A kongruens modellezés információs többletérték-termelő potenciálja

(Potential of the information added value based on congruent modelling)

Pitlik László, Pitlik Marcell, MY-X team

Kivonat: A modellezés kapcsán a matematikai apparátussal az információs többletérték-termelés szempontjából quasi egyenértékű (megkerülhetetlen) maga az az adatvagyon, amit fel kell dolgozni. Adat nélkül sincs eredmény és matematika nélkül sincs eredmény a modellezésben. A klasszikus nyers adatok (attribútumok) feldolgozását a származtatott változók bevezetésével sikeresebbé lehet tenni pl. a termelési függvények pontosságát illetően. A származtatás már ismert context free esetei a dupla-attribútum-készletű modellek, ill. a nyers attribútumok összegeit, különbségeit, arányait feldolgozó modellek (valamint ezek hibridizálása – vö. additív és/vagy multiplikatív modellek). Speciális (immár az objektumokat is érintő) származtatást jelent az input-objektumok attribútumonkénti különbségeinek új objektumként való értelmezése, különösen akkor, ha egyes objektumok rel. kis arányban fordulnak elő (vö. cégcsőd vs. élő cégek), de a speciálisan ritka objektumok bármelyike bármely többségi objektummal párba állítható és így a speciális objektumok hatása, aránya tetszőlegessé tehető. Az itt bemutatásra kerülő kongruencia-alapú modellezés során a nyers attribútum-értékek által alkotott OAM-mal azonos méretű, adott osztó melletti maradékosztályok értékei kerülnek az adatpozíciókba. Az eltérő osztókhoz tartozó modellek egyike/másika és/vagy zárómodellje képes meghaladni a nyers modell pontosságát – sőt, mindezt akár szűkített attribútum-készlet esetén is realizálni lehet – hiszen a több adat jobb modell elv nem triviálisan igaz. Amennyiben a következmény-változók is a nyers Y maradékosztályait tartalmazzák eltérő osztók esetén, akkor átlépünk a konzisztencia-alapú modellezés egy speciális területére, ahol a nagyobb osztók és ezek tényezői egymással szoros kölcsönhatásban állnak.

Kulcsszavak: hasonlóságelemzés, GPS, következtető gép, oszthatóság, maradékosztály

Abstract: The information added-value can be catalysed not only through different types of functions - y=f(Xi) – but also through the data being available. Without data, models can not be evaluated and without functions models can not be realized. The classic (raw) data can always be processed but it is also possible to derive new inputs based on the raw data and it is a fact that these new attributes are capable of increasing the modelling accuracy of production functions. New attributes can be derived in unlimited ways: it is possible to use doubled attributes where the raw attributes will be ranked in direct and inverse ways. It is also possible to make simple connections between the raw attributes like additions, multiplications, etc. Besides, the already listed possibilities can also be hybridized. Specific derivation is, if not only the attributes will be manipulated but the objects too: e.g., new objects can be built if the attributes-specific differences of two raw objects will be calculated. This technique is capable of supporting problem solving procedures with a little ratio of object types (see e.g., fraud detection challenges). The object-pairs can be selected so, that the low frequently elements are always one member in the pairs. Here and now, the congruence-based modelling will be observed whether a kind of information added value can be detected? This approach derives new OAMs from the raw OAM: the new OAMs has the same volume, but the cells consist residue-values after a division by a lot of divisors. It will be expected that – based on a random OAM – the information added value can be realized through congruent OAMs and/or through their combinations. It is also expected that the input-demand can be reduced without decreasing of the model accuracy because more data lead not to better models in general. Finally, the consequence variable (Y) can also be divided by different divisors and the new Ys can also be modelled. The partial results make possible to create a typical consistence problem where the similarity analysis can be seen as an inference machine.

Keywords: similarity analysis, GPS, inference machine, divisibility, residue-classes

# Bevezetés

Előzmények:

* <https://miau.my-x.hu/miau/271/e_money.docx> (vö. 2019 becslése 2015-6-7-8 éves és a 4 évet összegző zárómodell alapján)
* <https://miau.my-x.hu/miau/268/ceteris_paribus_domestication.docx>
* <https://miau.my-x.hu/miau/268/context_free_hermeneutics.docx>
* <http://miau.my-x.hu/miau/198/veletlen_e_a_veletlen_v2.docx>
* ...

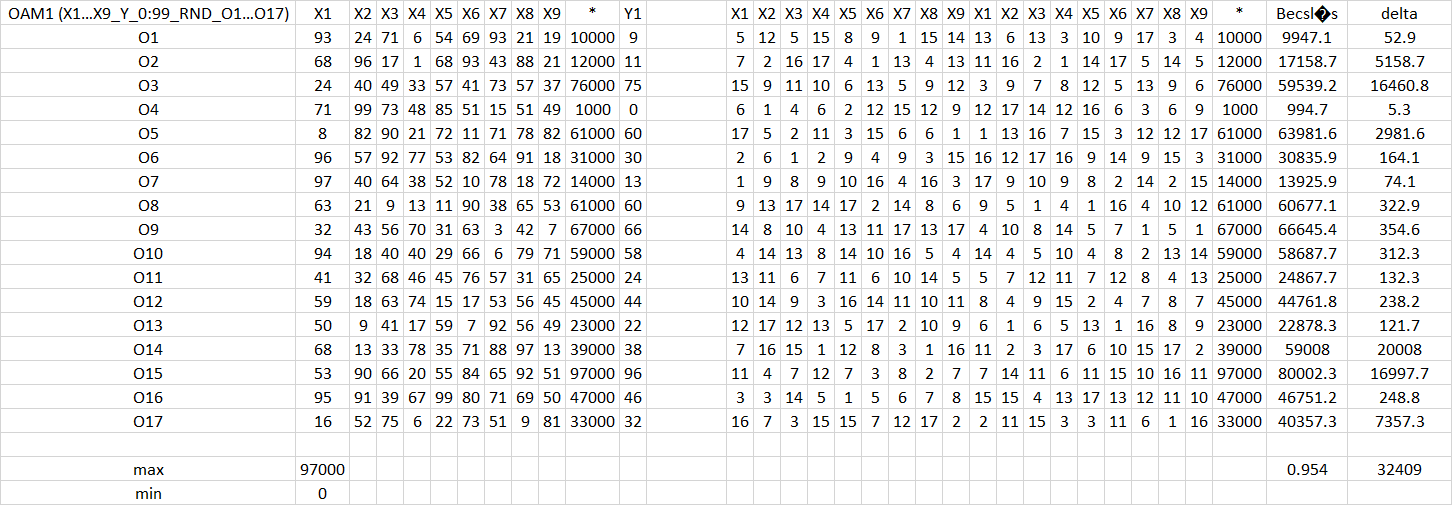
A GPS (general problem solving) egyik közelítése az OAM- (objektum-attribútum-mátrix)-alapúság, ahol az objektumok egymással összevethető (hasonló) esetek, s ezek egymáshoz képesti viszonyait (pl. lépcsős) függvényekbe sűrítő eljárások olyan kérdésekre is választ képesek adni, melyek adott pillanatban csak becsülhetők, s objektíven talán soha meg nem válaszolhatók. Ilyen kérdések természetesen az előrejelzések – melyek azonban a megfelelő idő elteltével objektív ellenőrzésen eshetnek át. S főleg ilyen kérdés a rendszerek genetikai potenciáljának mibenlétét firtató kérdés is, ahol a cél az ismert rendszer-outputok alapján azt megbecsülni, létezhet-e az ismert Y-maximumnál nagyobb output-érték? A tanulási minták optimalizálásakor a genetikai potenciál nem más, mint az az egyrétegű, és/vagy aggregált modell-helyességi mutató-érték, mely jelzi, hogy a rendelkezésre álló adatok valamely részhalmazaihoz tartozó modellezési sikeresség ismeretében elvárható-e, hogy létezzen olyan részhalmaz (optimális tanulási minta), melyre vonatkozóan a teszthelyesség nagyobb, mint a már ismert eddig legnagyobb teszthelyesség (mindenkor azonos tesztadatokra vonatkozóan). Amíg a genetikai potenciált jelentő objektum nem kerül bele a tanulási mintába, addig a genetikai potenciál-számításoknak jelezniük kell, hogy létezik jobb megoldás, s amint az ideális variáns a tanulási minta részévé válik, akkor a genetikai potenciál-számításnak azt kell kimutatnia, hogy nincs további esély az output-érték növelésére. Ilyen genetikai potenciál jellegű valós helyzetek pl. a Forma-1-es legjobb köridő kérdése, vagy éppen a genetikai potenciál fogalmát adó mezőgazdasági termelési potenciál kérdése, ahol keressük, lehet-e teoretikusan pl. több kukoricát termelni az ismert terméseredmény-maximumot meghaladóan?

A modellezés másik alapvető/klasszikus kérdése az azonos modell-struktúrák melletti tanulási pontosság maximumának becslése. Vagyis mennyi az adatokban (adott OAM-ban) rejlő mintázat-potenciál? A mintázat-potenciál polinomok modellezésben való felhasználása és ellentmondás-mentes alapadatok esetén mindenkor 100% illik, hogy legyen. Ha a modellező olyan modellt akar, ahol a polinomizáció foka korlátozott, akkor jutunk el a kiindulási állapotba, ahol adott egy OAM és adott egy modell-típus (pl. lépcsős függvény), s keressük a nyers Xi alapján az Y becsülhetőségének maximumát.

Már itt fel kell hívni a figyelmet arra, hogy a korábban jelzett származtatásra ugyanazon polinomizálást korlátozó elvek kell, hogy érvényesek legyenek, mint a modellezésre magára. Vagyis például a nyers Xi és a SIN(Xi) kapcsolatrendszer alapvetően már polinomizálásnak számít.

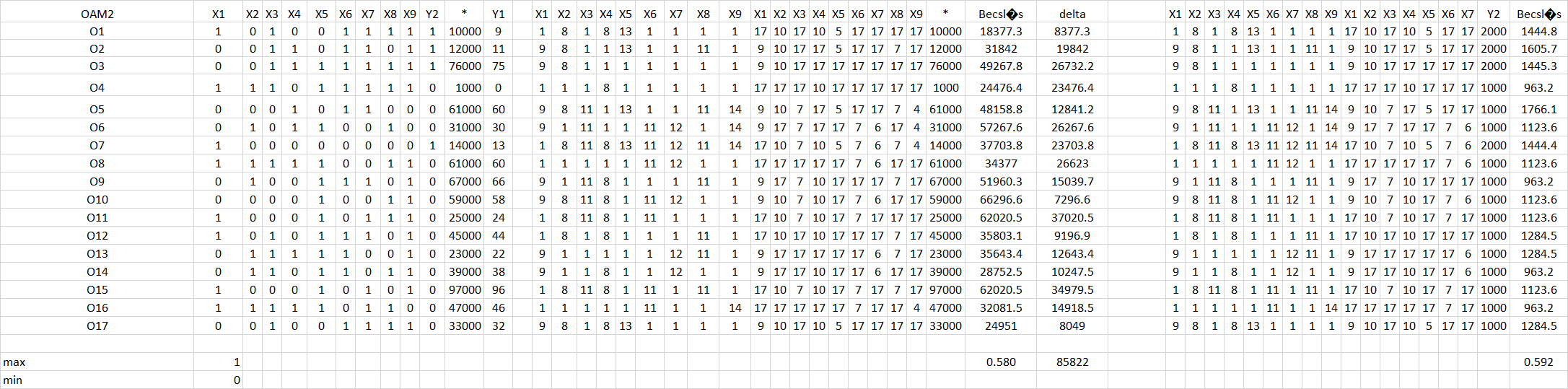
# A kísérleti adatvagyonról

Részletek: <https://miau.my-x.hu/miau/271/oam_kongruencia.xlsx>

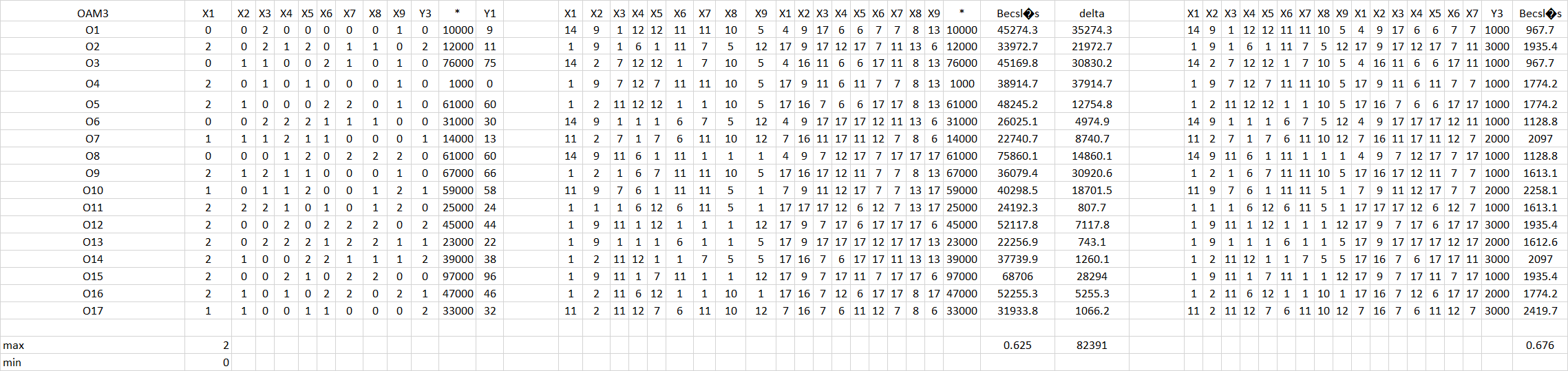


1. Ábra: A kísérleti véletlenszám-generált adatvagyon (forrás: saját ábrázolás)

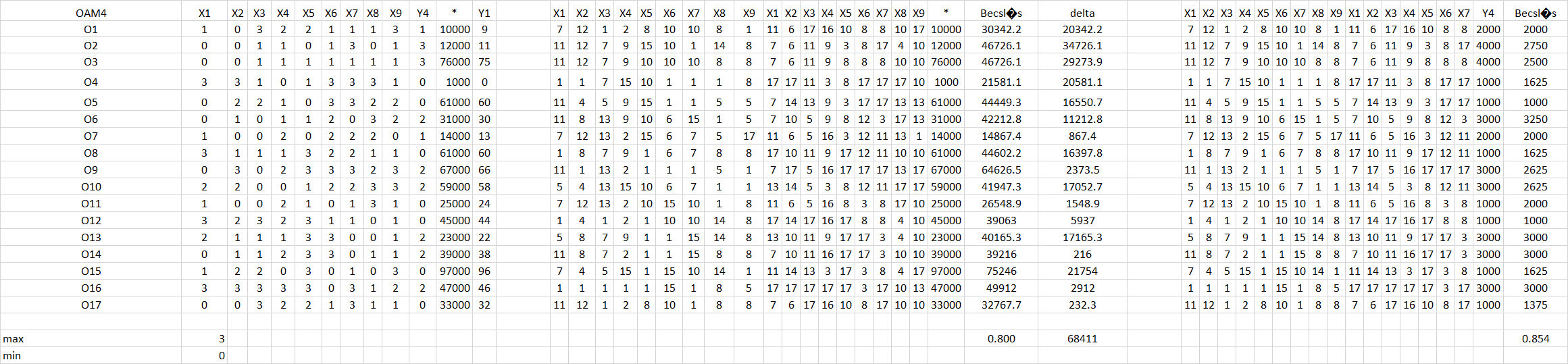
Az 1. ábra egy minden tekintetben (vö. objektumok száma, attribútumok száma, maximum és minimum értékek mibenléte) teljes mértékben véletlennek tekinthető állapotot mutat be, ahol a hasonlóságelemzés elvárásai szerint az Y=0 érték esetén eltoltást, s a rel. alacsony Y értékek miatt nagyítást kell végrehajtani (vö. \*-oszlop = Y(nyers)\*1000+1000) – s ez érvényes az Y maradékként való értelmezésekor is.



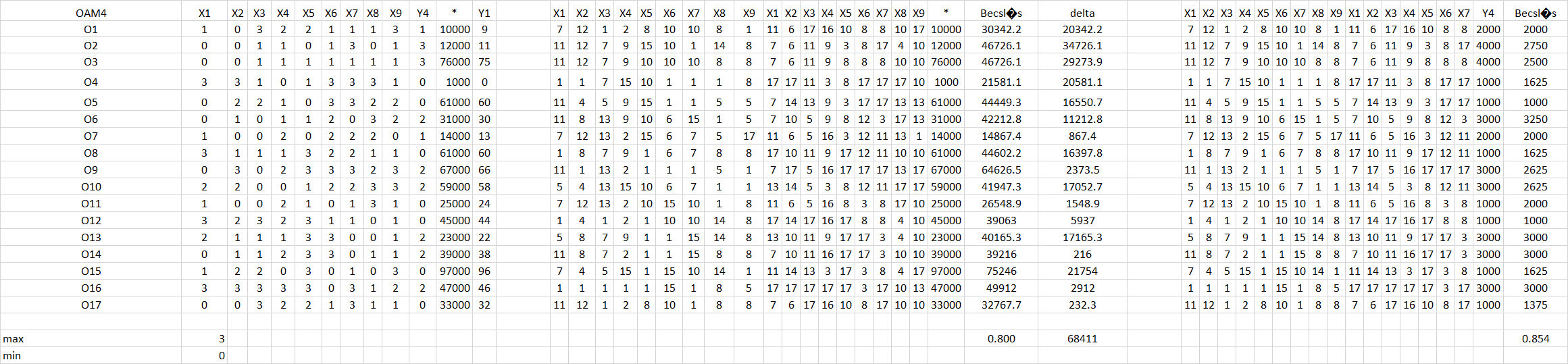
1. Ábra: OAM kettővel való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



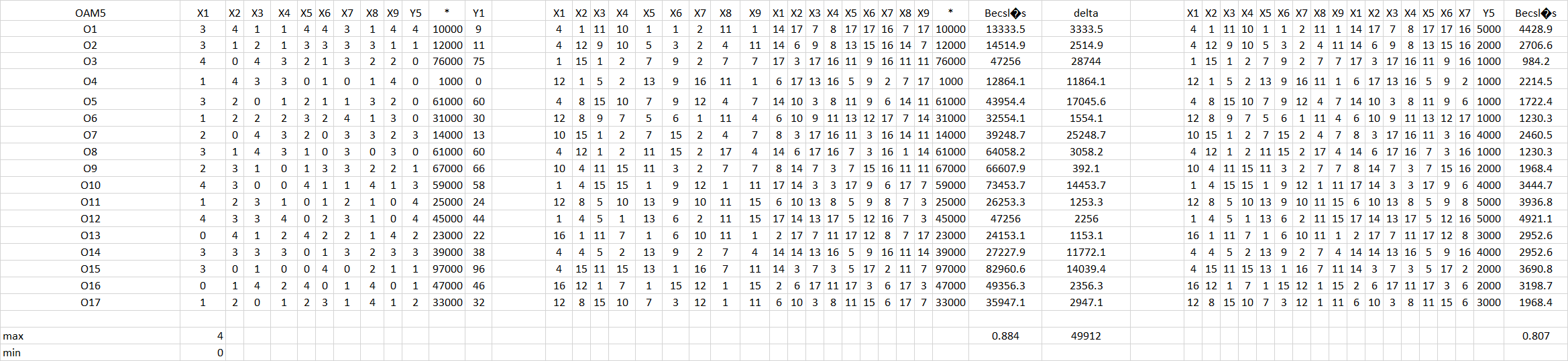
1. Ábra: OAM hárommal való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



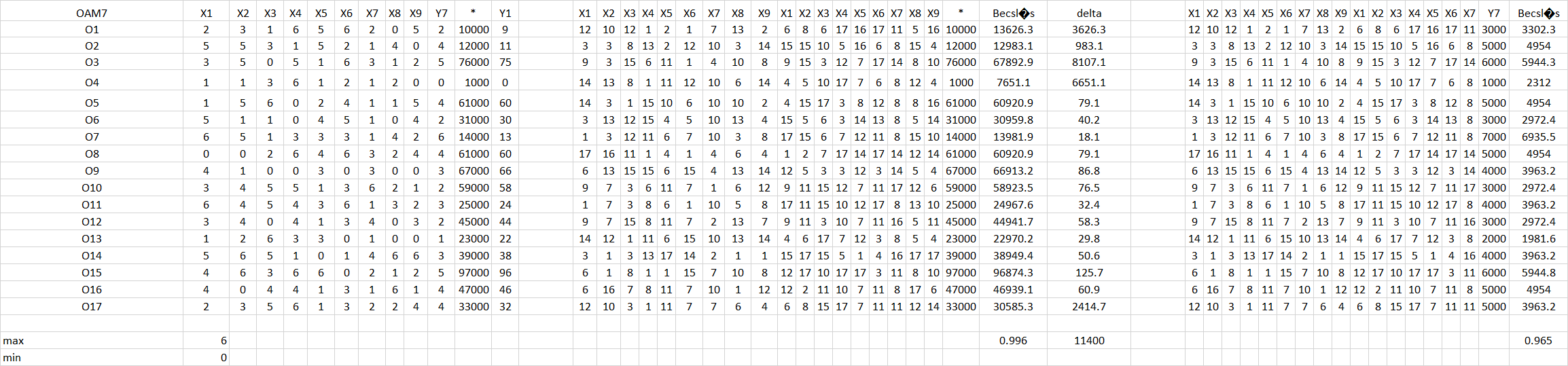
1. Ábra: OAM néggyel való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



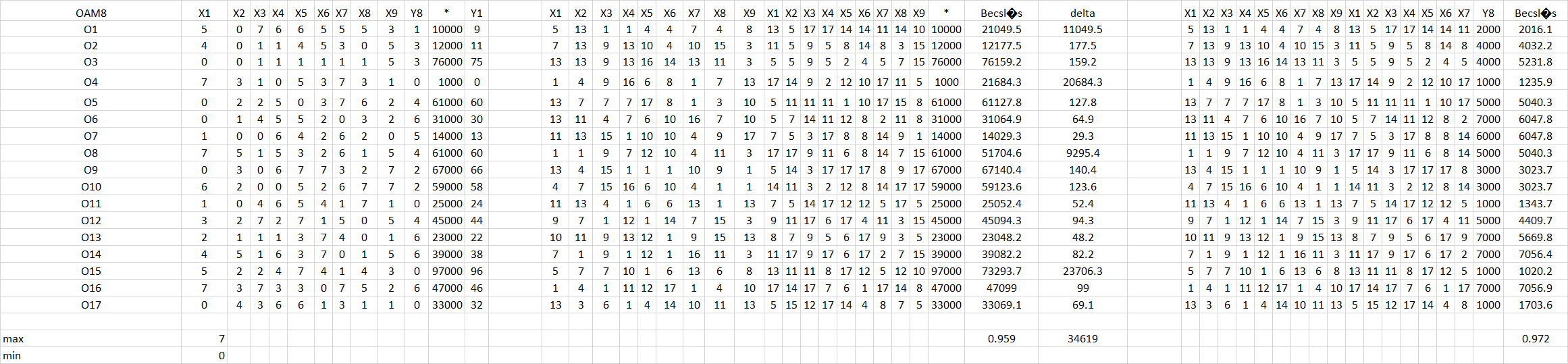
1. Ábra: OAM öttel való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



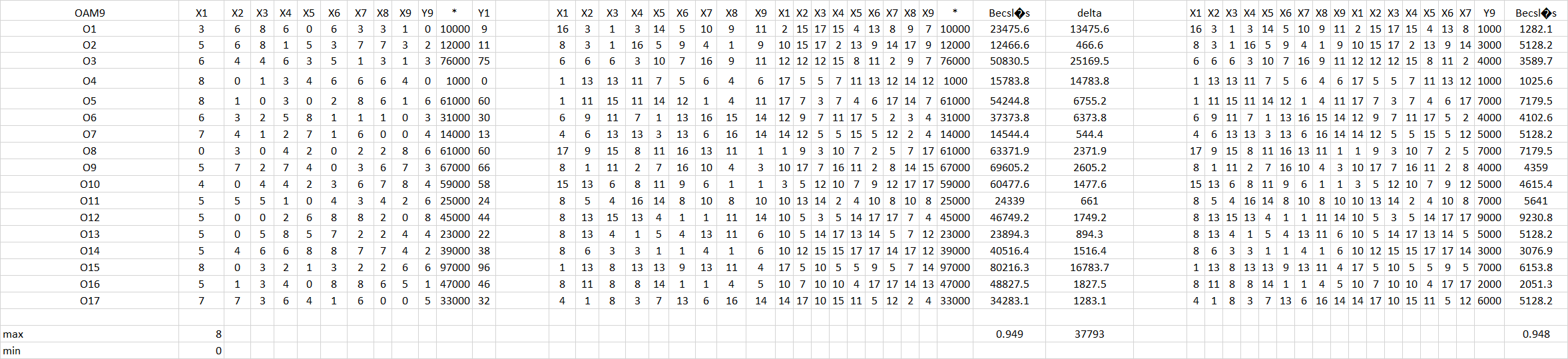
1. Ábra: OAM hattal való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



1. Ábra: OAM héttel való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



1. Ábra: OAM nyolccal való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)



1. Ábra: OAM kilenccel való osztás utáni maradékosztályokkal (forrás: saját ábrázolás)

Már itt utalni kell arra, hogy az a paramétersor, miszerint hány osztót és melyeket érdemes egy modellbe bevonni önálló optimalizálási problémaként egy önálló publikációban értelmezendő.

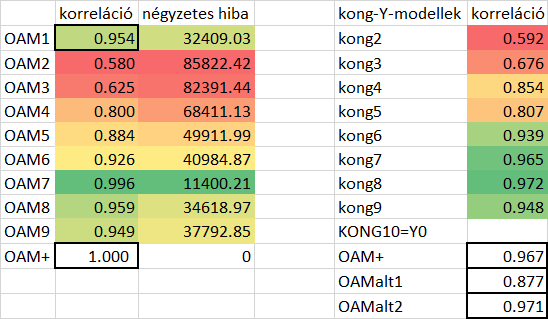
A 2-3-4-5-6-7-8-9. ábrák alapján jól látható, hogy a maradékosztály lényegében sorszámként értelmezhetők, ahol a kongruens, azaz egybeeső tulajdonság nem a numerikus szomszédság elvére alapoz, mint a normál sorszámozás esetében.

Hasonlóképpen itt kell jelezni, hogy a nyers Y, ill. ennek transzformált változata (\*-oszlop) mellett a nyers Y maradékosztálya is lehet outputja egy modellezési folyamatnak, ahol a mindenkori Y-t lényegében a rá jellemző oszthatósági viszonyokkal igyekszik leírni a modellező szakember. Ez a konzisztencia-orientált alapvetés egyrészt figyelembe veheti az oszthatóság tényét (meglétét) és nem teljesülését, ill. akár a maradékosztályokra vonatkozó becsléseket is a legvalószínűbb nyers Y maradékosztályokból történő levezetéséhez, ahol a hasonlóságelemzés maga egy fajta következtetőgépet lehet képes pótolni.

# Eredmények

Az eredmények kapcsán az alábbi kérdésekre keressük a válaszokat:

1. Milyen pontossággal modellezhető a nyers Y a nyers OAM alapján? (OAM1)
2. Milyen pontossággal modellezhető a nyers Y a kongruens OAM-ok alapján? (OAM2…9)
3. Milyen pontossággal modellezhető a nyers Y a kongruens OAM-okra alapozó becslésekből kiindulva? (OAM+)
4. Milyen pontossággal modellezhető az Y maradékosztálya kongruens OAM-ok alapján? (kong2…9)
5. Milyen pontossággal készíthető nem kongruens alternatív modellek?



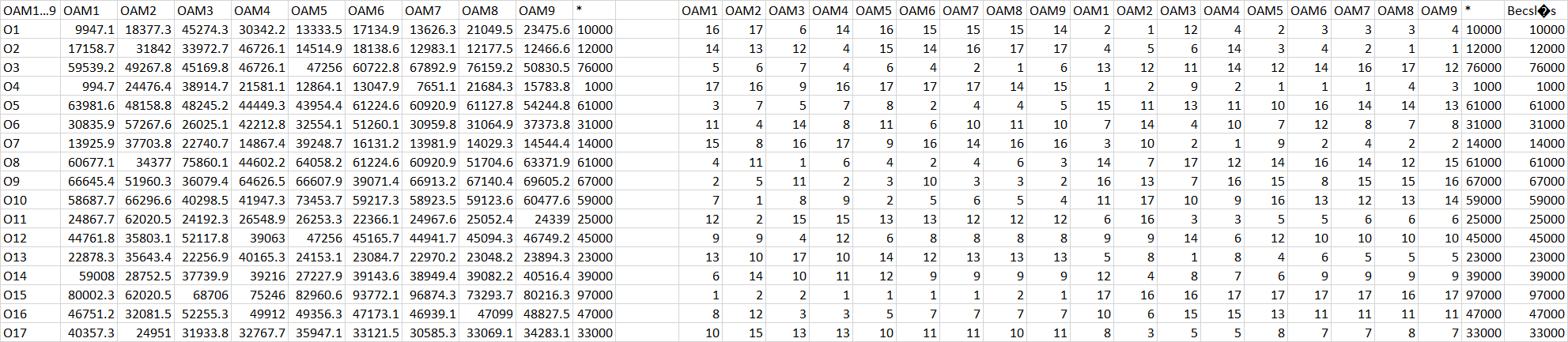
1. Ábra: A kongruens OAM-ok alapján nyers Y és Y-maradékosztályok becslése (forrás: saját ábrázolás)

A 10. ábra értelmezéséhez tudni kell, hogy minden modell standard hasonlóságelemzés volt dupla (direkt és inverz) attribútum-készlettel (azaz domesztikált polinomizáció mellett).

Az OAM1 korrelációja a tények és a becslések között 0.954 volt, ami nagyon magas érték véletlen adatminta esetén különösen (vö. 1. kérdés).

De itt és most a kérdés csak az, vajon létezik-e információs többletérték a kongruens modellek (2. kérdés) és ezek zárómodellje (3. kérdés) keretei között? A válasz triviális: OAM7 és OAM8 esetén OAM1 becslési pontossága meghaladható volt – de csak OAM7 esetén volt a négyzetes hiba (gyöke) is kedvezőbb. Így az ideális konkruens mintázata a nyers OAM-nak a 7-es osztó esetén állt elő – már önmagában is információs többletértéket hordozó módon (vö. 2. kérdés).

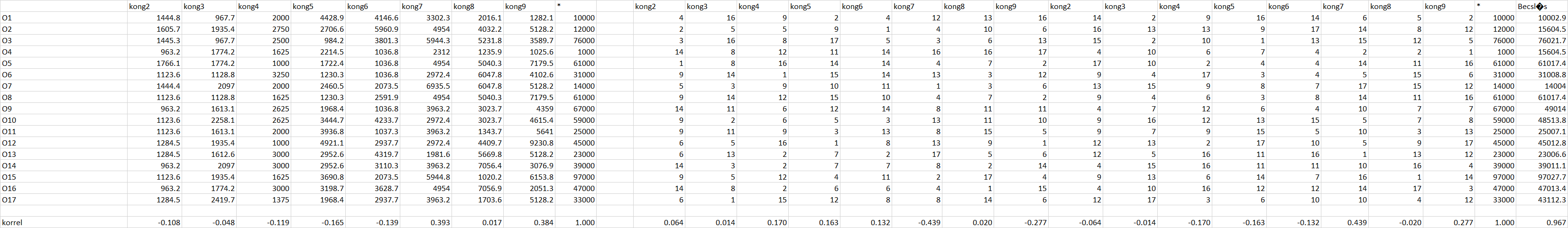
A 3. kérdés kapcsán az OAM+ modell (vö. 11. ábra) az OAM1…9 alapján kell, hogy az OAM7 esetében látható 0.996-os korrelációt meghaladja – s meg is haladja, vagyis ezen szinte már értelmezhetetlen becslési pontossági szinten is létezik az információs többletérték. Ennek egyszeri léte is elegendő annak bizonyítására, hogy a kongruens modellezés hasznos lehet.



1. Ábra: AZ OAM+ és ennek eredményei (forrás: saját ábrázolás)

Értelmezési támogatásként megjegyzendő: pl. a 9. ábrán a balról első Becslés oszlop (kép közepe) O1 sorának 23475.6-os értéke látható a 11. ábra O1 sorának és OAM9 oszlopának cellájában is.

\*\*\*



1. Ábra: Y-maradványosztályok becslése kongruens OAM-ok alapján (forrás: saját ábrázolás)

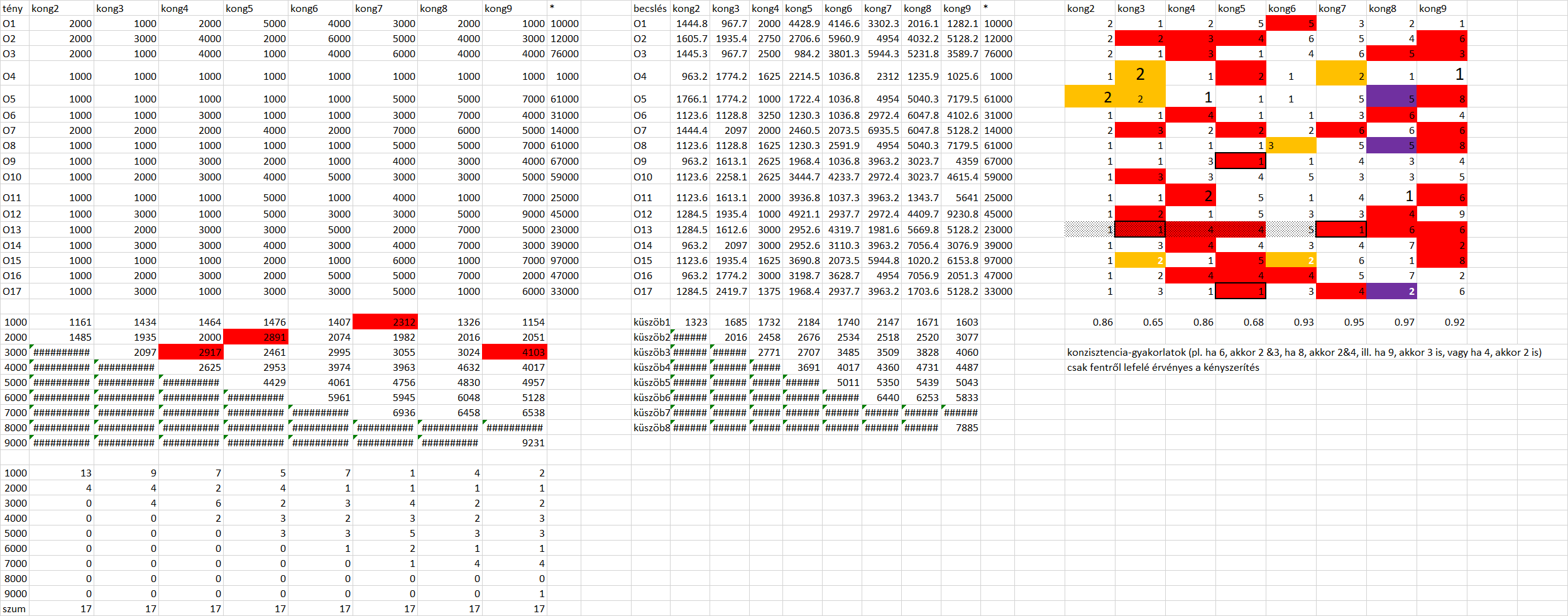
A 12. ábra kapcsán ellenőrzésként belátható, hogy a 12. ábra O1 sorának és kong9 oszlopának 1282.1-es értéke azonos kell, hogy legyen a 9. ábra jobb felső értékével.

Itt kell megjegyezni azt is, hogy az Y maradványosztályokat osztónként becslő értékek nagyon alacsony korrelációt mutatnak a nyers Y maradékosztályaival (vö. 12. ábra alsó sor), s mégis, ezen látszólag értéktelen inputok eredője egy 0.967-es korreláció a nyers Y becslésekor.

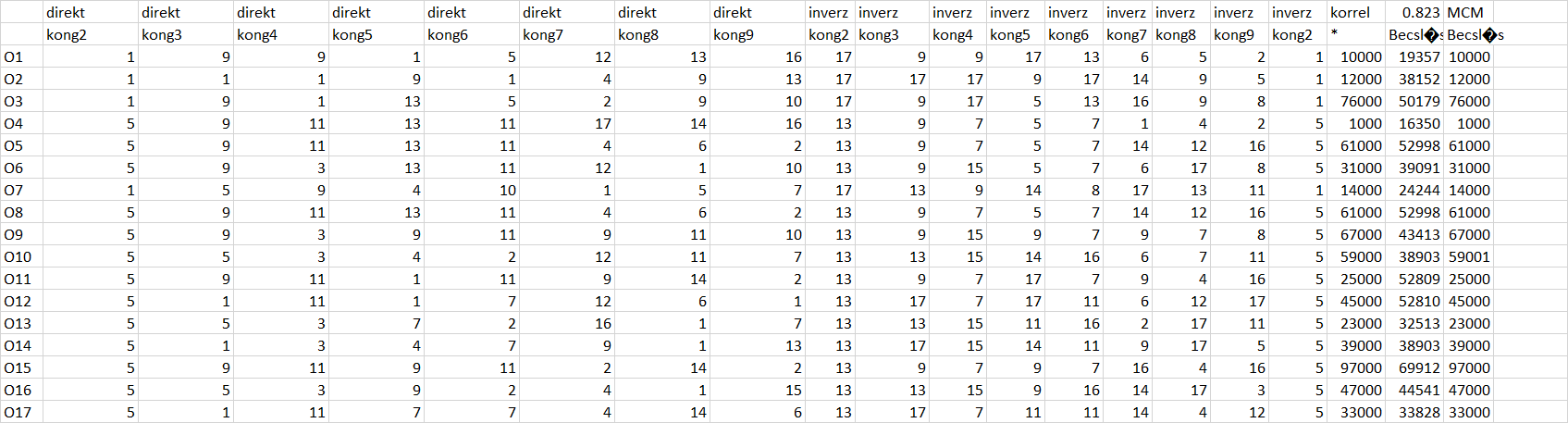
\*\*\*

A 13. ábra arra mutat példát, hogy az oszthatóság tényének modellezése és az osztók közötti kölcsönhatások milyen szabályszerűségeken keresztül milyen hibás becslésekre mutatnak rá.

A 14. ábra a 13. ábra bal felső OAM-jából indul ki.A 12. ábra a 13. ábra középső OAM-jából indul ki (1444.8). Mint látható: az Y nyers értékeinek tényleges maradékosztályai és a becsült maradékosztályokra alapozó modellek közül a becslések, mint inputok lényegesen jobb korrelációt mutatnak fel (0.967>0.823). Ennek oka, hogy a becslések közel sem olyan kevés distinct értékkel dolgoznak, mint a tények.



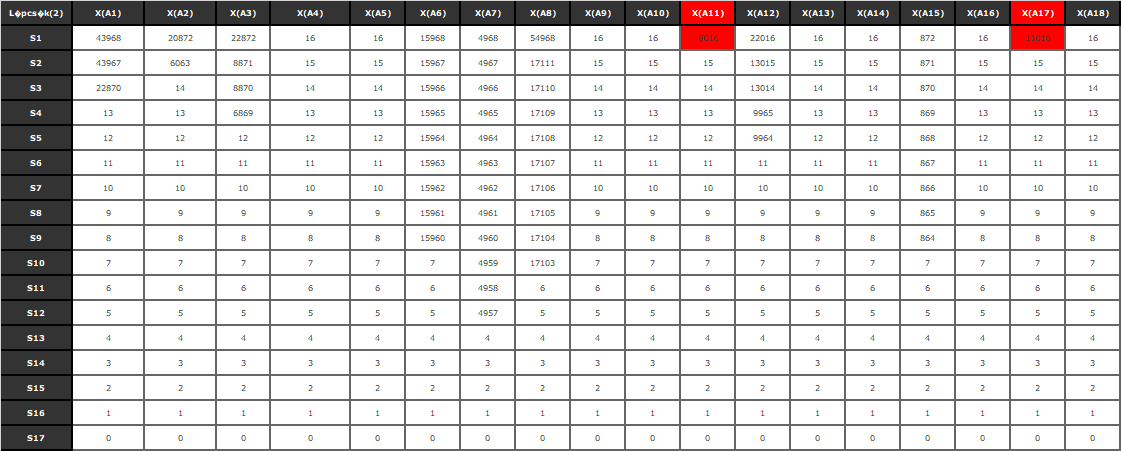
1. Ábra: Az oszthatóság hatása a modellezhetőségre konzisztencia-alapon (forrás: saját ábrázolás)



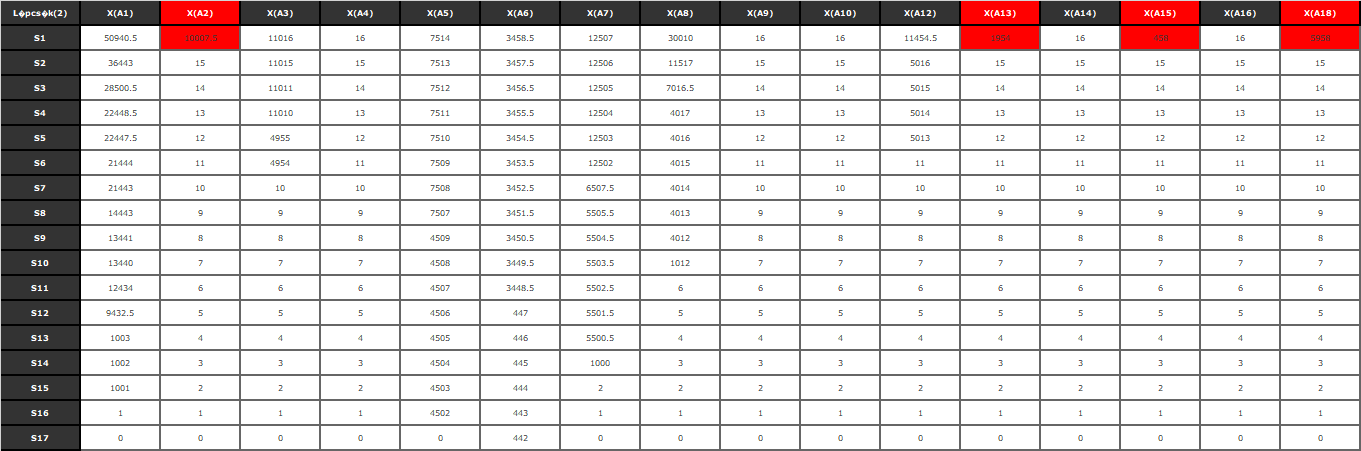
1. Ábra: Az Y valódi maradékosztályainak inputként való felhasználása az Y nyers értékeinek becslésére (forrás: saját ábrázolás)

\*\*\*

Az OAM1 esetében a dupla-attribútum-készletű modell kihasználta a lépcsős függvény specialitását, az S1-dominanciát.



1. Ábra: Az OAM1 modell attribútum-szűkítésre alkalmas változói (forrás: saját ábrázolás)



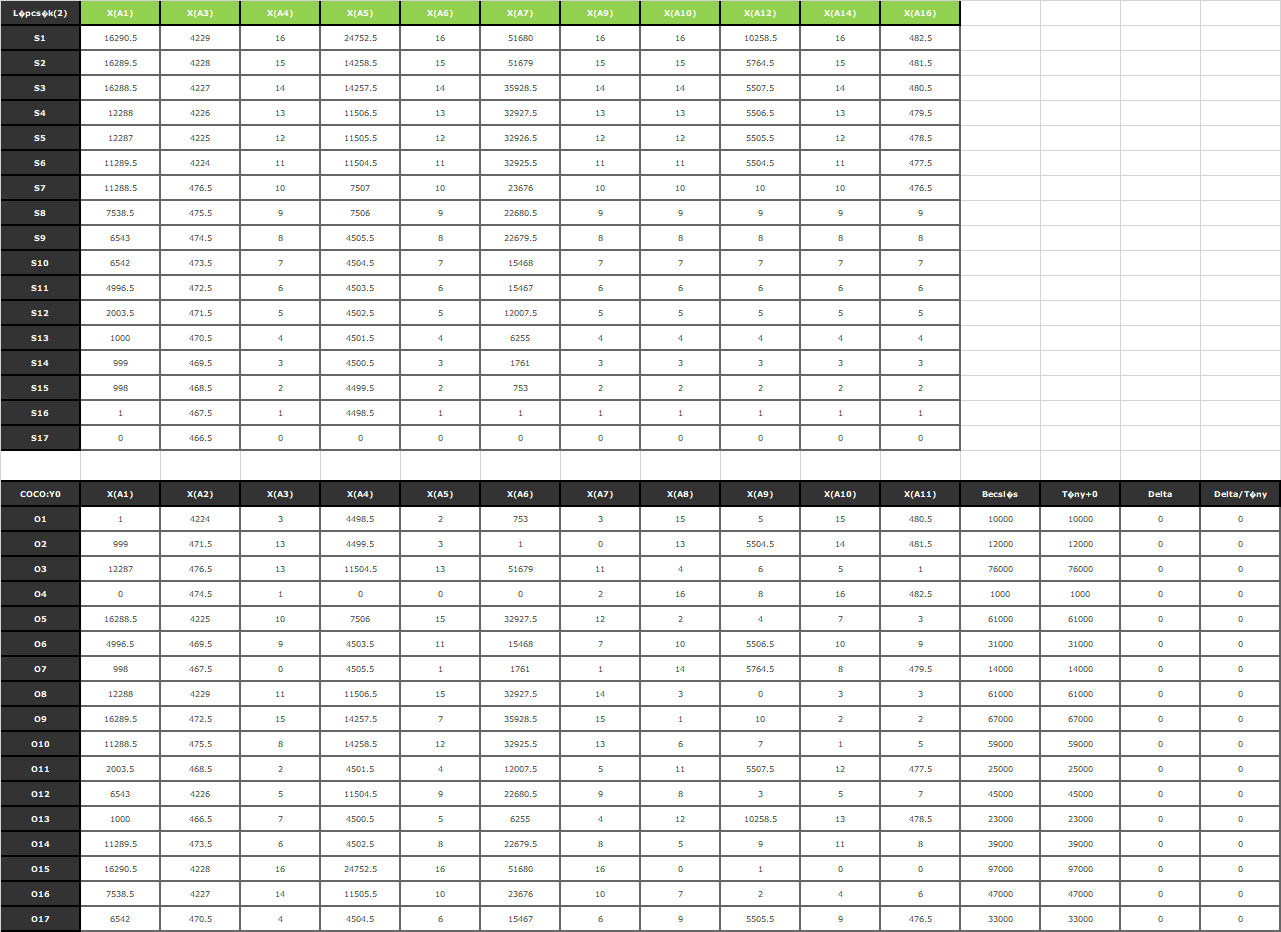
1. Ábra: Az OAM1 második szűkítési halmaza (forrás: saját számítások)



1. Ábra: Az OAM1 harmadik szűkítése (forrás: saját ábrázolás)

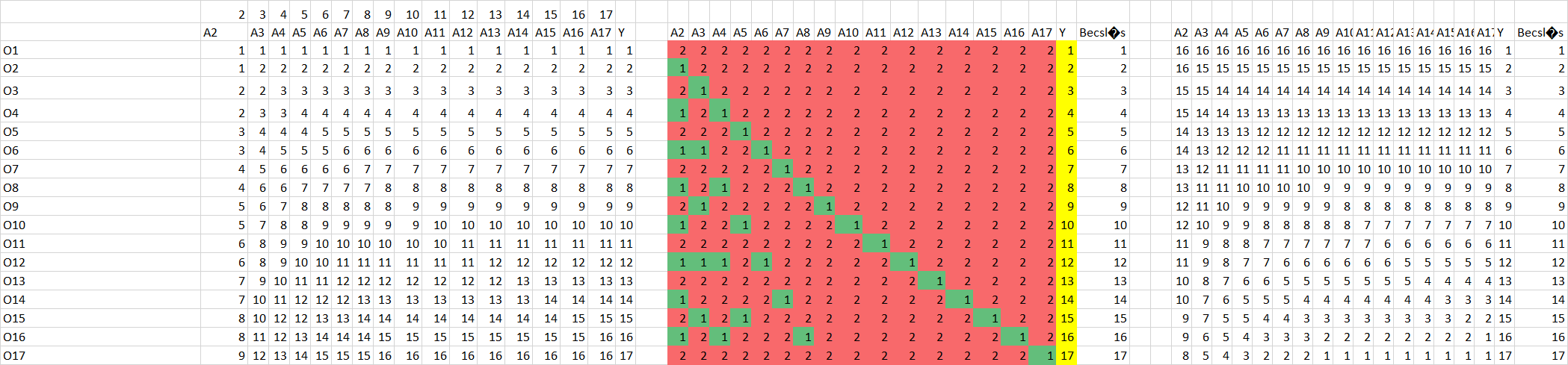
A 15-18. ábrák alapján látható, hogy az eredetileg 2\*9=18 attribútumos rendszerből 2+4+1=7 attribútum kiesett. Az A2(direkt):A11(inverz) páros teljesen kiesett – vagyis az eredeti OAM2 attribútumnak (vagyis a 2-vel való osztásnak) nincs érdemi (sem direkt, sem inverz) hatása arra, hogy a nyers Y tökéletesen előrejelezhető legyen. A nyers Y modellezésre elegendő 11 megmaradt változóból 7 direkt és 4 inverz, s a 11 változóból 5 darab hatástalan (vö. S1=16). Vagyis 7-3=4 direkt változó hatásos, ill. 4-2=2 inverz hatásos. S a direkt hatás mennyisége lényegesen meghaladja az inverz hatások mennyiségét.

Ez egy fajta hatékonyságnövelő modellezés, mely célja, hogy az alternatív lépcsős függvények között célirányosan redukáljon input-igényt úgy, hogy a modellezés pontossága ne változzon.



1. Ábra: Az OAM1 végső lépcsős függvénye (forrás: saját ábrázolás)

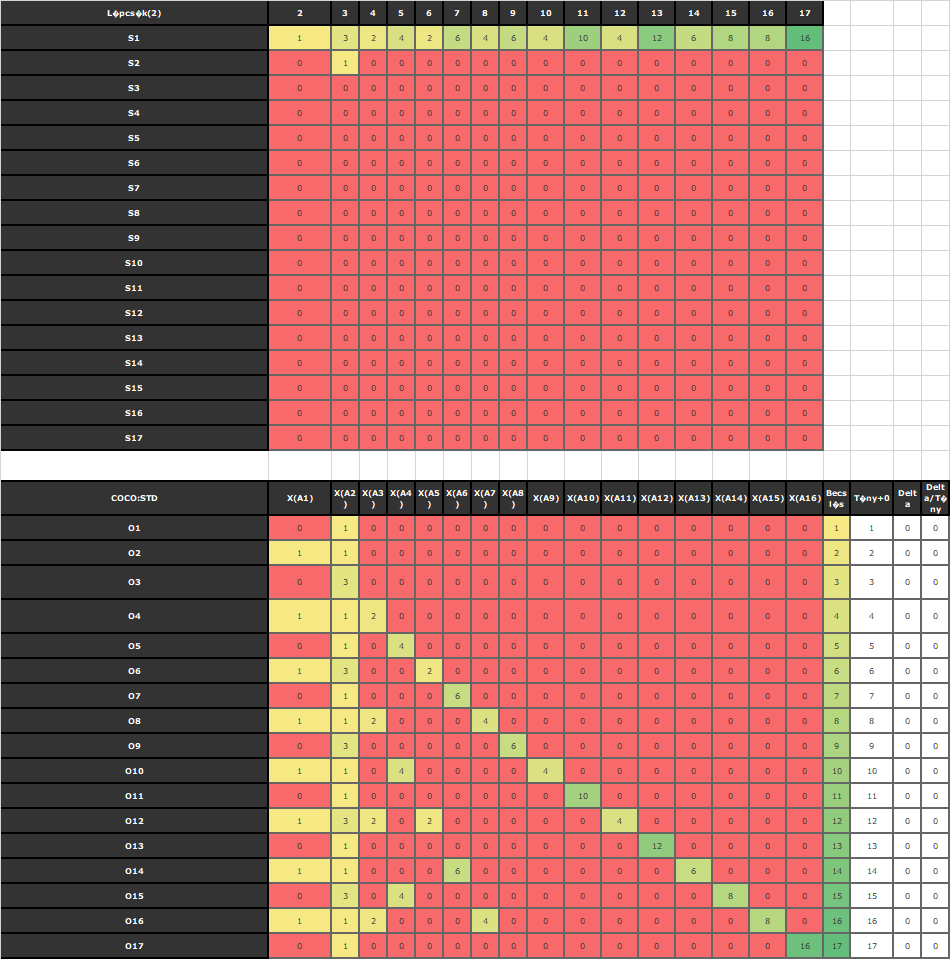
\*\*\*



1. Ábra: Oszthatósági teszt-OAM (forrás: saját ábrázolás)

A 19. ábra arra mutat példát elméleti alapozásként, hogy kiindulva az oszthatóság tényéből (1) ill. hiányából (2):

* ha Y-nak véletlen számokat választunk (jelen esetben 1…17 között)
* és az osztók maguk az Y-ok is egyben
* akkor egy speciális átló-domináns OAM-ot kapunk
* melynek azonban a hasonlóságelemzési megoldása nem kényszerűen átló-domináns (vö. 20. ábra):

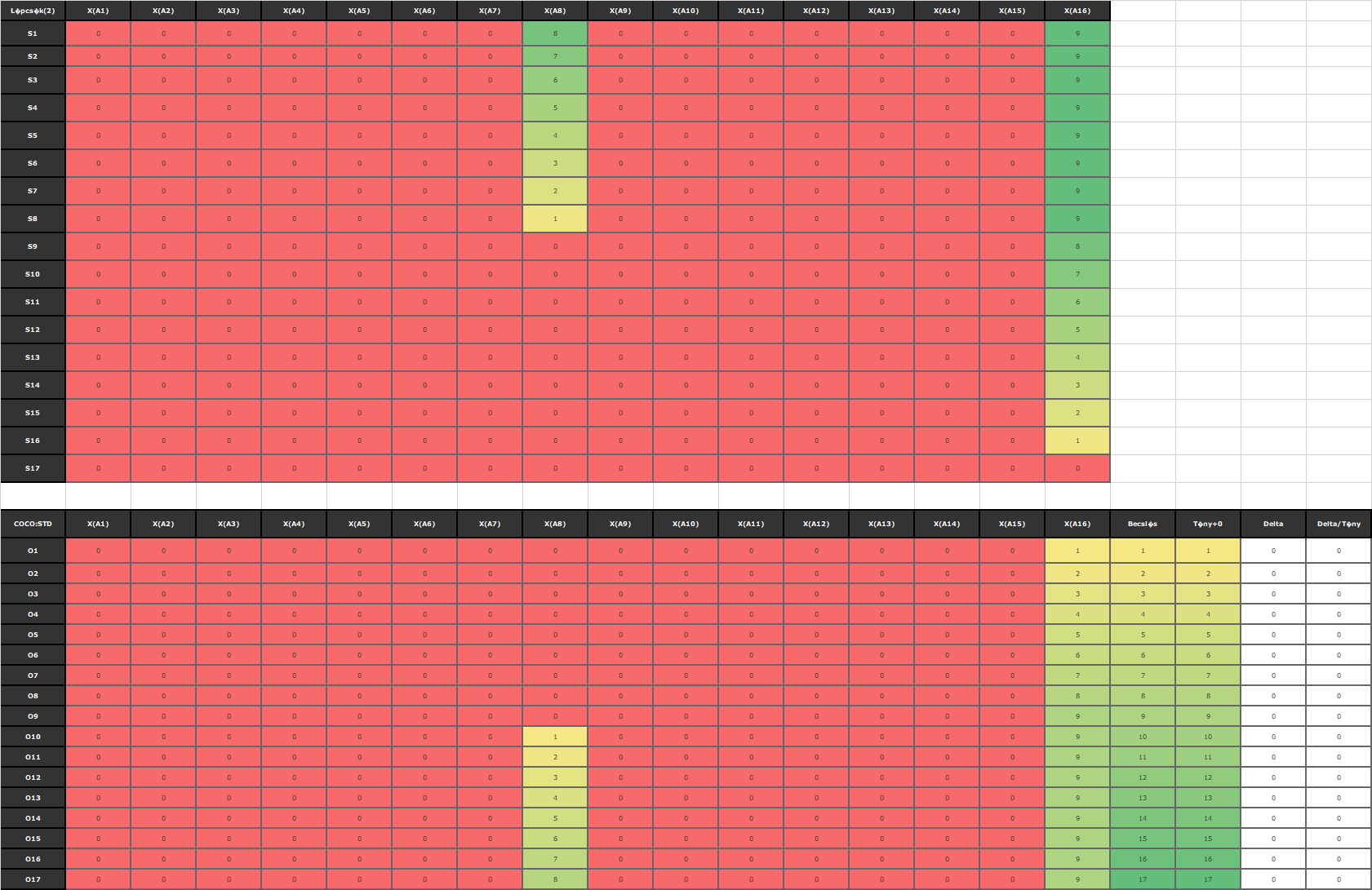


1. ábra: Átló-domináns OAM hasonlóságelemzési alternatív nézete (forrás: saját ábrázolás)

Az átló-domináns megoldás esetén S1 nem lenne más, mint a mindenkori Y értékek növekvő sora.

A 19. ábra jobb széle az osztás utáni maradékosztályokat használja fel OAM-ként, s ennek hasonlóságelemzési eredménye a 21. ábrán látható. A 9-cel és a 17-tel való osztás maradékai alapján a számok 1…17 között rekonstruálhatók.

A 21. ábra alapján jól látható, hogy a legnagyobb osztó esetén a maradékok quasi a keresett Y értékek, így a legnagyobb osztó információértéke közel azonos az 1-gyel, mint osztóval, mely értelemszerűen bele sem került emiatt az azonosság miatt az attribútum-készletbe. Az érdekesség azonban az, hogy nem csak a legnagyobb osztó kerül be az értelmező erőtérbe (vö. átló-domináns modellek alternatív megoldásai), hanem további osztó(k) is bevonódhatnak a megoldásba.



1. ábra: A maradékosztályokra alapozó teszt-modell (forrás: saját ábrázolás)

# Konklúziók

A kongruens modellezés OAM-jai kísértetiesen hasonlítanak tehát a lépcsős függvényekkel feldolgozható hasonlóságelemzések OAM-jaihoz (további sorszámozás nélkül is), hiszen a maradványosztályok tulajdonképpen sorszámként is értelmezhetők, csak nem szomszédsági alapon, hanem ahogy az a névben is benne van egybeesési (kongruens) alapon, ahol az egybeesés szabályai quasi bármik lehetnek – itt és most ezek a szabályok az oszthatóság és a maradvány-azonosság alapján képződnek.

A hatékonyságnövelés attribútum-szűkítő, de modell-pontosság őrző lépésekkel elérhető.

A konzisztencia-alapú modellezés érdekében a maradékosztályok és/vagy oszthatóság alapján készülő modellek következtetőgép-szimulátorként is értelmezhetők.

# Referenciák

…szövegközben…