy nyugat-európai és egy hazai) egy-egy évet átfogó vizsgálata csak a kitűzött célt szolgálja.

A *Journal of the American Statistical Association* 451–454. számaiban megjelent tanulmányok közül mindössze egy (!) olyan volt, amely közgazdasági jellegű modellezéssel (a valutaárfolyamok elemzésével) foglalkozott, értelemszerűen a legmagasabb aggregáltsági szinten. Az *Allgemeines Statistisches Archív* 1995-ös számaiban megjelent cikkek tematikus megoszlását vizsgálva megállapítható, hogy nem volt olyan publikáció, amelyet az alacsonyan aggregált adatbázison történő elemzések közé lehetne sorolni. A 16 elméleti matematikai–statisztikai elemzés nem használ adatbázist, fejtegetései elméletiek. Természetesen a sokrétű e témájú német szakirodalomban találhatók gyakorlati indíttatású publikációk, de azok módszertani adaptációs kérdésekkel nem célzottan foglalkoznak.

Hasonló a helyzet a magyar szakirodalomban is. A statisztikai modellezésben meghatározó *Statisztikai Szemle* publikációinak megoszlása 2000-ben azt mutatta, hogy – néhány ágazati szemléletű tanulmány kivételével – alacsonyabb aggregáltságú adatbázissal dolgozó publikáció nem található. Ez tehát azt jelenti, hogy a szűkebb értelemben vett vállalati szintű elemzéseken túlmenően eggyel magasabb szinten már találhatók írások, hiszen például az ipari termékszerkezettel, az építőipar jellegzetességével, mezőgazdasági kérdésekkel és a

* A dolgozat a szerzőnek az Osztrák Statisztikai Társaság és az osztrák statisztikai hivatal (Statistik Austria) Statisztikai Hét című konferenciáján tartott előadásának rövidített, átdolgozott változata.

nonprofit szektor különböző kérdéseivel foglalkozó tanulmányok rendszeresen jelennek meg a *Statisztikai Szemlében*, de az igazi alacsonyan aggregált adatokra épülő vállalati elemzések hiányoznak. Az említettekhez ugyanakkor azt is hozzá kell tenni, hogy a társadalomstatisztikai körben (háztartás-statisztikai felvételek, időmérleg-felvételek, jövedelemfelvételek) gyakori a mikroadatokra épülő elemzés, jóllehet ennek módszertana olykor igen lényegesen különbözik a mikrogazdasági elemzésekétől. Hasonló tendenciák tapasztalhatók más magyar szakmai folyóiratok (*Szigma*, *Gazdaság és Statisztika*, *Területi Statisztika* stb.) esetében is.

Az okokat keresve mindenekelőtt az adatok hozzáférési nehézségeit kell említenünk. A statisztikai tevékenység törvényi szabályozása sokak szerint féloldalasra sikerült: az adatvédelemnek a kívánatosnál nagyobb súlya megnehezíti a statisztika helyzetét. Az 1993. évi (és 1999-ben módosított) törvény leszögezi ugyan, hogy a hivatalos statisztikai szolgálat által végrehajtott adatgyűjtések – meghatározott kivételektől eltekintve – nyilvánosak, de az említett kivételek elég jelentősek. Számunkra ezeknek leglényegesebb eleme az egyedi adatok nyilvánosságra hozásával kapcsolatos. A törvény rendelkezései szerint egyedi adat csak statisztikai célra használható, és nyilvánosságra csupán az adatszolgáltató hozzájárulásával hozható. Az elterjedt nézetekkel szemben tehát a törvényi szabályozás nem tiltja, csak engedélyhez köti a publikálást.

Vállalati adatokról lévén alapvetően szó, az előbbiek értelmében a vállalati szintű adatbázis alkalmazásának kettős akadálya van:

– a statisztikai hivatalok az előzőkből következően nem szolgáltatnak ki adatokat,
– a vállalatok pedig csak nagyon ritkán, meghatározott feltételek mellett járulnak hozzá adataik publikálásához.

Ez utóbbiak félelmei sok tekintetben indokoltak. Nem kell részletezni azt az esetet, ha nyilvánvaló ellentmondás alakulna ki a tényadatok közlése esetén az adóbevallás által vélelmezhető adathalmazhoz képest. A konkurencia úgy jutna adataikhoz, hogy a sajátját nem tárja fel, és ez neki mindenképpen előnyös, számára versenyelőnyt is adhat. Ezek a félelmek hozzájárulnak ahhoz, hogy pusztán elemzési célra, anonim publikáláshoz, de még külső szerv által végzett nem publikációs céllal (tudományos kutatási, módszertanfejlesztési, modellkísérleti stb.) elvégzett vizsgálatokhoz sem szívesen biztosítanak adatbázist a gazdálkodó egységek.

Mint azt tapasztalhattuk, az adatok megszerzése is nehézségekbe ütközik, ám az adatbeszerzés nehézségeinek legyőzése után az eredmények közlése, szakmai megvitatása, tudományos teljesítményként való elismertetése sem problémamentes, hiszen módszertani kísérletekre, új modellek bemutatására csak ellenőrizhető adatbázisokon alapuló publikációkat fogad el a szakma, a vállalatok pedig általában nem járulnak hozzá a saját adatbázisaikon alapuló esetlegesen létező modelleredmények közléséhez sem.

A vállalatok félelme gazdasági tevékenységüket leíró modelleredmények publikussá válása kapcsán sem alaptalan, és az okok is hasonlók, mint az alapadatok nyilvánosságra kerülésénél. Mindezekhez hozzájárul még az is, hogy az esetlegesen prognosztizálható modelleredmények aktív beavatkozást tesznek lehetővé a konkurenciának a piaci versenyben.

Azt, hogy az alapprobléma szempontjából közelebbről mindez mit jelent, két példával szeretnénk illusztrálni. Első példánk, ahol az alacsony aggregáltságú adatbázisok sajátosságait taglaljuk, a szezonalitást is tartalmazó összetett idősori modellezés. Ezen a területen egy sor jól kidolgozott és bejáratott modell, illetve szoftver létezik (a klasszikus Census II, ARIMA -X-11, ARIMA-X-12, SEATS, TRAMO) melyek az első pillantásra univerzálisan felhasználhatóknak tűnnek, olyan interaktív beavatkozási lehetőségekkel, amelyek teljes mértékben kielégítik az alacsonyan aggregált adatbázisokon történő elemzési igényeket is. Nem szabad azonban szem elől téveszteni, hogy ezeknek a módszereknek számos alváltozata létezik, illetve az interaktív beavatkozási lehetőségek a szakirodalomban még nem kellően feltártak. Ebből a szempontból fontos lenne vizsgálni az alacsonyan aggregált adatbázisok sajátos igényeit is. Ezek közül csak a legfőbbeket említve:

– a szezonális kiigazítás versus trendbecslés kérdését illetően vállalati és ágazati szinten nagyobb az igény a szezonális kiigazításra, a véletlen hatást is magában foglaló rövid távú előrevetítés elsődleges cél lehet;
– az idényszám-variabilitás kérdéskörében a kidolgozott módszerek – bár tudják kezelni például a 6 napos munkahét hatását is – alapvetően a havi bontásos idősorok elemzésére kerültek kimunkálásra. A vállalati szintű elemzésekkor – de esetleg nemzetgazdasági szinten is – fontos lehet az alapvetően 5, 6 munkanapos vagy 7 napos, illetve speciális esetben (energiagazdálkodás, közlekedés, kiskereskedelem) 12, illetve 24 órabontásos elemzés;
– az outlier-kezelés, munkanapszám-váltás, szökőévhatás, ünnepszámváltás, húsvét, pünkösd szerepét tekintve vállalati, illetve ágazati szinten lényegesen nagyobb jelentősége van az outliereknek, hiszen makroszinten érvényesül ezek egymást kioltó, elimináló hatása. A módszerek alapvetően az outlier kiszűrésre törekszenek. Vállalati szinten sokkal nagyobb a szerepe ezek tipizálásának.

A másik példát csupán érintőlegesen említjük. A valutaárfolyamok idősoros vizsgálatánál az árfolyamok közötti kointegráció gyakran nem mutatható ki. Amennyiben meghatározott portfolió részei, tehát magasabb

aggregáltsági szintre kerül az adatbázis, szignifikáns kointegráció lép fel, illetve mutatkozik a modellezés során. Eliminálódik az outlierek módosító hatása, ami azt jelenti, hogy a mikro- és a makroelemzéseket egészen más kiindulópontból kell kezelni.[1]

A következtetések összefoglalása, a diagnózis elég egyértelmű: a makroszintű adatbázisok aránytalanul nagy szerepet kapnak a statisztikai modellkísérletek során, féloldalassá teszik a gazdaságelemzés statisztikai módszertanát. A terápia azonban már korántsem ilyen egyszerű, és jószerivel csak ötleteket lehet felsorolni.

– Input oldalról meg kell győzni a vállalatokat, hogy bizonyos esetekben, nagyon megfontoltan kiválasztott adatbázis esetében, a reklámérték, a publicitás adta előny nagyobb lehet a hátrányoknál.

– Output oldalról érv lehet, hogy a sikeres tevékenység, ami a modelleredményekben tükröződik, megerősítést adhat a gazdasági kapcsolatok során a partnereknek, esetleges részvényvásárolóknak, a finanszírozóknak.

– Élni kell az adatok anonimizálásának egyre terjedő és bővülő eszközeivel, módszereivel, amelyek lehetőséget adnak arra, hogy az egyedi információk igen nagy részét az egyediség felfedése nélkül fel lehessen használni. Ennek érdekében népszerűsíteni kellene ezeket a módszereket, hiszen vélhető, hogy ha a vállalatok megismerik, és elhiszik, hogy korábban említett félelmeik alaptalanok, talán meglátják az elemzések számukra is hasznos oldalát.

– A módszertani fejlesztéseket illetően támogatni kell a gazdasági egységeket alapul vevő gazdaságmodellezési tevékenységet. Ennek során potenciális támogatók lehetnek az akadémiák, tudományos intézetek, alapítványok, szakmai társaságok, például a statisztikai társaságok. Sokat segíthetnének a helyzet javításában a szakfolyóiratok is, tematikus számok megjelentetésével.

Herman Sándor

[1] *Körősi G. – Longmire, R. J. – Mátyás L. – Clayton, M.* (1996): *Aggregation and cointegration*. Monash University. Melbourne.