Adatminőség-ellenőrzés univerzális lehetőségei hasonlóságelemzésre alapozva

(Universal possibilities of quality management based on similarity analyses)

Pitlik László, Pitlik László (jun), Tóth Roland, Ficza Helga, Pitlik Mátyás, Pitlik Marcell, MY-X team

Kivonat: A tanulmány üzenete többrétegű: egyrészt az ellenőrzések ellenőrzésének szükségszerűségére kell rámutatnia minden olyan gondolatsornak, mely kritikai/minőségbiztosítási elveket kíván megfogalmazni. Másrészt kijelenthető: a komplex, nem lineáris összefüggések létezhetőségének tesztelésére ismételten alkalmasnak mutatkozik a hasonlóságelemzés irányítatlan tanulást támogató megközelítése homogén/kombinatorikailag teljes mintázatok esetén. S végül a heterogén/kombinatorikailag lyukas input-mintázatok kapcsán: a szakértői rendszer elvű átvilágítás képes a kombinatorikai tér minden egyes opciójának következményváltozóra gyakorolt átlagos hatásmértékét feltárni, s az egyes jelenségek opcióin belül ezek rangsorát, távolságát érzékeltetni, mely speciális nézet lehetővé tesz egy fajta humán szakértői globális ellenőrzést, ill. az ellenőrzési szabályok knuth-i elvek szerinti kódolása után akár robotizált ellenőrzést is. A tanulmány speciális eredménye annak felismerése, hogy a teljes kombinatorikai terekben a rekordok sorszáma is lehet információ értékű, amennyiben a variációkat leíró bináris kódok quasi növekvő sorrendben, azaz szisztematikusan kerülnek megadásra, ugyanis a következményváltozó és a rekord-id között magas korrelációk léphetnek fel ilyen esetekben, ami további ellenőrzési lehetőségként aknázható ki.

Kulcsszavak: hasonlóságelemzés, irányítatlan tanulás, ki vigyáz az őrzőkre elv, id-alapú változók hatása

Abstract: The paper has a multilayered message set. First, it is always relevant to declare and to expect, that each control mechanism should also be controlled. Parallel, similarity analyses (without any restrictions) are capable of exploring the existence of quasi arbitrary patterns in form of expert systems, if the inputs present each variants of a full combinatorial pattern. In case of inconsistent input-patterns, it is possible to derive a kind of average impact for each variable and all they options. Therefore, it is also possible to rank the impact-volumes of each options within a variable. These information units can be evaluated both by human experts and through algorithms, if the knowledge engineers could transform the rules existing in the human brains into source codes. A special result is, that in case of a full combinatorial space derived as a kind of binary ranked series, the record-id can be used as an impactful variable having high correlation to the output variable. This special effect can also be involved into the control processes.

Keywords: similarity analysis, learning without restrictions, who-should-control-the-controllers-principle, impact of id-based attributes

# Bevezetés

**Az adatvagyonok minőségét, konzisztenciáját tartalom-függetlenül tesztelő eljárás** igényét banki adatok elemzésére való alkalmasság megítélésekor bárki átélheti, ahol a feladat nem más, mint teljesen sor- és oszlop-fejléc nélküli táblázatok alapján a legkevésbé racionális sorok szűrésére vonatkozó algoritmusok kialakítása úgy, hogy a táblázatot kiadó pénzintézeteknek van olyan tudása, mely alapján egyes rekordok anomáliának minősítendők. Ez természetesen nem jelenti azt, hogy nincs ezen halmazon túl is anomália, ill. nem csak egyféleképpen lehet anomáliáról beszélni (nem vitatva, hogy üzletileg, adott döntéshozó számára nem minden anomália releváns). Amennyiben tehát adott projektekben olyan adatvagyonok (big data) kerülnek átvételre, melyekben az objektumok és az attribútumok logikai kapcsolata sem ismert esetleg, **azaz quasi context free alapon kell egy adatvagyon konszolidáltságát megítélni, akkor is lehetséges a potenciális elgépelések, adattranszfer-hibák, tizedesjel hiányok, stb. feltárása**. (vö. <http://miau.gau.hu/miau/33/wam/sugb1.doc>, ill. <http://miau.gau.hu/miau/162/outliers.doc>, valamint <http://miau.gau.hu/miau/170/2.docx>).

Az adatvagyonok minőségének ellenőrzéséhez szükséges következő motivációit anno a kérdőívezés során vélelmezhető potenciális csalások lehetősége generálta. A kérdőívezés, különösen attitűd-feltáró, vélemény-letapogató kérdések esetén már eleve kétes értékű [vö. reprodukálhatatlan, ad hoc/intuitív, szómágikus, a válaszadó számára esetlegesen félreértett, a felkínált kérdésopciókkal (pl. nem-tudom-opció, semleges álláspont, stb. hiányában) nem lefedhető] válaszokat szállítanak, már amennyiben a kérdőívezés az ún. alvállalkozók/partnerek által valóban végrehajtásra került. **De mit tehet/tegyen a megrendelő, ha fennállhat annak veszélye, hogy számára a kérdőíves válaszok véletlenszám-generátorral lettek előállítva az alvállalkozók?** (vö. <http://miau.gau.hu/miau/158/la158.doc>)

Hasonló kérdésként értelmezendő az a jelenség is, amikor egy-egy publikációban egy fajta szakértői rendszer kerül leírásra, mely inputvariánsaihoz nem (triviálisan, feltétlenül, hozzáférhetően) reprodukálható módon következményértékek kerülnek hozzárendelésre. Ilyen esetekben felmerülhet annak igénye a közlemények befogadóiban, hogy egy fajta gyors ellenőrzést futtassanak le azt feltárandó, **vajon a következményértékek kellően egyenszilárdnak tűnnek-e az inputadatok viszonyrendszerében?** (vö. <http://miau.gau.hu/miau2009/index.php3?x=e0&string=framingham>, ill. <http://miau.gau.hu/miau2009/index.php3?x=e0&string=czabadai>).

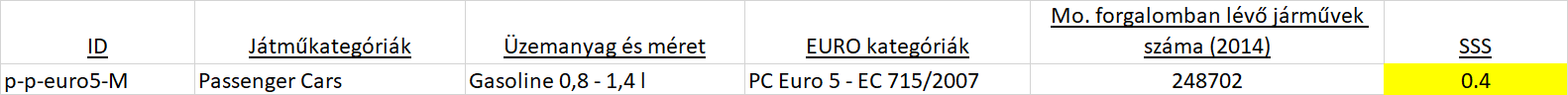
A közlekedésoptimalizálás kapcsán a környezeti adatok, ill. az alvállalkozók (vagy akár a saját alrendszerek) által mért adatvagyonok minőségbiztosítása automatikus kihívásként merül fel. Ez a tanulmány egy immár 11 elemből álló sorozat 12. része, ahol egy GINOP projekt keretében azon módszertani lehetőségek kerülnek kidolgozásra, melyek a mesterséges intelligencia-alapú forgalom-/környezeti terhelés-optimalizálás kapcsán újszerű megközelítésnek számítanak: vö.

1. <http://miau.gau.hu/miau/233/kvant_monitoring_v5.docx>   
2. <http://miau.gau.hu/miau/235/kvant_simulation_v1.docx>   
3. <http://miau.gau.hu/miau/235/kvant_geneticpotential_v1.docx>   
4. <http://miau.gau.hu/miau/238/Manuscript_Template_2017_myx.docx>   
5. <http://miau.gau.hu/miau/239/kvant_behavior_patterns_v1.docx>   
6. <http://miau.gau.hu/miau/240/lampa_ki_be.docx>   
7. <http://miau.gau.hu/miau/240/traffic_layers_1.docx>   
8. <http://miau.gau.hu/miau/240/lampaprogramok_versenye.docx>   
9. <http://miau.gau.hu/miau/241/traffic_szimulaciok_ertekelese.docx>  
10. <http://miau.gau.hu/miau/241/only_one_engine.docx>   
11. <http://miau.gau.hu/miau/242/onvezeto_parkolas_v1.docx>   
12. (<http://miau.gau.hu/miau/242/egyenszilard_adatvagyon_v1.docx>)

A tanulmány célja annak bemutatása, milyen operatív lépések mellett milyen gyanúgenerálási, kockázatkezelési feladatok kapcsán milyen eredmények várhatók el annak érdekében, hogy az adatvagyon-transzferek által okozható félreértések minimalizálhatók legyenek pontszerűen, alrendszerek, ill. a teljes adatvagyon szintjein.

# Esettanulmány: Számolótáblák által generált adatok „ellenőrzése”

Elvileg függvényszerű kapcsolatokból levezetett következményadatokhoz lehet jutni akkor, amikor pl. egy adott motortípus (vö. 1. ábra) esetén a NOX g/km károsanyag-kibocsátás kalkulált értékei kerülnek értelmezésre:



1. ábra: A motortípust leíró adatok (forrás: saját ábrázolás, ahol az SSS = start-stop-system)

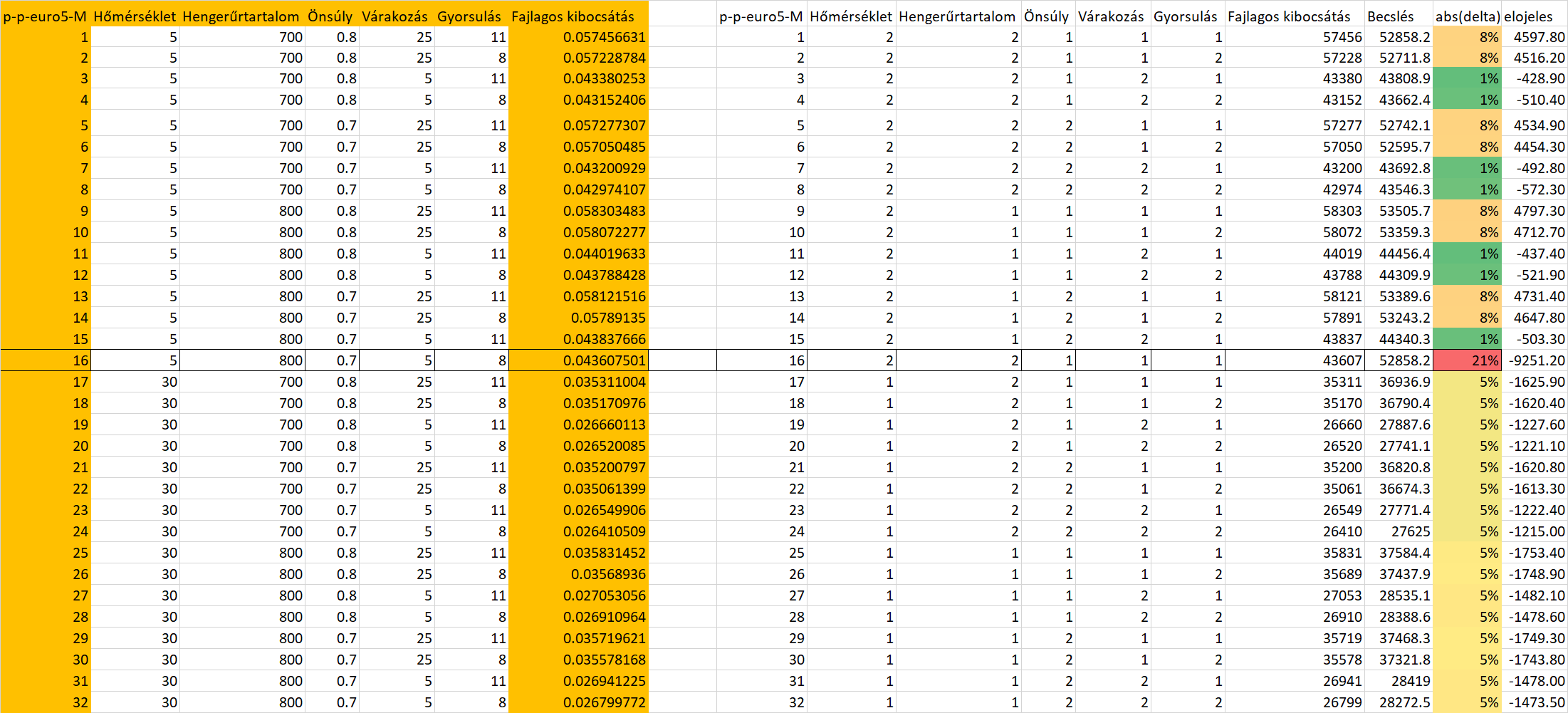
Feltételezve, hogy a károsanyag-kibocsátást befolyásoló tényezők függvényláncolata helyesen számol, vagyis az inputok és a becsült kibocsátásértékek között valóban függvényszerű a kapcsolat, s ez a függvényszerűség alapvetően értelmezni engedi pl. az egyenes és a fordított arányosság alapvetéseit, akkor quasi nincs az a bonyolult függvénylán (jelen esetben akár védett know-how formájában, vagy közelítőleg esetleg leírható alakzatként – vö. <https://hu.wikipedia.org/wiki/Taylor-sor>[[1]](#footnote-1)), mely logikusságát ne lehetne vizsgálni anélkül, hogy a szállított és az ellenbecslés keretében létrejött adatok között ne kelljen sikeres számításmenet és adattranszferálás esetén korlátlanul magas korrelációt megállapítani tudni megfelelő ellenőrző keretrendszer (vö. restrikciómentes és standard hasonlóságelemzéspárok) esetén (vö. <http://miau.gau.hu/myx-free/coco/index.html>).

A példa kedvéért a kibocsátás függjön (többek között) az alábbi tényezőktől (Xi):

* hőmérséklet (Celsius fok)
* hengerűrtartalom (köbcentiméter)
* önsúly (kilogramm)
* hideg várakozási idő (másodperc)
* gyorsulás ideje (másodperc)

A következményváltozó (Y) legyen a NOX g/km-re vetített károsanyagkibocsátás maga.

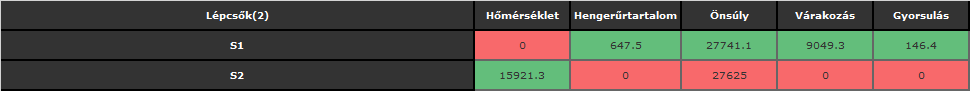
Anélkül, hogy a NOX számítása tételesen bemutatásra kerülne, az ellenőrző számára pl. a következő OAM (objektum-attribútum-mátrix) létezhet:



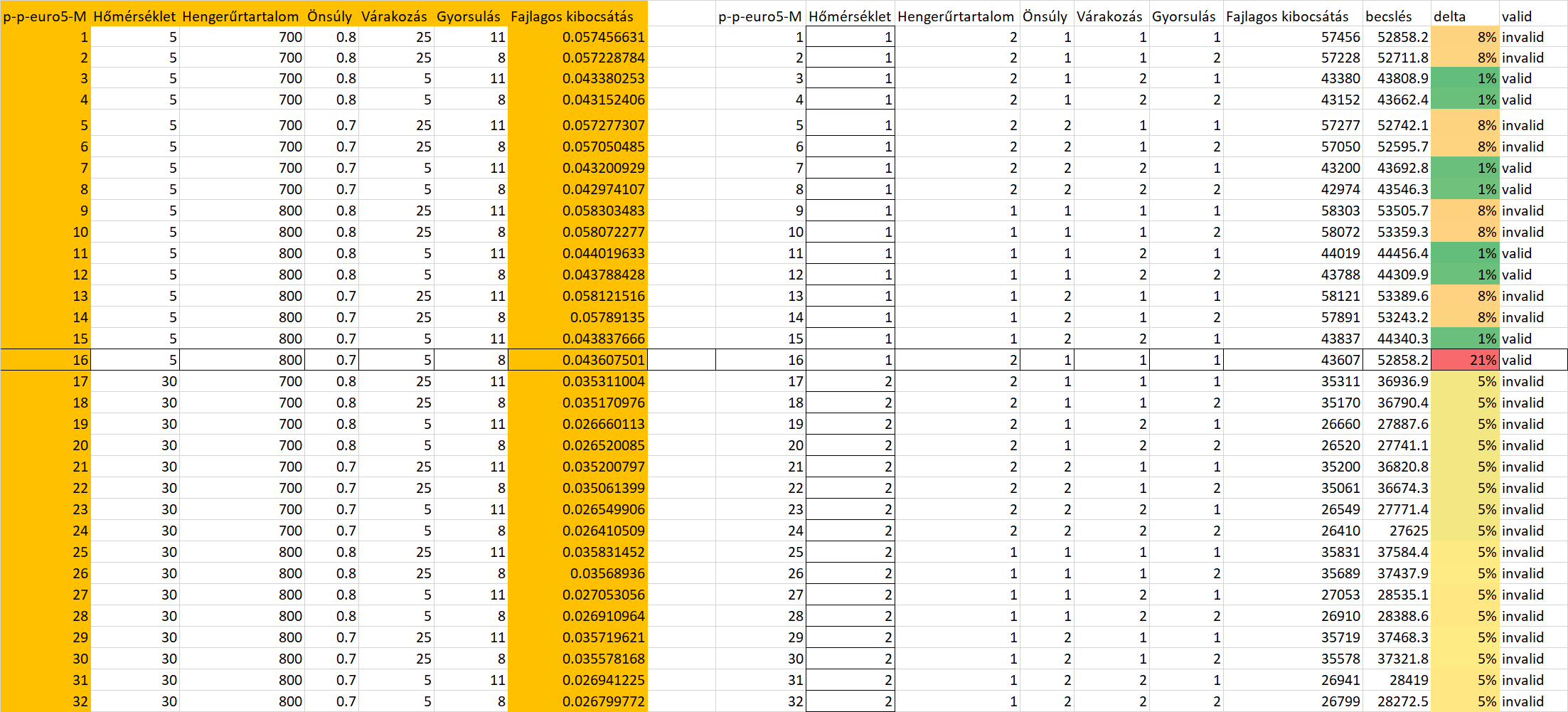
1. ábra: Irányítatlan online tanulási minta és eredményei (COCO-MCM számára – forrás: saját számítások)

A 2. ábra világos képet ad az 5 attribútumot és ezek 2-2 opcióit kezelő teljes kombinatorikai térről a számítás inputoldalán. A hasonlóságelemzések érdekében a numerikus input sorszámokká alakítandó attribútumonként. A restrikció nélküli, vagyis az irányítatlan tanulás lényege, hogy az attribútumok opcióira vonatokozóan semmilyen előírást nem kap a modellező keretrendszer (vö. COCO MCM). Így nem csak tetszőlegesen jó közelítést lehet elérni a becsült és a kapott NOX értékek között, hanem az összefüggésrendszer struktúráját is illetően is lehetséges impulzusokat kapni.

Jelen esetben a numerikus és strukturális közelítés egyaránt racionális eredményre vezetett:

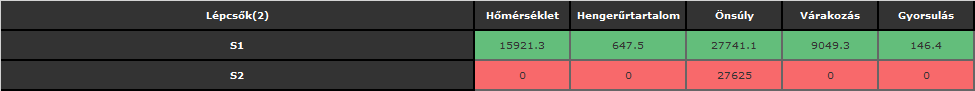


A korreláció a becslések és a tényleges NOX értékek között: 0.975, s a rendszer felismerte, hogy a hőmérséklet csökkenése növeli a kibocsátást: vö. y = -0,0167x + 1,1667, ahol x a hőmérséklet és y a kibocsátás.



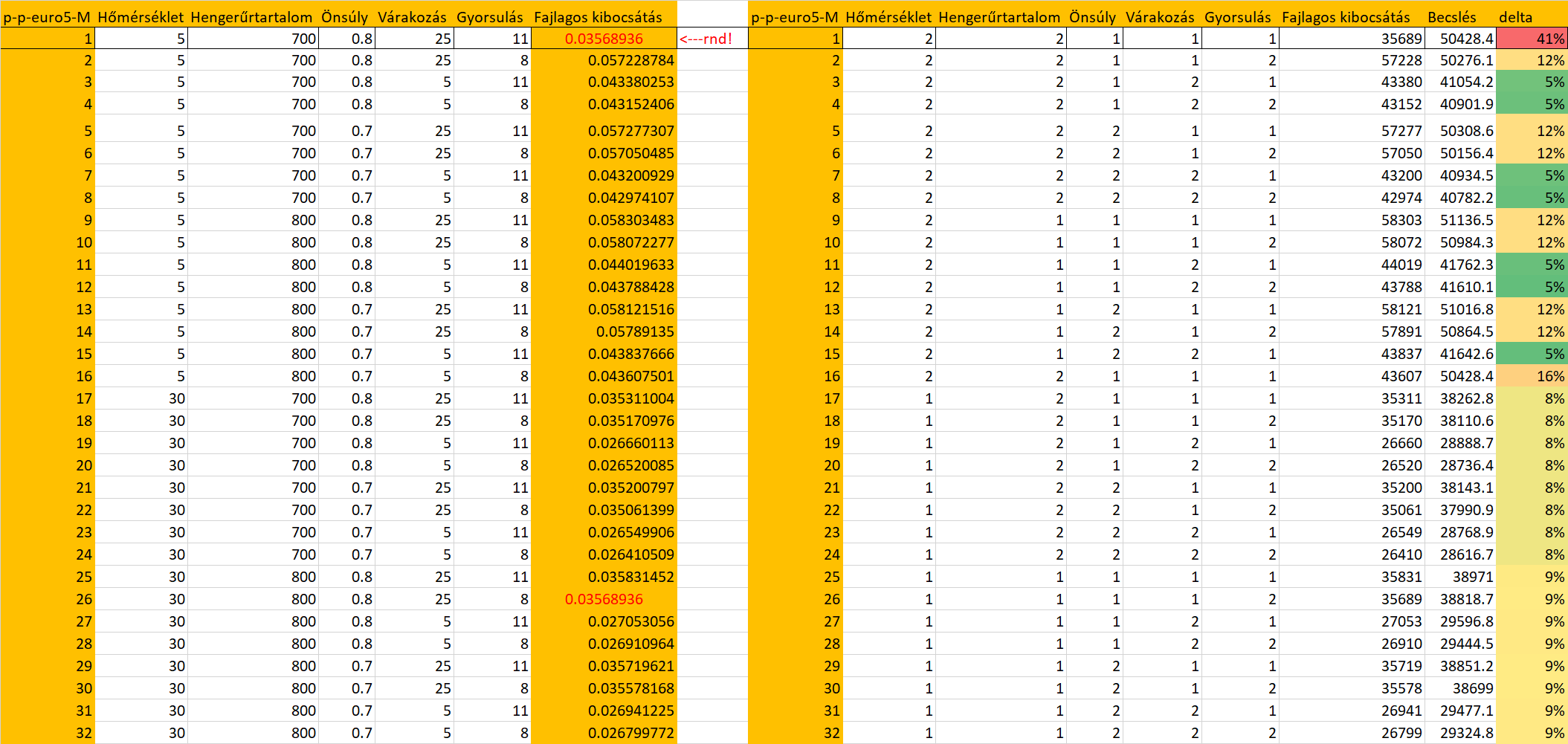
1. ábra: Irányított online tanulási minta (COCO-STD számára – forrás: saját számítások)

Miután egy irányítatlan tanulási modell feltételezni engedte, hogy a hőmérséklet fordítva hat, így ez a tudás átvezethető az inputoldal sorszámozására is, ahol ezekután mindenkor az 1. sorszám a nagyobb kibocsátást, míg a további sorszámok (vö. 2.) az egyre csökkenő kibocsátást szimbolizálják:



Az irányított tanulási eredmények teljes szinkronban állnak az irányítatlan tanulás eredményeivel, hiszen a hőmérsékleti sorszámok inputoldali megfordításának hatására a hatásmechanizmusok iránya is megfordult, de a mértékek nem változtak, így a becslések és a korreláció sem változott.

Mielőtt a 3. ábra 16. sorát megvizsgálnánk, mely arra mutat rá, hogy a kombinatorikai tér egyik rétege (variánsa) kevésbé következetesen értelmezhető, mint bármely másik 31 variáns, érdemes megnézni, miként reagált volna az összefüggések létét feltárni hivatott elemzés egy véletlenszerű elírásra az Y, azaz a kibocsátási oldalon, ahol a 4. ábra bemutatja, miként hat, ha pl. a 27. variáns értékét veszi fel az 1. variáns a maga helyes NOX értéke helyett:

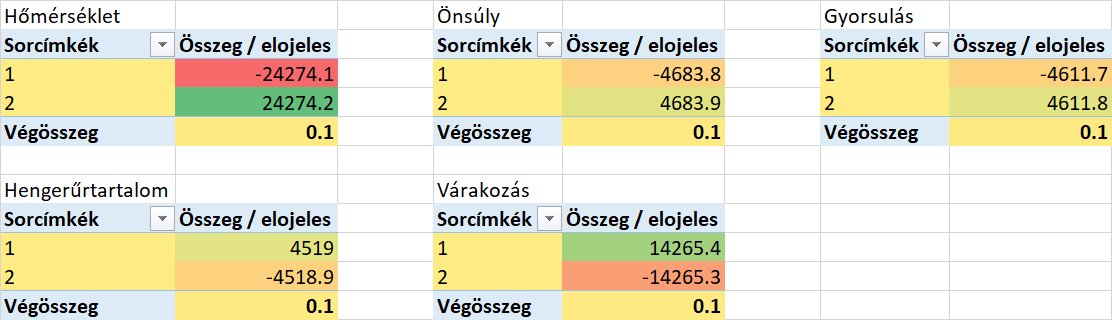


1. ábra: Hibás következmény-adat online feltárása (forrás: saját számítások)

A hibát a feltételes formázás keretében piros háttérrel kiemelve látható, hogy a hibás (véletlenszerű, elírt, téves) rekord értelmezhetősége a többi függvényében tényleg a leginkább irracionális tény:becslés alakzatot engedi vélelmezni (vö. 41%-os hiba az abszolút becslési eltérések esetén a tényhez képest). Továbbra is magas értéket kap azonban a 16. rekord.

Az ellenőrző mechanizmus direkt módon nem ad tanácsot arra, vonatkozóan, mi is lehet az oka a 16. rekord zavarának: lehet, hogy nincs is ennek oka, csak egy fajta számmisztikus zavarról van szó a numerikus rendszer 32 elemű kombinatorikai térként való butított ellenőrzése okán (vö. HD kép-minőség vs. pixeles kép). De az is lehet, hogy valódi probléma gyanújára sikerült bukkanni a mindösszesen magas korrelációjú rendszerértelmezés egy rétegeként.

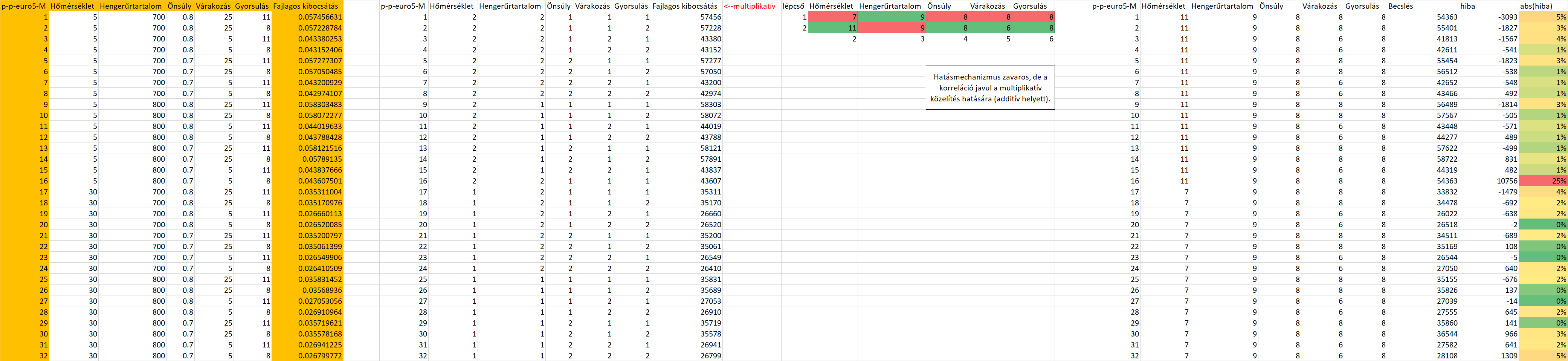
Az. 5. ábra arra mutat rá, miként alakul attribútumonkként az előjeles becslési hiba a véletlenszerűen feltárt, ill. irányítottan is azonos eredményre vezető rendszerben:



1. ábra: Hatás-dashboard (forrás: saját számítások)

Mint látható, a legnagyobb zavaró hatás a hőmérséklethez kötődik, vagyis a teljes kombinatorikai térben, ahol minden jelenség 2-2 opciója 16-16 alkalommal fordul elő, a legnagyobb rendszerinstabilitás a hőmérséklet szerinti 16-os rekord-csoportokhoz kötődik. A hőmérséklet előzőekben felismert előjelváltó hatása szintén erősíti azt a gyanút, hogy a 16. rekord, ahol az 1. ábra alapján csak 5 Celsius fok van, rel. dinamikus a gyorsulás (8mp), de a várakozás csak 5mp (ill. 800 kg, 700 ccm), ott a becslés nem feltétlenül racionális pontosságú.

Amennyiben a 16. rekord megértettségének kockázatát tovább akarjuk vizsgálni, akkor az eddigi additív, alapvetően kevésbé rugalmas, de éppen ezért rel. stabil context free közelítés helyett multiplikatív közelítést is lehet választani. A két közelítés viszonya olyan, mint a polinomok alacsonyabb és magasabb fokszámainak kapcsolata: a magasabb polinomizáltsági fokú egyenletek szinte tetszőleges alapadatok magyarázatára is képesek lehetnek, szemben a lineáris alakzatokkal. A multiplikatív hasonlóságelemzés itt és most Solver-alapon (azaz Excel-támogatás mellett) került kivitelezésre, míg a MY-X FREE COCO keretrendszer egy online megoldás. A 6. ábra egy multikplikatav megoldást mutat be az 1. ábra párjaként:



1. ábra: Multiplikatív offline közelítés (forrás: saját számítások)

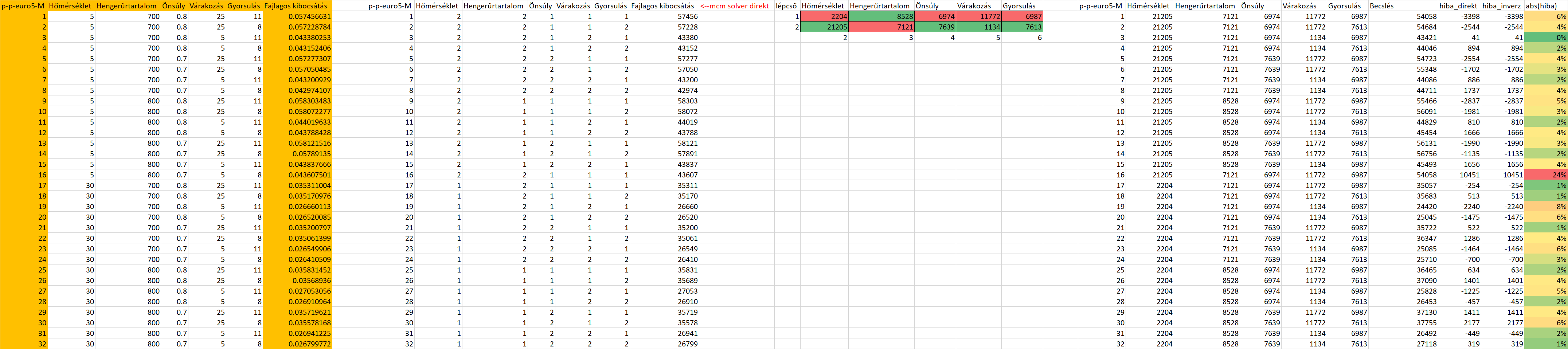
A 6. ábra jobb szélén ismét csak a 16. rekord piros vészjelét lehet beazonosítani, ami azt sejteti, hogy a 16. rekord valóban megérdemli a részletes értelmezést.

A 6. ábra középső zöld-piros színkódokkal ellátott kétsoros rétege azt mutatja be, hogy itt az egyenes és a fordított arányosságra vonatkozó alapvetések nem igazolódtak vissza az 5 attribútumból csak a baloldali kettő esetén.

A multiplikatív modell közelítése jobb, mint az additív modell közelítése, vagyis pl. a korreláció a tény és a becsült adatok között magasabb (0.982).

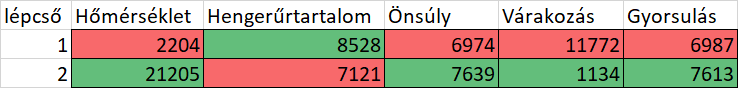
Mivel gyanút generálni nem illik alacsony komplexitású rendszerekkel, így a 16. rekord kérdése kapcsán további elemzési rétegek kialakítása is indokolt: egy ilyen réteg a még a 3. ábrán látható validitás-vizsgálat. A hasonlóságelemzések a függvényszimmetria alapján irányított modellek esetén képesek lehetnek inverz tanulási minta esetén inverz hibastruktúrát kialakítani a tények és a becslések között a direkt modellbecslések hibaelőjeleivel összevetve (vö. tagadás tagadása).

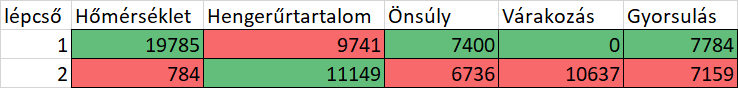
A 3. ábra alapján minden alacsony eltérésarányú (vö. 1%) becslés valid csak, ill. emellett a 21%-os hibaarányt felmutató 16. rekord, ami azt sejteti, hogy a rendszer maga alapvetően a 16. rekordra vonatkozóan nem ad egyértelmű gyanúra okot. S egy teljes kombinatorikai rendszerben valódi összefüggést sejtve a háttérben, s a 16. rekord NOX értékét számszakileg leellenőrizve valódi gyanúnak nem triviális, hogy legyen alapja. Ilyen kevert indikációjú esetekben a robot-gyanúgenerátor egyre inkább hajlik a rendszerszintű nem-tudom válasz felé, már ami a gyanú stabilitását illeti…



1. ábra: Irányítatlan offline modell (forrás: saját számítások)

A 7. ábra, ahol az online COCO-MCM helyett egy offline Solver-modell került kialakításra ismét csak a 16. rekord inkonzisztens leképezésénak gyanúját erősítit. A 7. ábra egyben rámutat arra is, hogy függvényszimmetriáról csak irányított esetben lehet beszélni, mert az irányítatlan modellek automatikusan felveszik a megfelelő lépcsős függvény-paramétereket, s azonos becslési értéket inverz lépcsőkkel képeznek le:

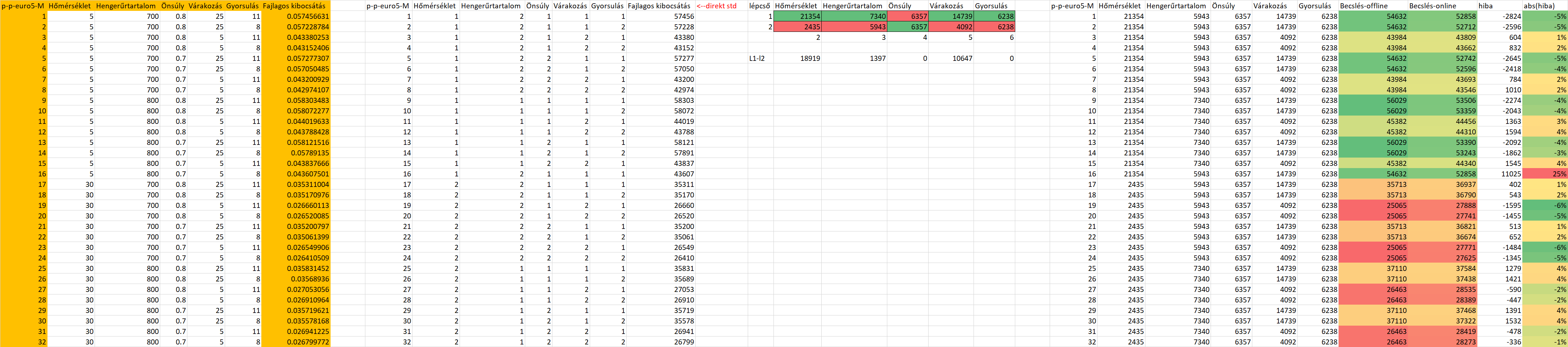




1. ábra: Alternatív irányítatlan modellek kikényszerítése (forrás: saját számítások)

A 8. ábra a színkódok szintjén az elvárt inverz jelleget visszatükrözi, de a numerikus értékek kapcsán már alternatív megoldást sejtet, ami a hasonlóságelemzések sok-paraméterhelyes lépcsősfüggvényeinek sajátja.

Az alternativitás nem volt tetten érhető az irányítatlan és a megfelelőképpen irányított modellek kapcsán, de megfigyelhető a hasonlóságelemzés futtató motorjainak és az ezeket kiszolgáltó paramétereknek a váltása kapcsán (vö. 9. ábra):

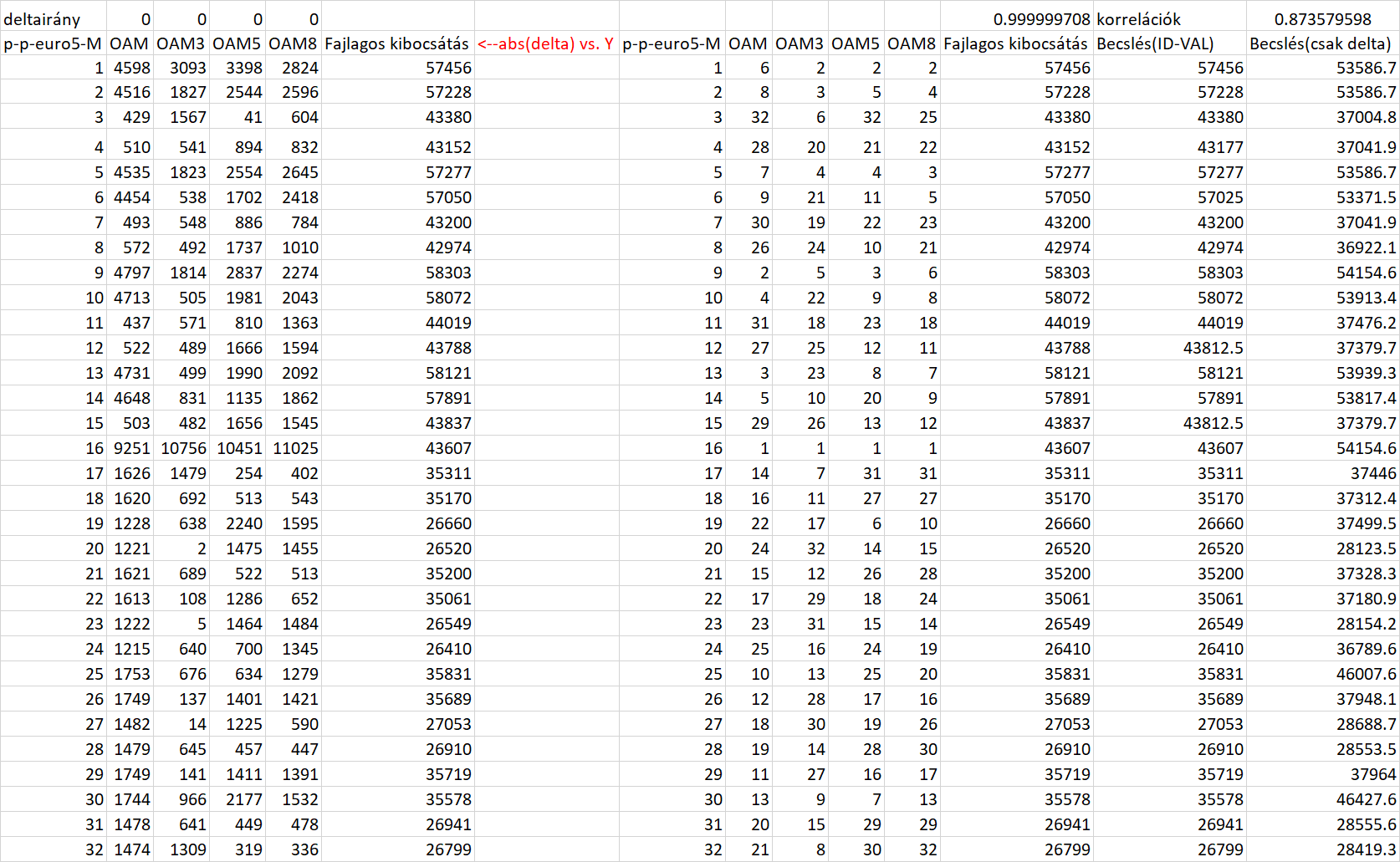


1. ábra: Offline additív modell (forrás: saját számítások)

A 9. ábra jobb szélén az online és az offline becslések egymás mellett kerültek színkódokkal összevetésre: a mintázat nagyon hasonló, de nem azonos! S különösképpen kiemelendő, hogy a 9. ábra közepén látható lépcsős függvény egyenes és fordított arányosságot mutató zöld-piros cellái egy újabb hatásmechanizmus esélyét vetik fel: öt változóból a 2., 4. és az 5. esetén helyes viszonyt leírva. A korreláció a Solver-alapú, legkisebb hibanégyzetet minimalizáló megoldásban 0.976, ami nagyobb, mint az online megoldás esetén, ahol az abszolút hiba minimuma kerestetik egy online LP-motor bevonásával. A 9. ábra is tovább erősíti a 16. rekord körüli gyanúmomentumot.

Az alternatív megoldások rendelkezésre állása lehetővé teszi az alternatív megoldások alapján való további becslés-pontosságnövelést:

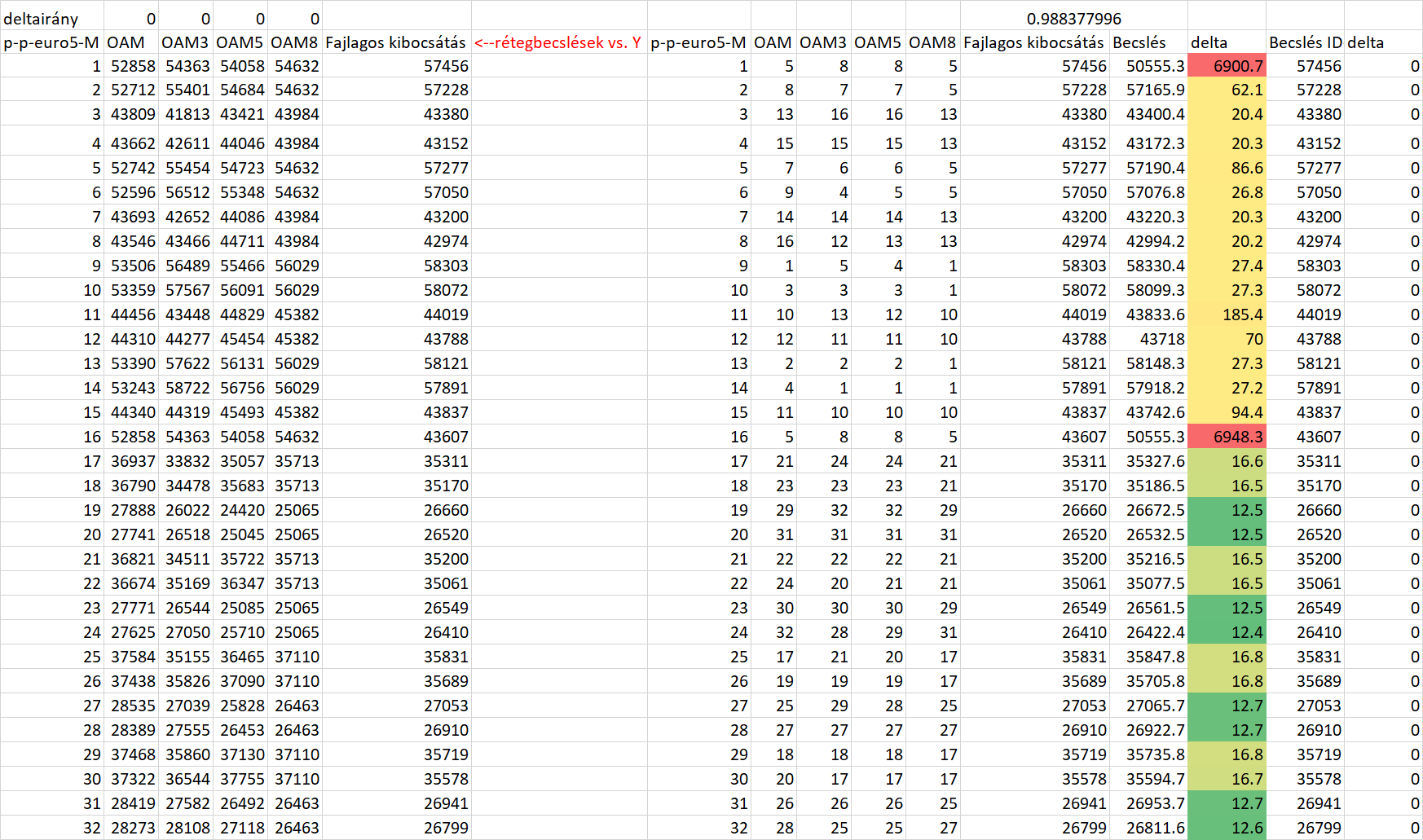
A 10. ábra egy tudatosan több szinten is speciális kísérletet mutat be. Az OAM-ba 4 alternatív megoldás került: vö. OAM az 1. ábra alapján, OAM3 a 6. ábra alapján, OAM5 a 7. ábra és az OAM8 a 9. ábra alapján. A hibák minél nagyobb abszolútértéket mutatnak, annál nagyobb NOX értékhez illik, hogy tartozzak – legalább is a hibák irányítása erre utal, ami egy nagyon komplex hipotézis, mely vélhetően ritkán igaz. S ezt meg is erősíti a 0.873-as korreláció, ami nem jelenti az elvárt előrelépést a 0.97-0.98-as szintekhez képest.



1. ábra: Hiba-alapú finomhangolás (forrás: saját számítások)

A 10. ábra másik specialitása a kombinatorikai tér quasi véletlenszerű rekordazonosítóinak inputként való bevonása, melyek eredményeként a becslések és a tények között a korreláció 0.999 értékre ugrik fel. Ez a kombinatorikai tér zártságát, s mint ilyen a 16. rekord körüli gyanúk elvetését támogatja.

A 11. ábra az alternatív modellek becsléseiből és a tényadatokból levezett abszolút hibák helyett magukat az alternatív becsléseket vonják be az OAM-képzésbe (ID-nélkül):



1. ábra: Hibridmodell az alternatív megoldások alapján (forrás: saját számítások)

Az alternatív becslések alapján kialakított másodlagos becslés egy olyan hibridmodell a hibák alapján megkísérelt hibirdizáció után, mely azonnal képes a korrelációt növelni: 0.988-ra szemben az eddigi legjobb modellkorrelációval (0.982). A korrelációk változásakor nem az abszolút eltérések nagysága a fontos, csak és kizárólag a növekedés tettenérése számít. Az így nyert modell immár az 1. és a 16. rekord esetén hasonló mértékű kockázatokra mutat rá, ami a 16. rekord kockázatainak új értelmezés szerinti kimutatását jelenti, vagyis a 16. rekord egy olyan rekord, mely véletlenszerűen egy meglévő, jelen esetben az 1. számú rekorddal mutat fel logikai ütközést: vagyis a rendelkezésre álló alternatív becslések alapján azonos a várható következményérték, noha a tényleges NOX értékek jelentősen eltérnek egymástól.

A 11. ábra mögötti modellaktivitás a 10. ábrához hasonlóan egy párhuzamos modellbe bevonta a kombinatorika tér variánsainak sorszámát, ami ismét csak hibátlan modellhez vezetett. A kombinatorikai tér azonosítását szolgáló (formálisan nem numerikus azonosítók információértékének ilyen speciális alakzatára már nem egy 2-2 opciós, 5 attribútumos, azaz 2^5=32 elemű térben kialakított 2-lépcsős függvény esetén volt tetten érhető, hanem a klasszikus 32lépcsős rendszerben úgy, hogy az ID a teljes összefüggésrendszer újra formálását kényszerítette ki (vö. 12. ábra):



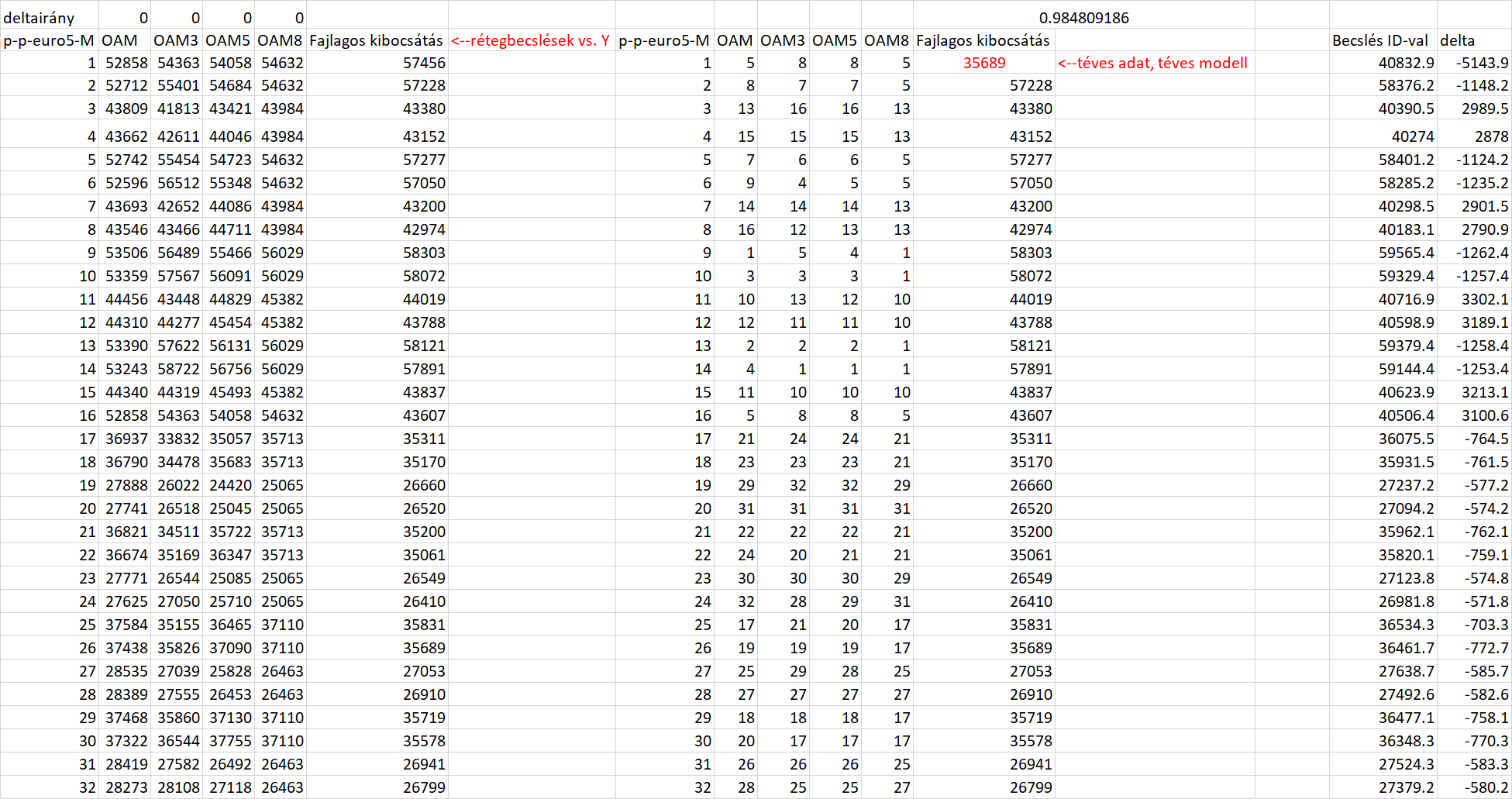


1. ábra: A teljes kombinatorikai tér leírásának hatása (forrás: saját számítások, ahol felül az alternatív becslések, alul ezek hibái kerültek inputként az OAM-ba)

A 11. ábra kapcsán az 1. és a 16. rekordok antagonizmusa kapcsán akkor is fel kell figyelnie immár az elemzés által érintetteknek, ha eddig a 16. rekord körül kialakult ellentmondásos, de mégis csak állandósuló probléma kezelése nem lett volna önmagában eredményes. Az ellenőrző algoritmusok alapvetően szimmetria-vezéreltek, így egyetlen egy rekord hibássága mindenkor rendszer szinten gyanús. Amint felismerésre kerülnek antagonizmusok, akkor azonnal vizsgálni kell, vajon az alapmodell nem volt-e már antagonizmusok által érintett (vö. 2. ábra)? Ugyanis semmi garancia nincs arra, hogy egy ellenőrzés nem lehet hibátlan. Így magának az ellenőrző mechanizmusnak az ellenőrzése is szükséges minden esetben (vö. ki vigyáz az őrzőkre). Amint az a 2. ábra alapján világosan megállapítható: az 1. és a 16. rekordok input-rétege azonos (noha az X2 értékének változnia kellett volna), s következésképpen az output értékek is azonosak. Vagyis a tanulmány eddigi folyamata arra akarta felhívni a figyelmet, hogy az ellenőrzés önmaga kontrolljára is ki kell, hogy tudjon terjedni, s ennek egyik eszköze a függvényszimmetria, ill. ennek hiánya.

Mielőtt azonban a korrigált inputokkal a teljes eddigi elemzési útvonal újra bejárásra kerülne még érdemes a függvényszimmetria erőterét aktiváló lépéseket lefolytatni annak érdekében, hogy az ID speciális szerepét jobban meg lehessen érteni (s mely megértést a korrigált inputokkal folytatandó kísérletek természetesen tovább fognak lendíteni):

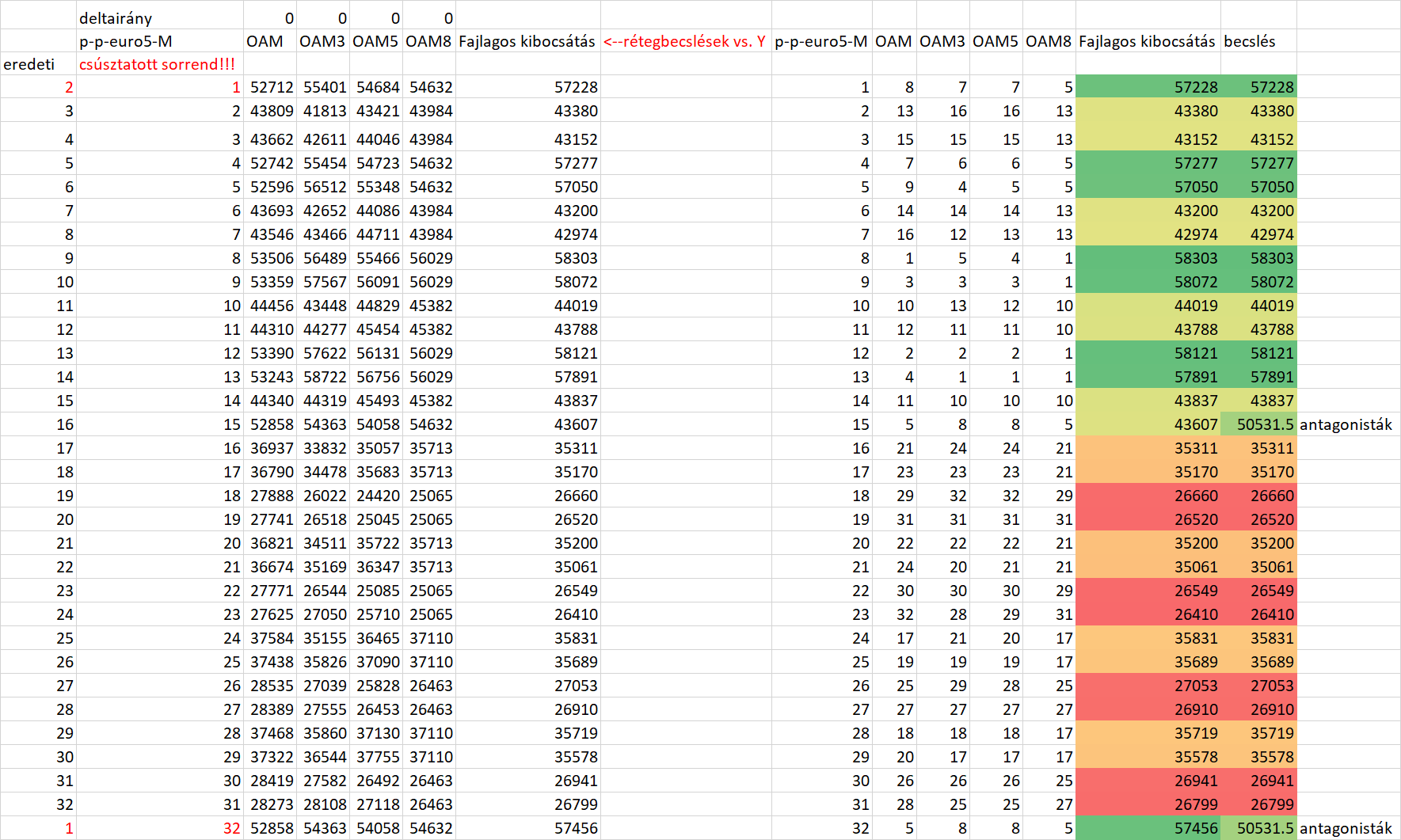
Az ID szerepének vizsgálatához érdemes egy téves output megadásának hatását is feltárni (vö. 13. ábra)



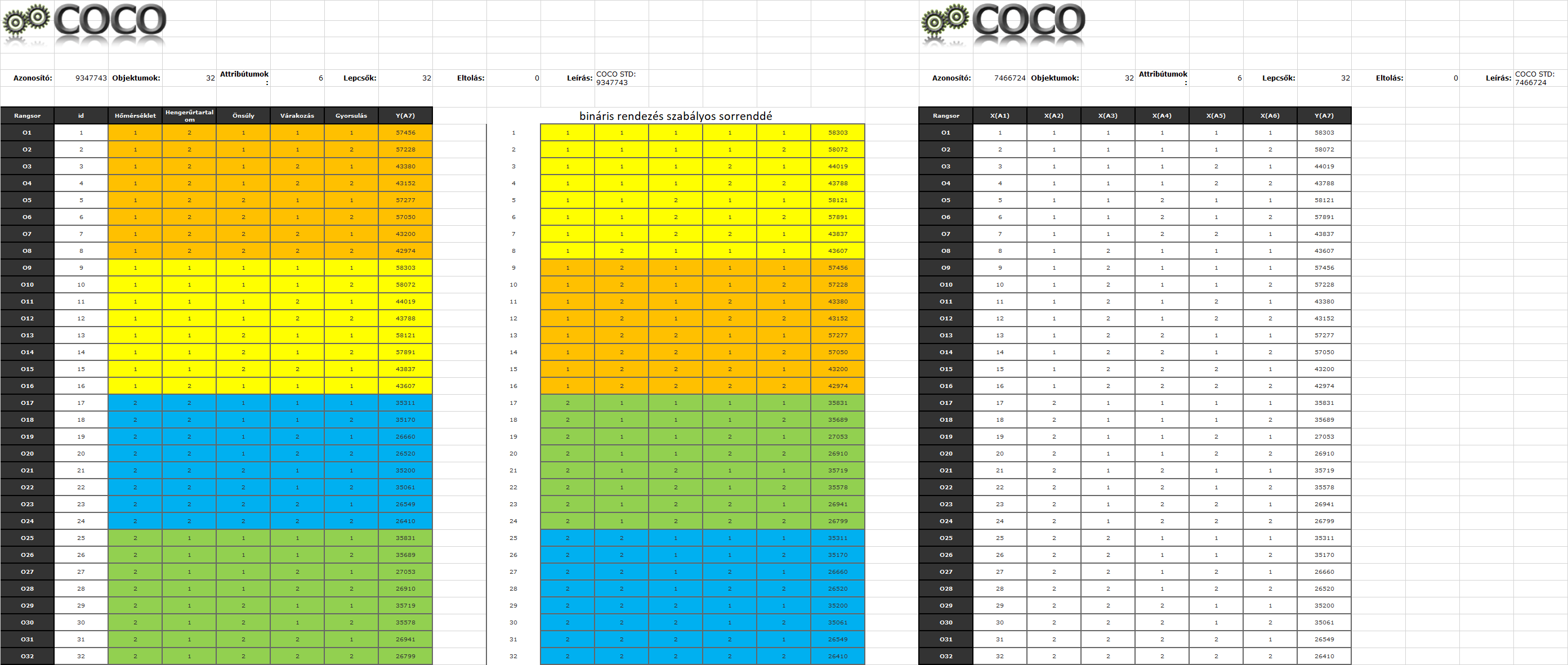
1. ábra: Téves output hatása az ID-t is feldolgozó modellben (forrás: saját számítások)

Mint a 13. ábrán látható: a téves OAM-cella (téves NOX) érték azonnal felborítja a specális ID-hatást, tehát az ID-nak tartalmilag is van információértéke.

A 14. ábra az ID-hatás ellenőrző letapogatásának eredményét mutatja be, ahol (a rekordsorrend-változást lekövető ID-változás eredményeként korábban tökéletes becslések álltak elő, de) a valódi ID-eltolás kapcsán már az ID-alapú modell tökéletessége repedezni kezd, s nem mellesleg ismét az eredetileg 1. és 16. rekordok antagonizmusa kerül a felszínre:



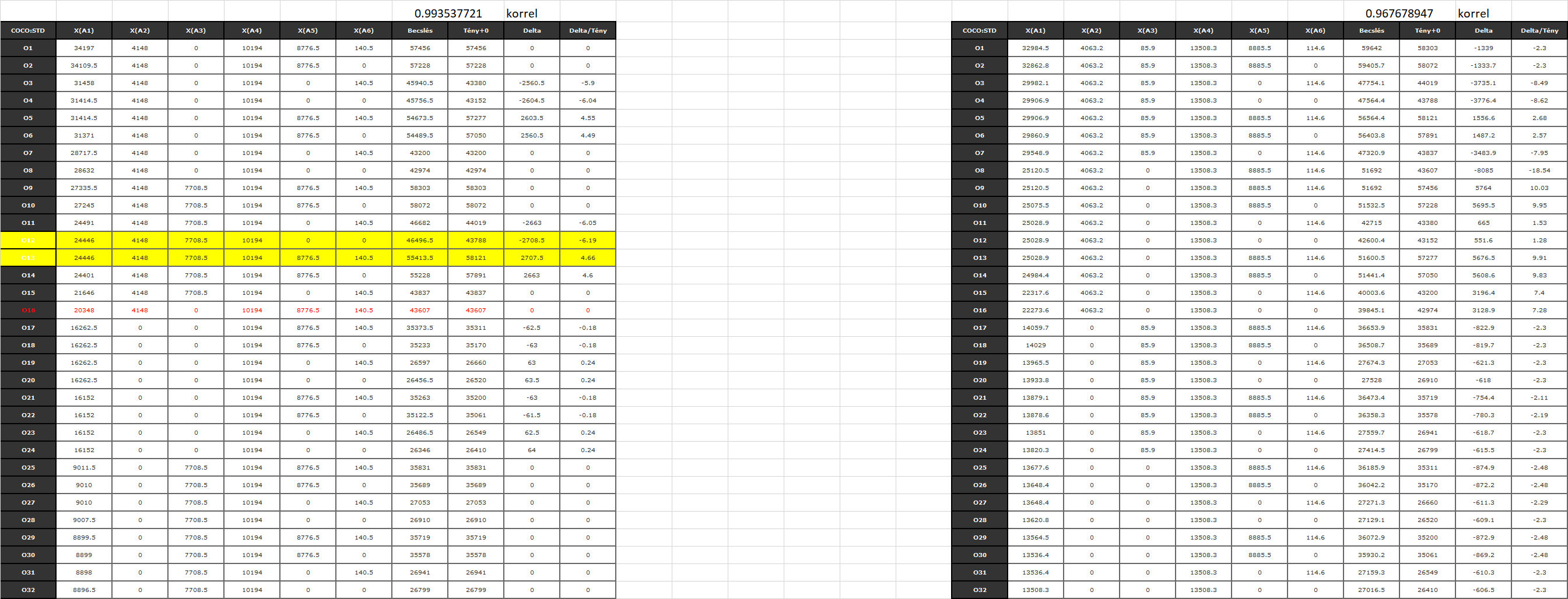
1. ábra: Az ID-változás által előhívott antagonizmusok (forrás: saját számítások)



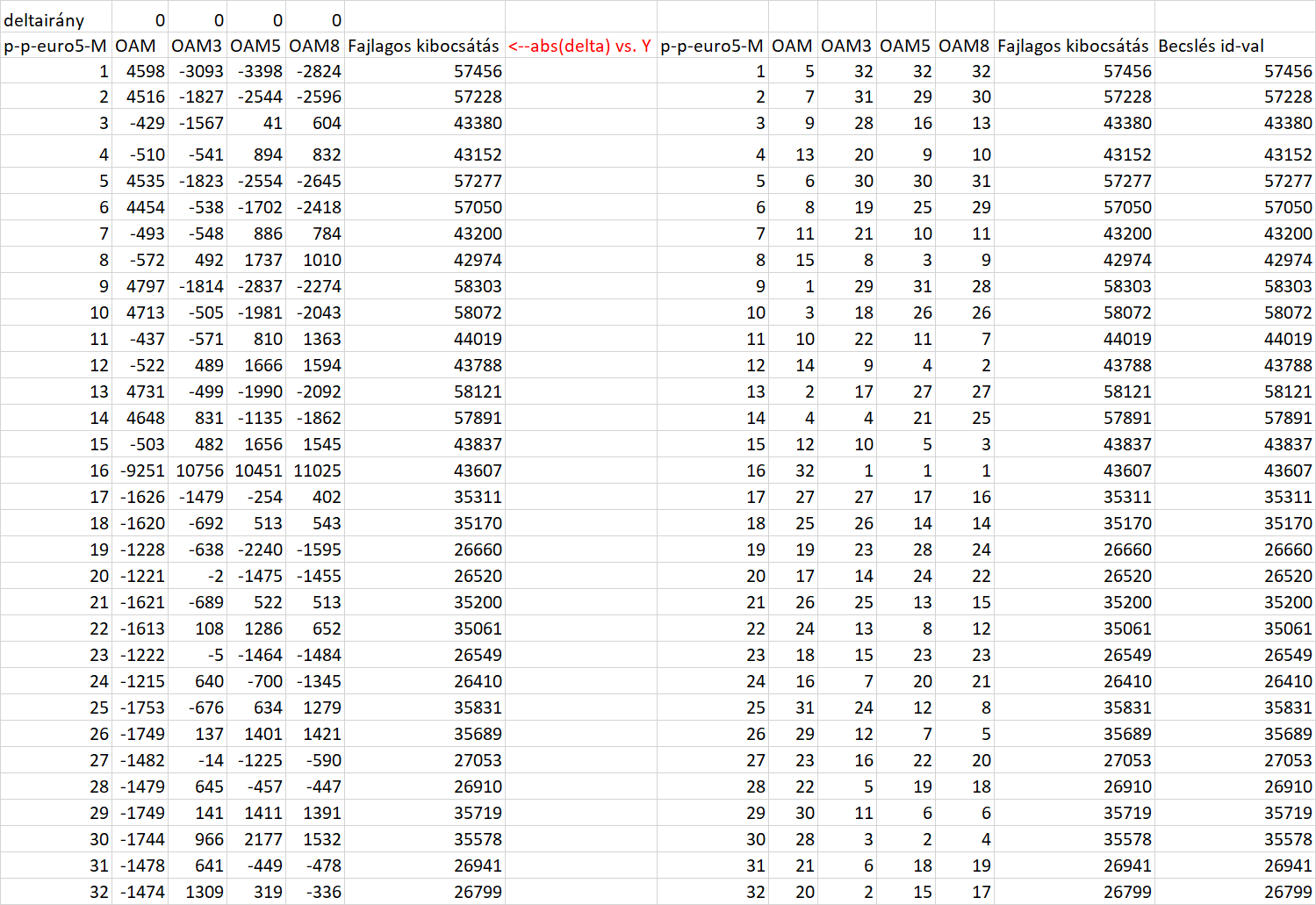
1. ábra: A bináris inputleképezés sorrendjének változása (forrás: saját számítások)

A 15. ábra annak a hipotézisnek az értelmezését támogatja, miszerint az inputok speciális (quasi növekvő) sorrendje az oka az ID-hatásnak. A korrelációs értékek szerint (vö. 16. ábra) a bináris rendezettség nem hat pozitívan (vö. 0.99>0.96).

A 15-16. ábra irányított tanulás keretében azt vizsgálja, vajon a tökéletes ID-hatás, melyet 32-lépcsős függvények tártak fel eddig, segíti-e a csak 2-2 inputos alapadatok és a 32 inputos ID-változó esetén is a jobb becslések kialakulását. A korrelációs érték 0.99-es szintje a bináris rendezettség növelése előtt = alapállapotban magasabb, mint a korábbi, ID-nélküli értékek (vö. 0.975).



1. ábra: A bináris input-mintázat hatásai (forrás: saját számítások)



1. ábra: Előjeles tanulási minta hatása (forrás: saját számítások)

A 17. ábra az abszolút hibákkal és az alternatív becslésekkel dolgozó modellek által feltárt tökéletes ID-hatás mellett immár az előjeles hibák esetét is megvizsgálni engedi: az elvárásoknak megfelelő az előjeles hibákkal dolgozó OAM is tökéletes ID-hatáshoz vezet. Az előjeles és az alternatív becslésekkel dolgozó OAM-ok egymás függvényei, így a tökéletes hatás ezek esetében triviális. Az abszolút hiba azonban már egy olyan OAM-torzulás, mely esetén trivialitásról nem eshet szó.

Itt és most egy regresszió-alapú alternatív értelmezés illene, hogy következzen azID-hatás mibenlétének feltárása érdekében, vagyis a hasonlóságelemzési hatásmechanizmus kikapcsolását lehetővé téve. Ezt azonban érdemes, hogy megelőzze immár a tökéletes bináris inputra alapozó elemzési rétegek eredményeinek feltárása:

# Esettanulmány: Számolótáblák által generált adatok ellenőrzése

Az eddigi „ellenőrzést”, mely célja az volt, hogy az ellenőrzés ellenőrzésének fontosságát erősítse, következzék immár egy input-oldalon hibátlanná alakított OAM-ra alapozó vizsgálatsor:

Az előző ábrák, különös tekintettel a 2. ábra eleve kiemelt 16. sora a gyakorlott szakértői rendszer-szakértők, vagyis a kombinatorikai terek ismerői számára azonnal jelezték, hogy az input-oldali kombinatorikai tér zavaros, hiszen nem minden oszlopban van jelen 16:16 egyes és kettes jelzés.

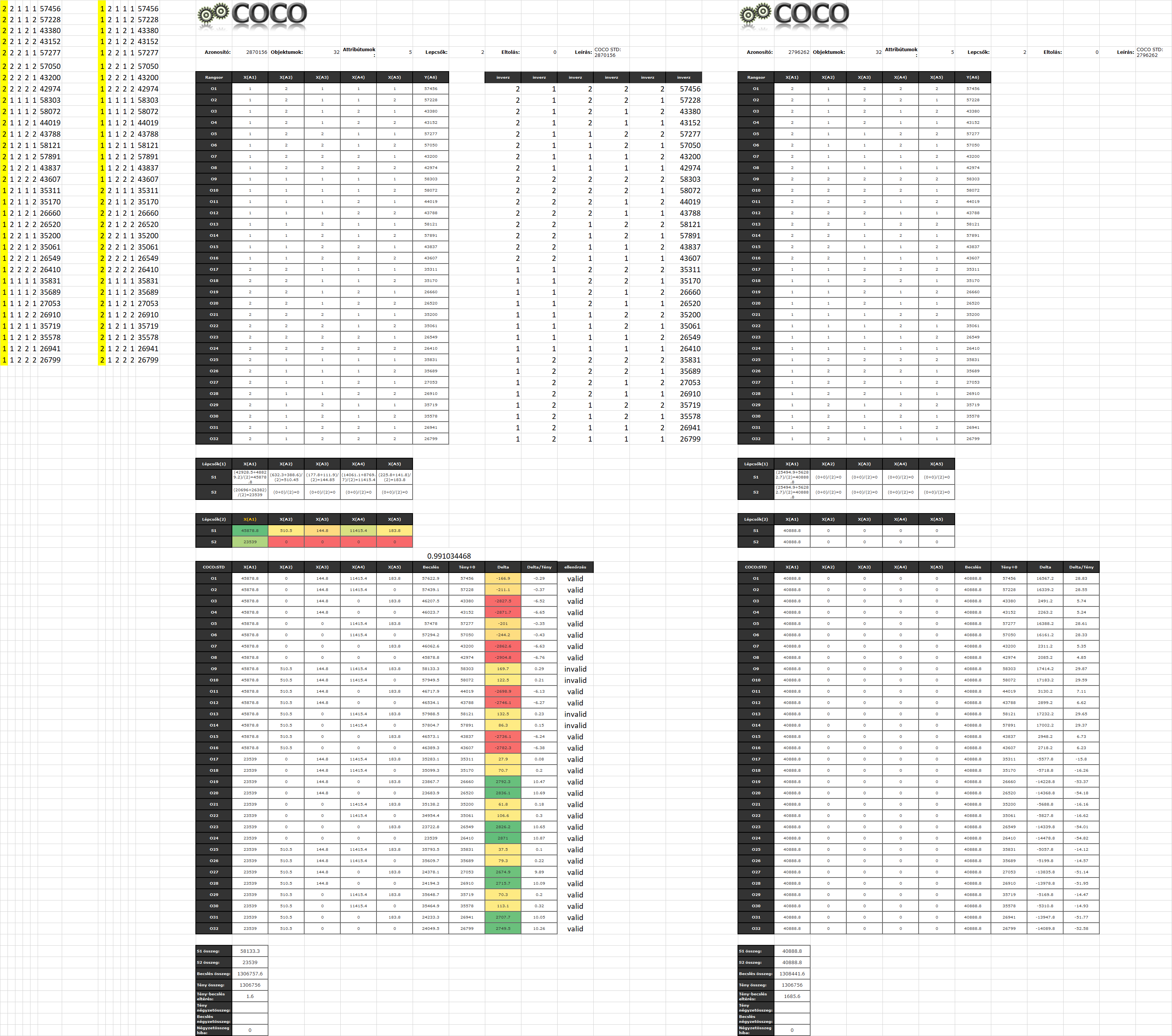
A 18. ábra immár a korrigált input-oldalt mutatja meg az egyelőre változatlan output-oldallal együtt:



1. ábra: A korrigált input-oldal hatása a modellekre (forrás: saját számítások)

A 18. ábra alapján elmondható, hogy

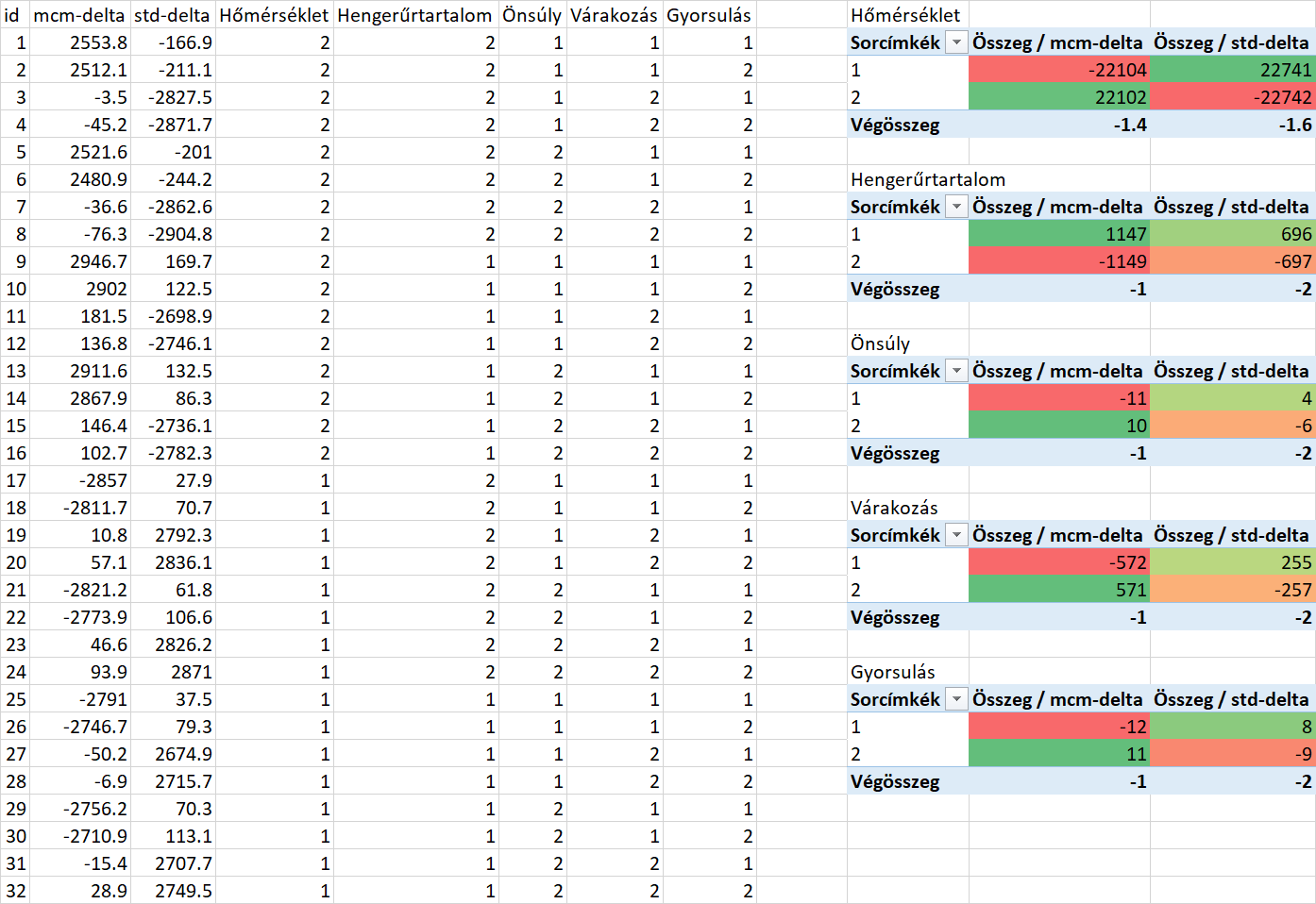
* az irányítatlan modellezés (MCM) korrelációs értéke 0.99-nél is nagyobb, vagyis a hibás inputoldalt jellemző 0.975-ös értéket meghaladó, ahol messzemenő meghaladásról ilyen telítettség esetén nem lehet soha beszélni, ellenben a magasabb korreláció világos utalást ad arra nézve, hogy az input-output viszonyrendszer értelmezhetősége erőteljesebb,
* az irányított modellezés (STD) korrelációs értéke 0.5, s a lépcsők az első változó, vagyis a hőmérséklet esetében azonos értékűek (középső szekció), ami egyértelműen jelzi, hogy oda fordított arányosságnak megfelelő értékek kellenek, amit az MCM modell vissza is tükröz (jobb oldal),
* míg a becslések és a tények eltérése az irányított tanulás esetén egyértelműen a hőmérséklet mintázatát követi (vö. zöld-piros színkódok), addig az irányítatlan tanulás esetén a mintázat a várakozás mintázatához és a hőmérséklet inputoldali leosztásához egyszerre hasonlít. Vagyis a komplex, nem lineáris összefüggésrendszer eredményeként előálló Y bináris mintázat formájában való leképezése egyre jobban közelít a mintázattalanság elvárásához,
* a hőmérséklet fordított arányosságának MCM-alapú leképzése mellé a többi változó egyenesarányosságának elismerése társul itt is, mint korábban, ami nem meglepő, hiszen csak egyetlen egy inputsor volt zavaros,
* a helyes input-mintázat esetén nincs kiugróan hibás rekord (vö. 16-os szindróma),
* a validitás mértéke (vö. 3. ábra megduplázódott: 8-ról 16-ra emelkedett), ahol a kezdeti magas invaliditás oka a hibás input-oldal és a fordított arányosság tudatos kizárása volt, míg a helyes inputoldal kialakítása után már csak a kényszeres egyenes arányosság okozza a validitási zavarokat…



1. ábra: Az irányított tanulás invertált inputokkal (forrás: saját ábrázolás)

A 19. ábra az irányított modellezés keretében az első input-változó (hőmérséklet) esetében az invertált nézetet képes kezelni, így bináris inputok esetén az irányítatlan tanulás tükörképét képes előállítani. Emellett:

* a validitás jelentősen megnőtt, már csak 4 invalid eset tárható fel ott, ahol az inverz modell lényegében a teljes tanulási (következmény-) értelmezési tér átlagát tudta csak felvállalni racionális becslésként a tükrözött inputok esetére egy teljes kombinatorikai térben,
* a korreláció hasonlóan magas, mint az irányítatlan tanulás esetén (0.99),
* a kifordított hőmérséklet-minta értelmében az egyenes és fordított arányosságok helyesek,
* a becslések és a tények eltérésének struktúrája hasonló, de ebben az esetben a felső 16 rekord esetén vannak piros és alul a zöld eltérések,
* a lépcsőértékek nem azonosak a 18. ábra lépcsőivel

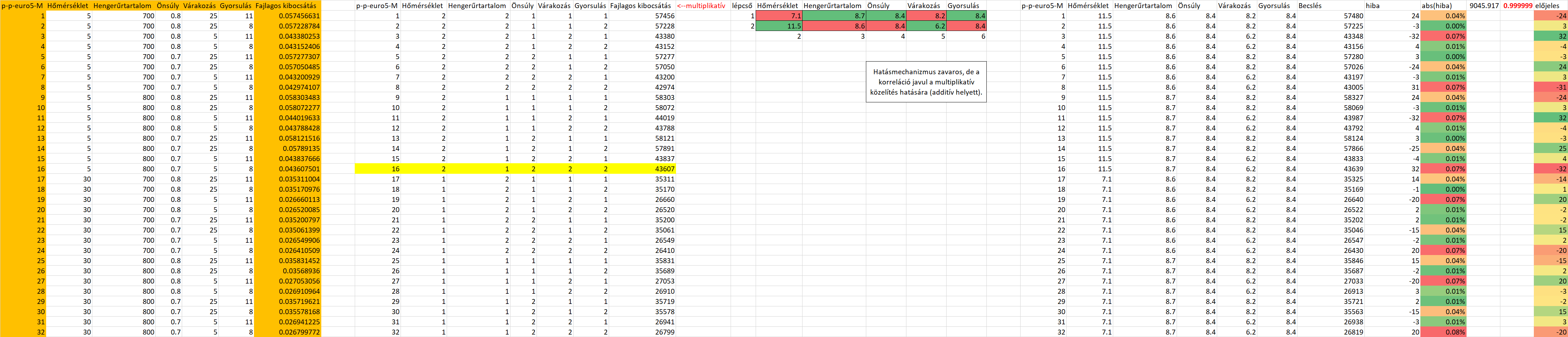


1. ábra: A hatás-dashboard-ok (forrás: saját számítások)

A 20. ábra értelmében:

* a hőmérséklet hatása csak látszólag más az irányított és az irányítatlan modellek esetén, hiszen éppen itt történt a bináris jelek invertálása a fordított arányosság irányított leképezése érdekében,
* az 5. ábrával összevetve a delta (tény-becslés) értékek szinte minden viszonylata más, lévén az 5. ábra mögött az 1-es és a 2-es értékek aránytalansága állt és a korrigált inputú tanulási folyamat a becslésekre is érdemben kihatott,
* az mcm és a std modellek hatásmechanizmusai is eltérők (bár ezek azonos input-mintázatra vonatkoznak), vagyis alternatív megoldásként értelmezhetők…

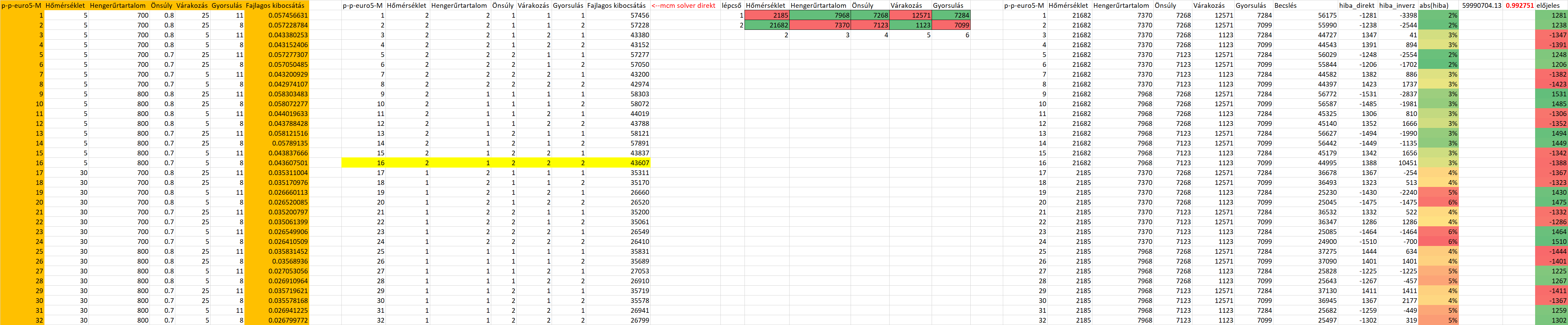
A 6. ábra a hibás input-mintázatra vonatkozó multiplikatív tanulást mutatja be, míg a 21. ábra ennek párjaként a korrigált inputú tanulás szorzatos modelljét ábrázolja:



1. ábra: Multiplikatív modell (forrás: saját számítások)

A 21. ábra tanulságai:

* a korreláció quasi 1 (vö. 0.999999), vagyis a nem lineáris NOX-számítási rendszer szorzatos lépcsősfüggvényekkel helyes input-mintázat teljes kombinatorikai tere esetén quasi tetszőlegesen pontosan közelíthető,
* igaz, az egyenes és fordított arányosságok alaplogikája a várakozás esetében nem tárult fel automatikusan teljes mértékben, de a hibás inputokhoz képest 5 változóból 4 esetben az irányok is helyesek,
* a tény-becslés eltérések előjeles mintázata (jobb szélen) nem mutat érdemi kombinatorikai tagozódást a színkódok alapján a 18-19. ábrák becslési hibáival összevetve
* a rel. kis hibák jelenléte nem zárja ki a további műveletek (vagyis nem lesz az egyik részeredmény tényekkel való tökéletes fedettsége miatt oka-fogyott a részeredményekből történő hibridizáció)…

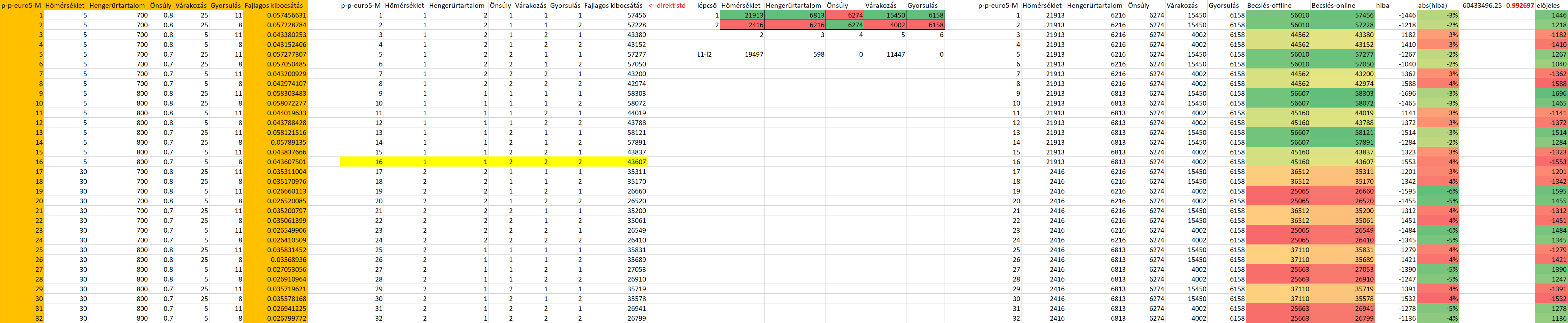


1. ábra: Az mcm modell solver-alapú nézete (forrás: saját számítások)

A 22. ábra offline keretrendszerben levezetett 0.992-es korrelációja nagyobb, mint az online megoldás esetében (vö. 6 ábra). Emellett:

* a várakozás kauzalitása itt is fordított arányosságot mutat, szemben az online megoldással, de párhuzamban a multiplikatív megoldással,
* a tény-becslés előjeles eltérések mintázatot mutatnak, de eltérőt az online megoldásétól…

A 23. ábra a 9. ábra párjaként jött létre:



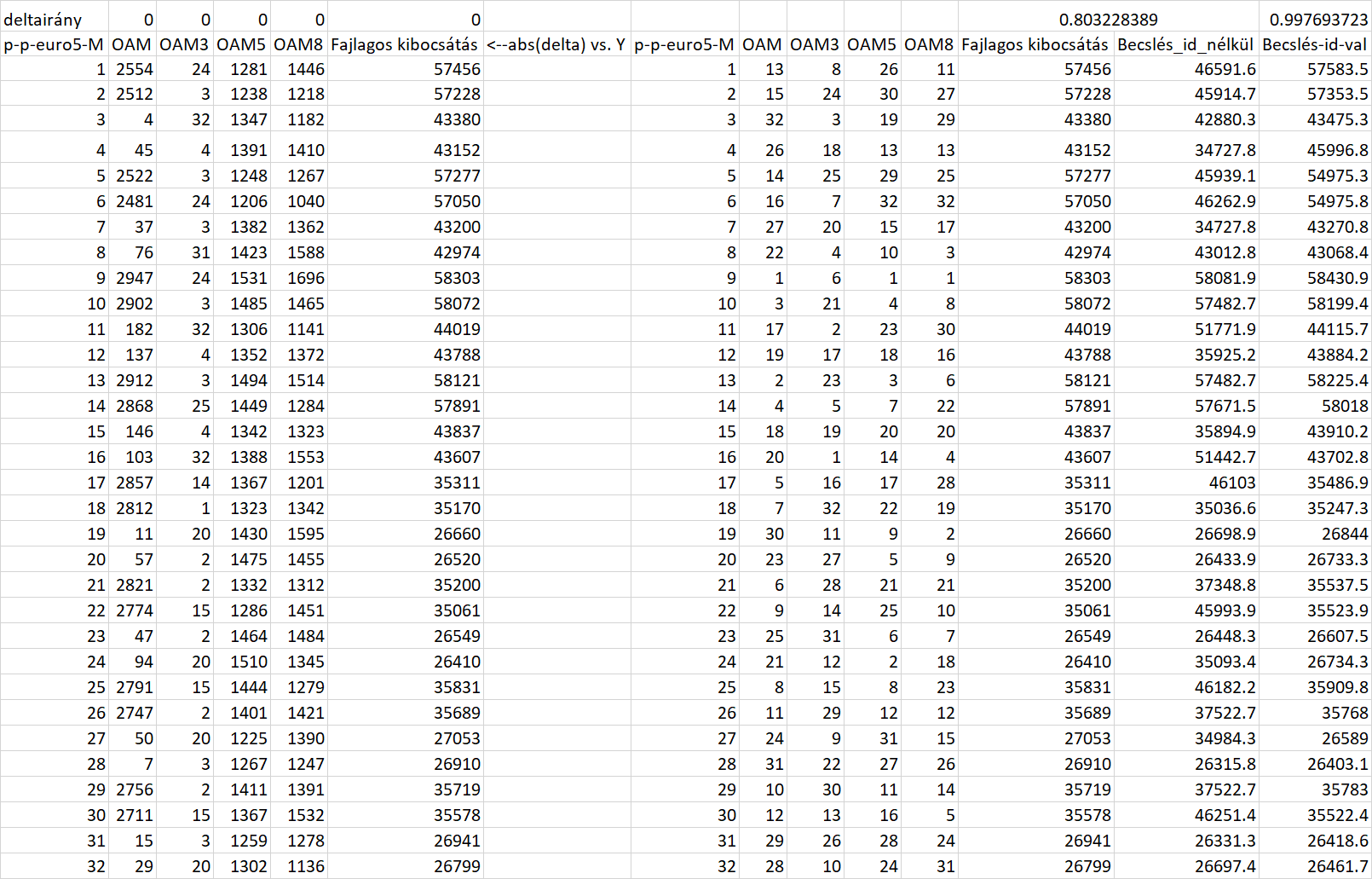
1. ábra: Offline standard modell (forrás: saját számítások)

A 23. ábra értelmében:

* a korrelációs érték jobb, mint a 9. ábra esetében, ami megfelel a javított input-minta nyomán elvárható hatásoknak,
* itt (offline esetben) az önsúly kauzalitása nem került felismerésre,
* a korrelációs érték jobb, mint az online esetben (mindkét OAM-ban javított inputokkal – értelemszerűen), ami az offline négyzetes-hibára-alapozó tanulás rugalmasságát erősíti, hasonlóan az irányítatlan esetekhez (online vs. offline)…

S ezzel elő is állt a hibridizációhoz szükséges 4 becslés:

* az online additív MCM (OAM)
* az offline multiplikatív MCM (OAM3)
* az offline additív MCM (OAM5)
* és az offline additív STD (OAM8)

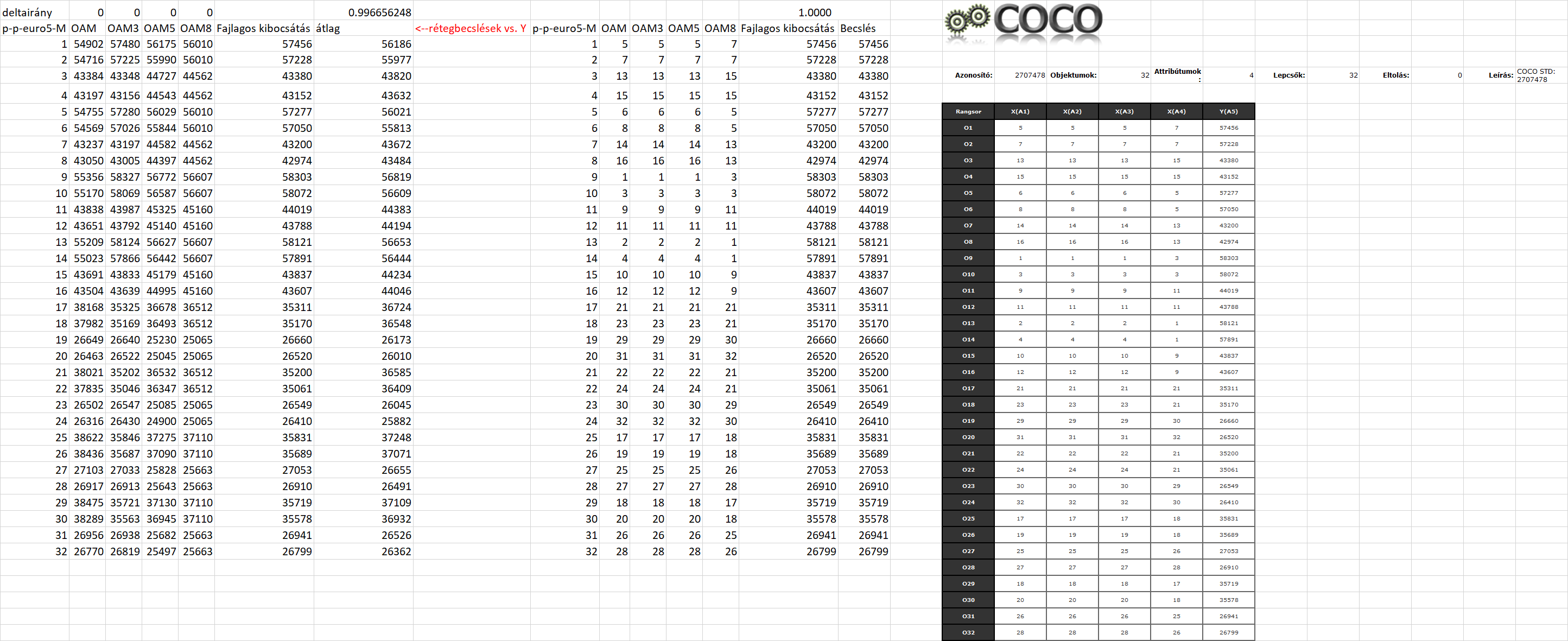


1. ábra: Hibridizáció id-val és id nélkül (forrás: saját számítások)

A 24. ábra üzenetei:

* A 10. ábra párjaként, ahol szintén abszolút hiba alapján kerültek a becslések finomhangolásra, az id nélküli alapmodell helyes inputminta esetén kisebb korrelációt mutat (0.80<0.87),
* az id-val együtt készülő minta korrelációja is kisebb (0.999<0.997)
* de az id-hatás itt is tetten érhető

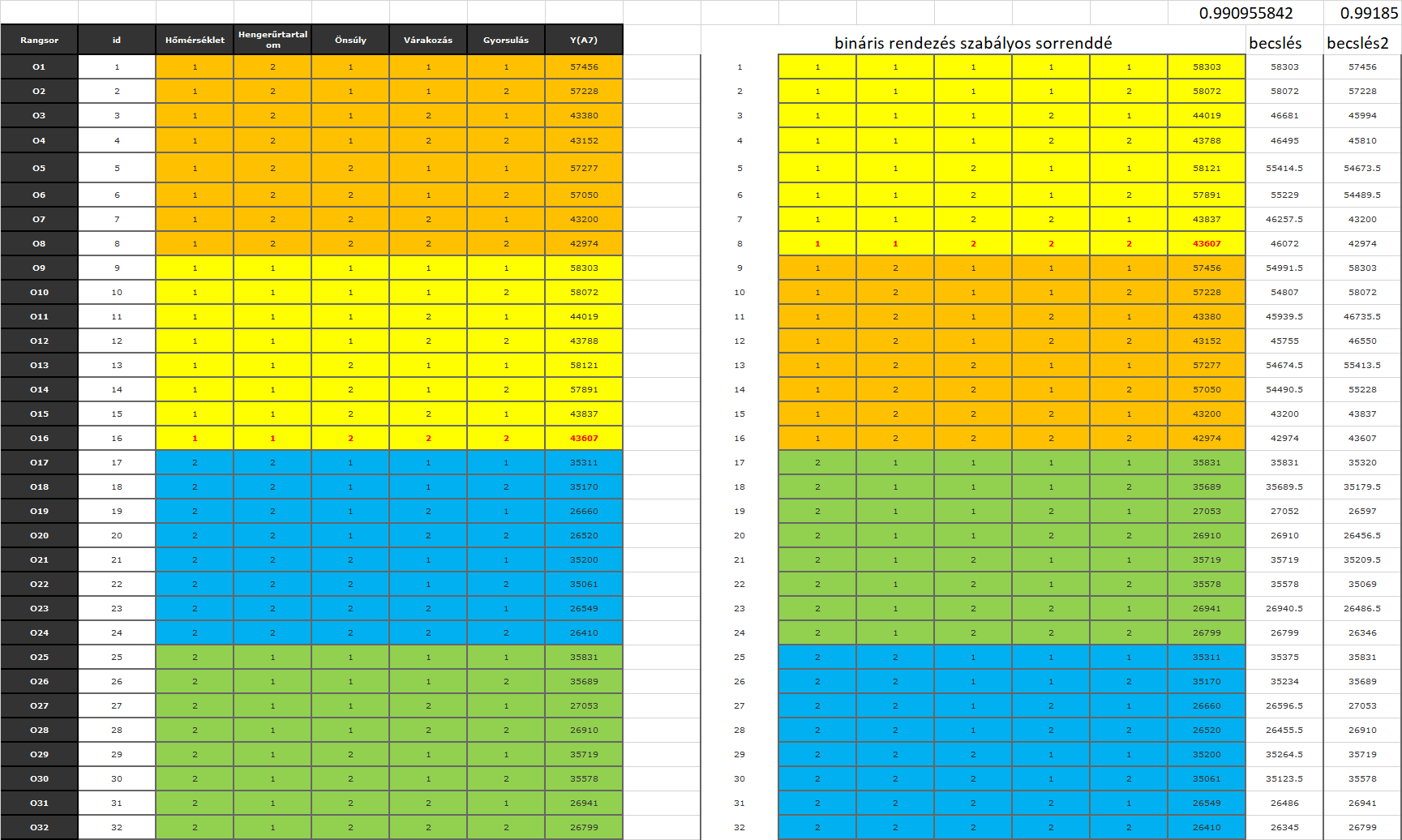
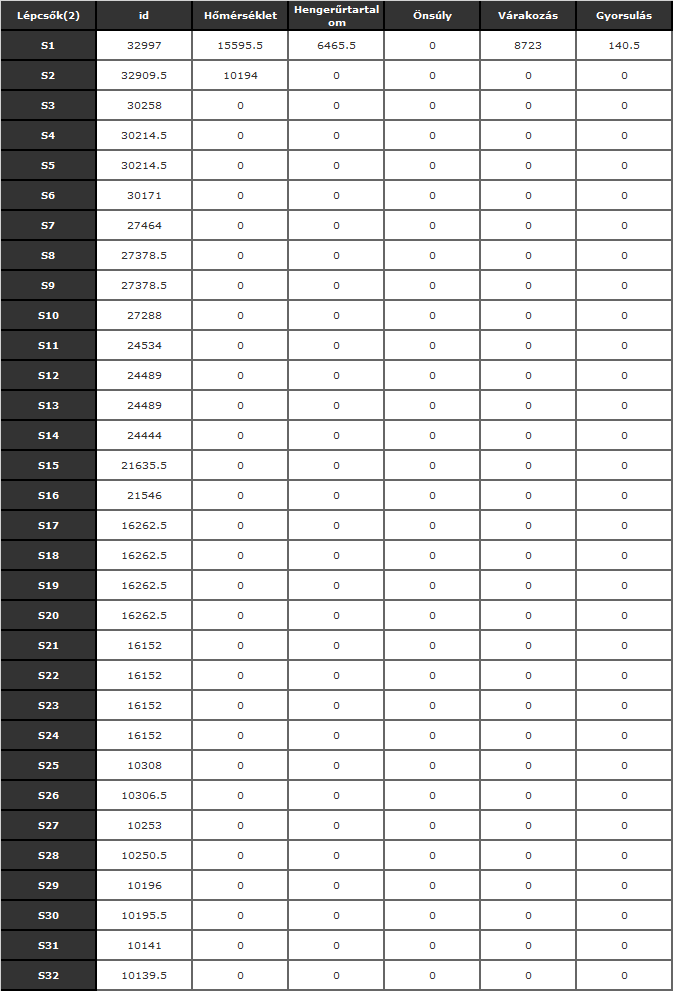
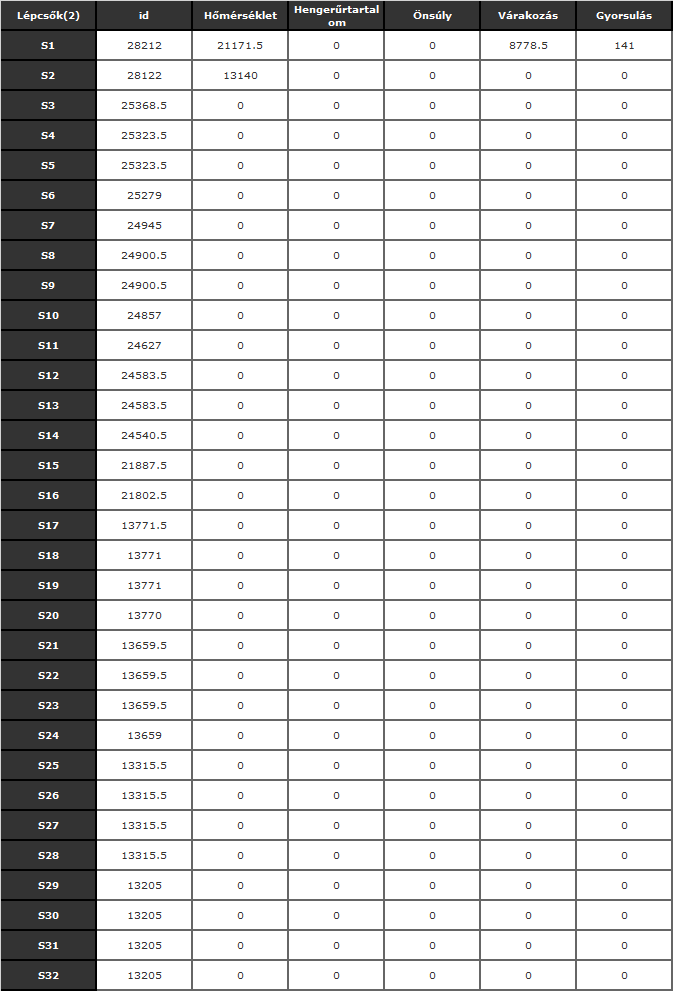
A 25. ábra a hibridizáció azon vetületét mutatja be, ahol az abszolút hibák helyett a direkt becslés-rétegek adják az inputokat. A hibridizáció abszolút hibával való lefolytatása éles esetben nem létező kalibrációs rendszer, mert egy ismeretlen Y esetén nincs hibaszámítási lehetőség. Az alternatív becslésekre hagyatkozó hibridizáció mindenkor létező modellt ad, hiszen alternatívákat mindenkor elő lehet állítani.



1. ábra: A becslésekre alapozó becslés (forrás: saját számítások)

A 25. ábra tanulságai:

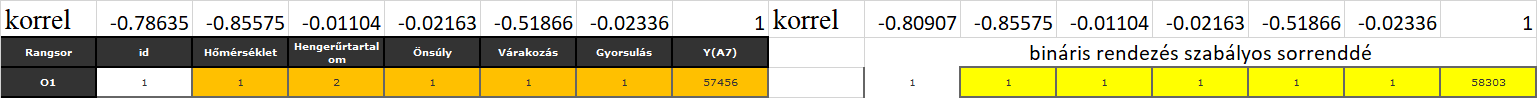
* Három alternatív becslés sorszám-nézete is azonos és ráadásul azonos az tények sorrendjével…
* Így egy hibrid modell kényszerűen hibátlan kell, hogy legyen,
* Vagyis itt már a sorszámok hatása nem vizsgálható
* A három (OAM, OAM3, OAM5) modell becsléseinek átlaga nem kell, hogy jobb modellt adjon a legjobb modellnél (vö. majdnem tökéletes multikplikatív modell), hiszen a hibridizáció lényege éppen a legjobb részmodell-arányok fellelése lenne,
* A hasonlóságelemzés ideális sorszám-inputoknál már nem segít tovább, ilyenkor vissza kell lépni pl. a regressziós modellezés világába, ahol tetszőleges numerikus kölcsönhatások kényszeríthetők ki…

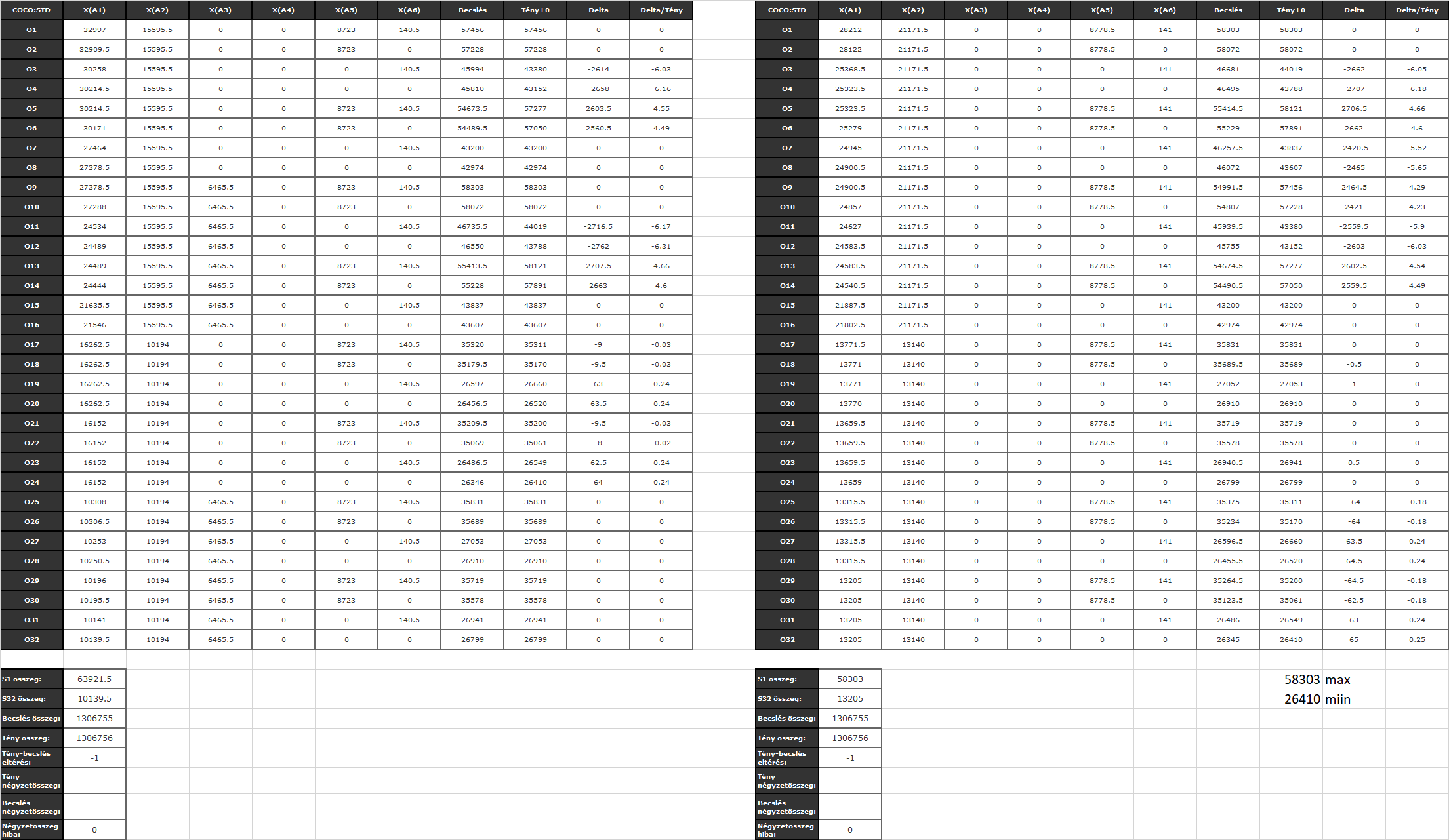
1. ábra: Az id hatása az alapproblémára (forrás: saját számítások)

A 26. ábra az alapprobléma (vö. 15-16. ábra, ill. 19. ábra), vagyis az online additív invertált hőmérséklet-mintázatú OAM sorainak sorrendjét megváltoztató, quasi bináris sorrendbe rendező új OAM hatásait mutatja be az id bevonása mellett:

* az id bevonásával a korreláció (0.991) nem javult a 19. ábrahoz képest,
* ellenben a kombinatorikai teret alkotó változókból 2 darab is kiesett, míg az id dominánssá vált,
* vagyis mintha információértéket vett volna át, ill. más megfogalmazásban: mintha eleve információértéket hordozna,
* ami nem csoda, ha kiszámoljuk az id-változók és az output-változó közötti korrelációkat, melyek az elvárásoknak megfelelő a rendezett sorrend esetén magasabbak (-0.786 vs. -0.809), de a konkrét esetben ez a jelképes többlet nem manifesztálódik becslése pontosságnövekményben:



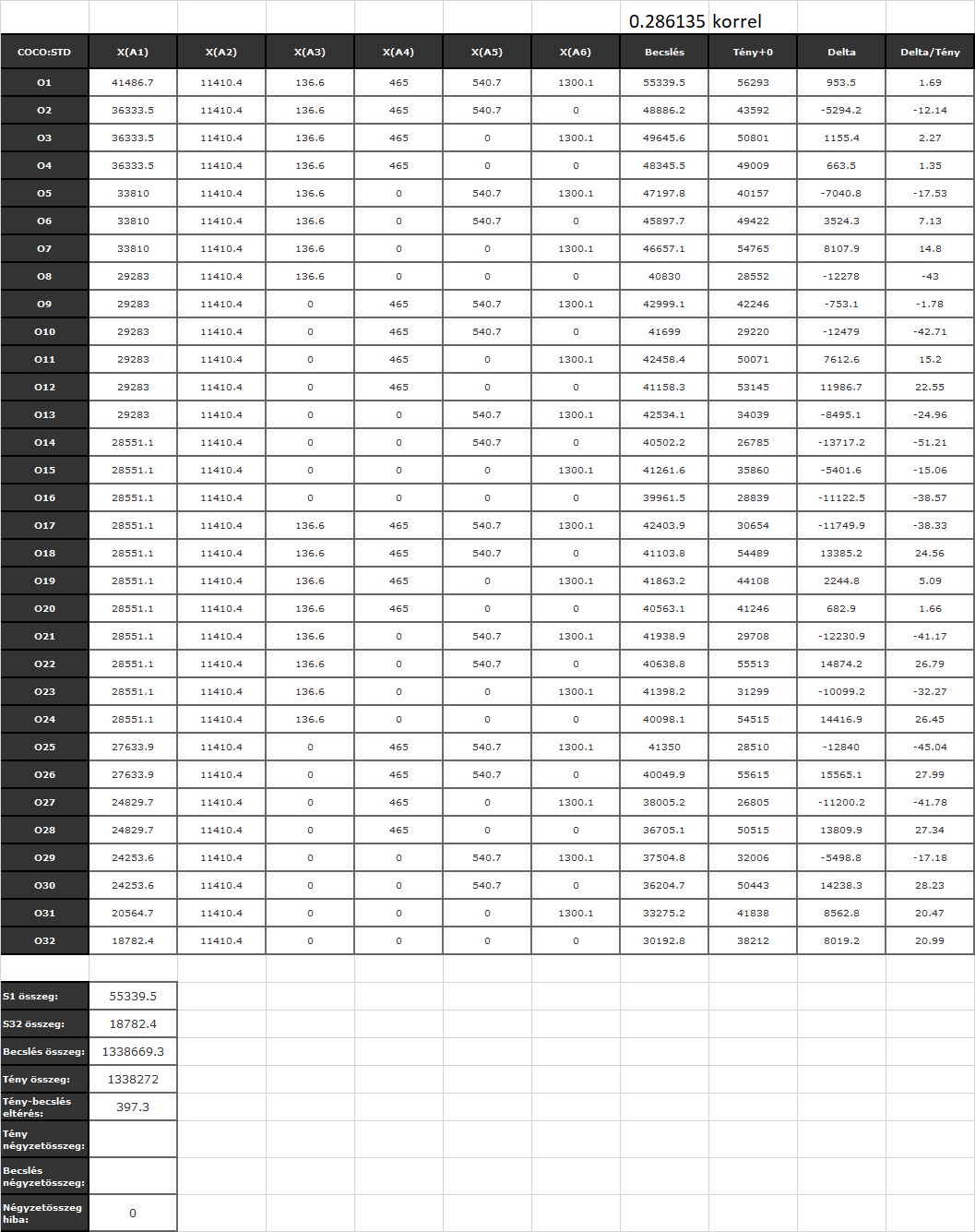
* az információt azonban nem a sorrend mibenléte hordozza, lévén a bal oldali nézet a quasi véletlenszerű alapsorrendje az OAM rekordoknak (becslés2: korrel = 0,9918), míg a jobb oldala a quasi binárisan nagyságrend szerint rendezett rekordsorrend (becslés: korrel = 0.9908) – vagyis a jobbnak vélt rekordsorrendnek nincs pozitív hatása a korrelációs értékre,
* ellenben a jobbnak vélt OAM-sor rendezettség esetén az ID hatására két változó esik ki, míg véletlenszerű sor-sorrend esetén csak 1 változó,



1. ábra: Az id hatás és a rekordsorrendek kapcsolata (forrás: saját számítások)

A 27. ábra a 26. ábra kiegészítéseként a becslés-levezetés részleteit mutatja meg:

* A bal oldalon látható a quasi véletlen sor-sorrendű OAM, ahol a teljesen hibátlan (nulla hibás) becslések száma nagyobb, mint a jobb oldali rendezett sor-sorrendű esetben
* s az 1%-os relatív hiba felett is a bal oldalon van kevesebb a jobb oldallal szemben – az elvárásokkal ellentétben a sorrendiséget illetően, de az elvárásoknak megfelelő a bevont változók számát illetően,
* a jobb oldali leíróképesség közelebb van a leírandó max-min tartományhoz, mint a bal oldali…



1. ábra: Egy véletlen következményű OAM (forrás: saját számítások)

S végül a 28. ábra egy ideális (input-sorrendű) OAM-hoz társít a már ismert Y max-min értékei közé véletlen szám generátorral létrehozott Y értékeket. Az eredmény magért beszél: az ID-hatása sem segít a korreláció (0.286) értelmes szintre emelésében:

Konklúzió: az ID-hatás reális input-output párok feltárásának új megközelítését rejti magában!

# Több motortípus összevetéséből származó adatok ellenőrzése

Amennyiben quasi véletlenszerűen állnak csak input-output kapcsolatokról rendelkezésre adatsorok, akkor is lehetséges az input-output kapcsolatrendszer függvényszerűségét (az összefüggések létének esélyét, vagyis a véletlentől való eltérés vélelmét) vizsgálni. A 29. ábra egy gondolatkísérlet alapján annak vázlatát mutatja, miként írhatók le heterogén/nem-teljes kombinatorikai terek a konkrét esetben. A 30. ábra pedig már egy konkrét, kísérleti adatösszeállítást tartalmaz, mely abban mindenképpen eltér a gondolatkísérlettől, hogy hibrid-járműveket nem sikerült belekényszeríteni a tanulási mintába (vö. a konstrukciós változó tartalma homogén), ellenben megjelenik egy új változó: az SSS, vagyis a start-stop-system rendelkezésre állásának valószínűsége.

A kérdés nem más, mint kiszámítható-e, s ha igen, értelmezhető-e az egyes attribútumok minden egyes opciójára hatásmértékek a NOX, mint output kapcsán?

Az átlagos hatásmértékek becslésére irányítatlan hasonlóságelemzésekre van szükség, melyek tetszőleges opciókhoz tetszőleges hatásmértékeket képesek rendelni az opciók potenciális viszonyrendszerét egyelőre nem értelmezve.



1. ábra: A heterogén inputok egységes leírásának terve (forrás: saját ábrázolás)



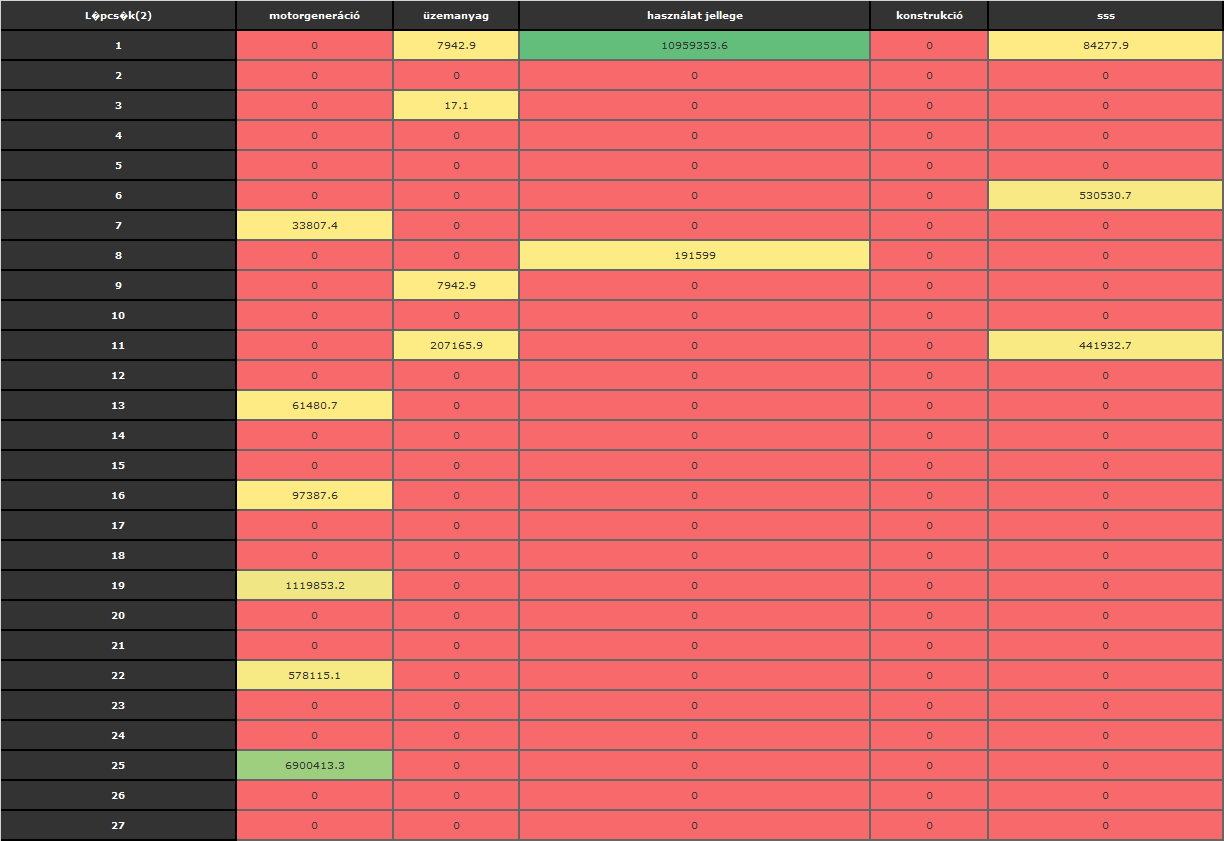
1. ábra: Tényleges input-output-kapcsolatok a számolótábla alapján (forrás: saját ábrázolás)

A 31. ábra a 30. ábra alapján készült az apró, dőlt betűs ceteris paribus alakzat kihagyásával és a ténylegesen vizsgálandó attribútumok/opciók attribútumonkénti rangsorolásával:



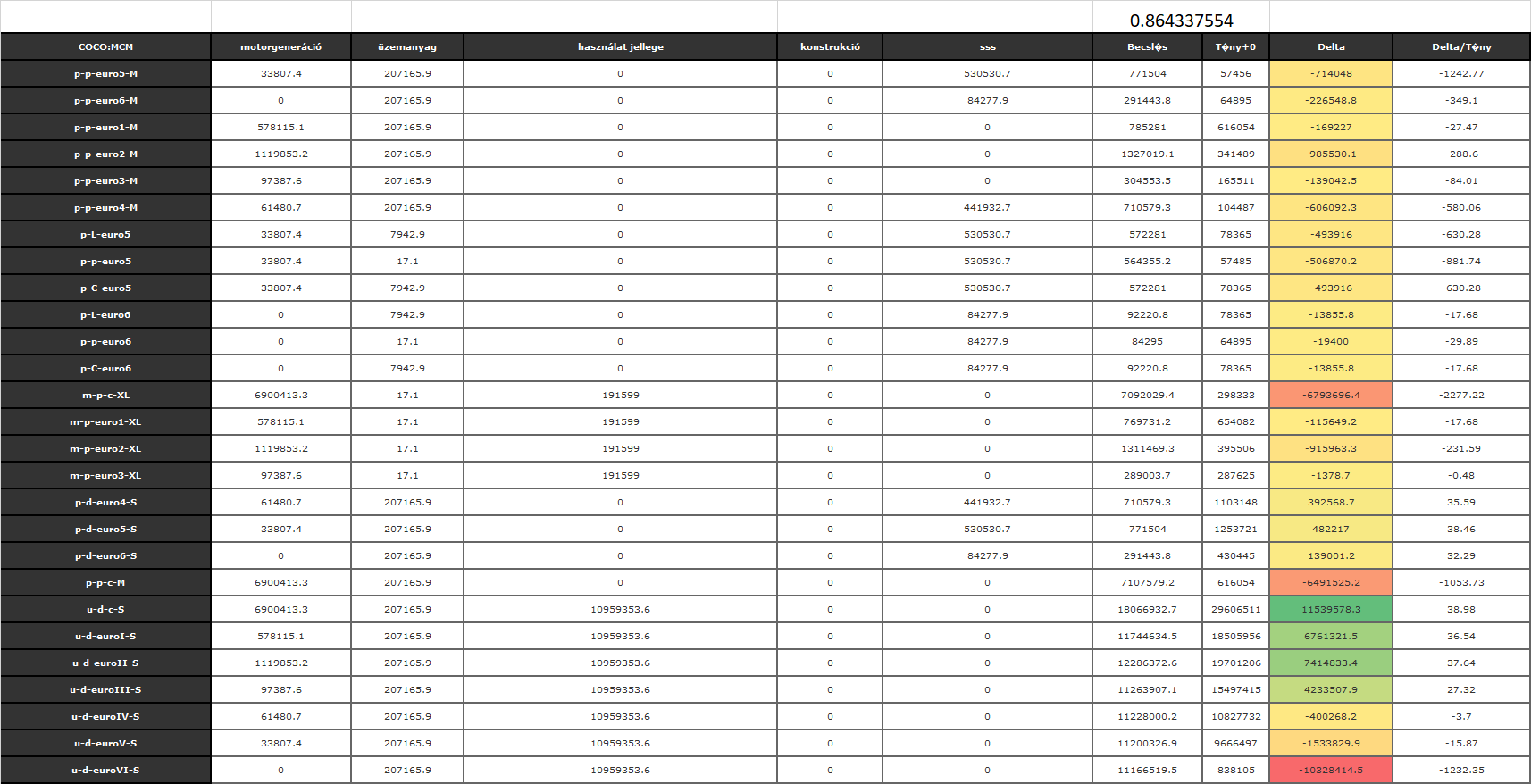
1. ábra: A 30. ábra rangsornézete az átlagos hatásmértékek feltárása érdekében (forrás: saját számítások)

A 32. ábra az irányítatlan tanulási folyamat nyers eredményét mutatja:



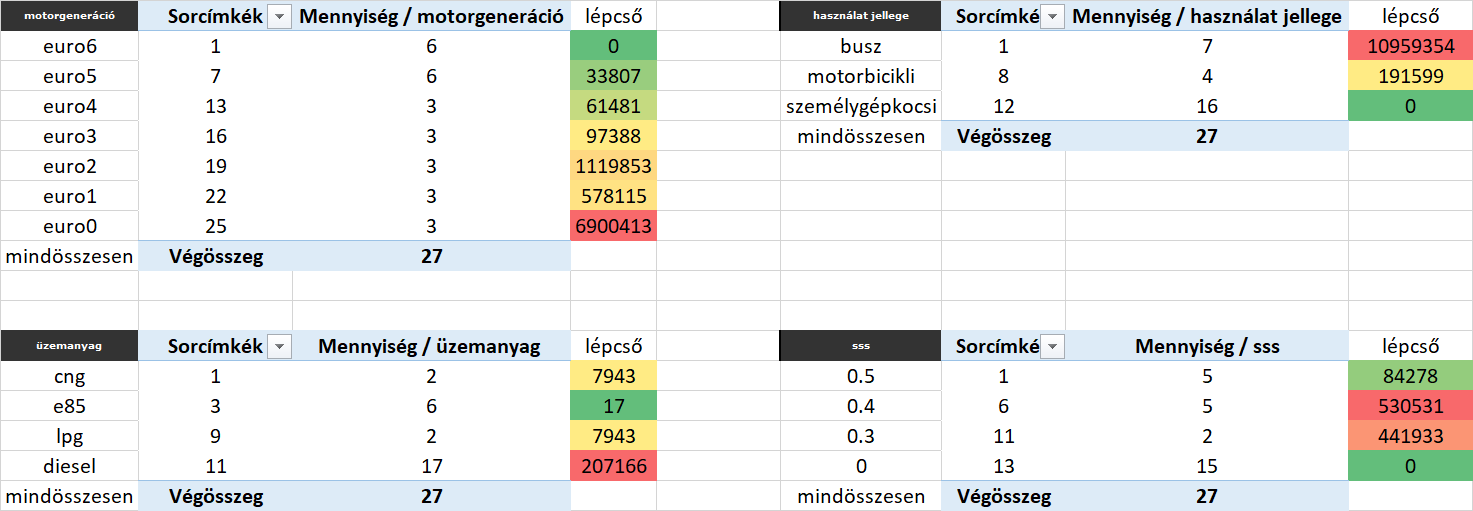
1. ábra: A heterogén inputú összefüggés felismert képe (forrás: saját számítások)

A 33. ábra a nyers lépcsős függvény feldolgozott rekordokra vetített nézete:



1. ábra: Az egyes motortípusok egyedi becslési hibái (forrás: saját számítások)

A 34. ábra pedig megadja az átlagos hatásmértéket az attribútumok minden egyes opciója esetén:



1. ábra: Átlagos NOX-hatásmértékek (forrás: saját számítások)

A 34. ábra humán szakértői értelmezései:

* az euro\* rangsorból az euro2 lóg ki, noha a nyersadatok szintjén még az euro6 vs. euro5 viszonyban is láthatók voltak tételes kérdőjelek, hiszen a szakértői vélemények szerint az euro\* fejlődési skálának illene minden egyes objektum/opció esetén kedvező NOX-hatást felmutatnia
* a busz-motorkerékpár-személyautó nézetben a buszok kapcsán az egy utasra vetített arányokon keresztül a fajlagos környezetterhelés még lehet kedvező, de a motorkerékpárok átlagos NOX-hatása elgondolkodtatóan magas
* a cng és az lpg azonos hatása legalább a gáz-üzem átlagos NOX-hatásán keresztül racionális, s a diesel negatív hatása szoros kapcsolatban látszik állni a járműgyártók botrányaival és a diesel járművek várható piaci részarányának csökken(t)ésével
* a sss-opciók NOX-hatásai kapcsán lehet leginkább arra utalni, hogy a valószínűségek nem feltétlenül helyesek
* mindösszesen a heterogén (quasi véletlenszerű) input-mintázatok alapján is masszív összefüggés léte vélelmezhető a NOX és ennek befolyásoló tényezői, ill. a motortípusok leíró adatai között…

# Irodalmi hivatkozások

…a hivatkozások a szövegközben találhatók…

1. „A matematikában Taylor-sornak nevezünk hatványfüggvényeknek egy speciális alakú függvénysorát. A Taylor-sorok határértékben gyakran előállítanak bonyolultabb függvényeket (például trigonometrikus vagy hiperbolikus függvényeket), melyek közelítő értékei így pusztán hatványozással kiszámíthatók. A függvények Taylor-sor alakjában történő felírását a függvények hatványsorba fejtésének nevezzük.” [↑](#footnote-ref-1)