Valós és szimulált adatvagyonok egyenszilárdságának vizsgálata

(Analysis of homogeneity of real and simulated data assets)

Pitlik László, Pitlik Marcell, Pitlik Mátyás, Pitlik László (jun) – MY-X team

Kivonat: A big data projektek gazdasági értelembe vett kockázataként értelmezhető, ha az éles adatvagyonhoz való hozzáférés elhúzódik. Az elemző algoritmusok fejlesztése szerencsére azonban már (többé-kevésbé) véletlen számokból álló adatvagyonon is megkezdhető. Ez nem csak felgyorsítja a fejlesztést, hanem egyben szinte kényszerűen nagyobb rugalmasság felé tereli a hermeneutikai, adatvizualizációs alrendszereket. Hiszen a véletlen számokból álló konstellációk általában a valós adatvagyonoknál nagyobb kombinatorikai teret fednek le, s az így kialakuló szcenáriók értelmezését is kikényszerítik a fejlesztőkből. Amint azonban vannak valós mérések, azonnal meg kell kezdeni a véletlen számokra alapozó részeredmények összevetését ezekkel. Ennek az összevetésnek az egyik lépése az, hogy a valós és a véletlen adatok karakterisztikáinak különbségei feltárásra kerülnek. Ideális esetben a véletlen számokból álló minta és a valós adatokból vett minta egyenszilárdsága adott, vagyis a két halmaz hasonlósága a halmaz-egyedek hasonlósági indexeinek szintjén azonos átlagra vezet. Ilyen típusú elemzésekre a COCO anti-diszkriminatív eljárása alkalmas.

Kulcsszavak: hasonlóságelemzés, homogenitás-vizsgálat, big data, véletlenszám-alapú adatvagyon

Abstract: The access to real measurements in case of big-data projects can be disturbed through arbitrary impacts, and this can lead to specific risks for the projects. Fortunately, in most of the challenges, analytical, hermeneutical, and visualization modules can already be developed based on (quasi) random data assets. This way makes possible not only the increasing of the project’s speed, but the flexibility of the interpretation potential will also be extended. It should mostly be so, because the covered part of the combinatorial space in case of randomized constellations is wider than in case of real measurements and their combinations. Therefore, it is necessary to be capable to interpret them too. If real data are available, then it is necessary to compare the to the randomized data assets. One of the first steps is, to proof, whether the records of the two sources can be seen being elements in the same set, where anti-discriminative similarity analyses are able to deliver index values. In ideal case, the average similarity index values for the two groups should be identical.

Keywords: similarity analysis, homogeneity analysis, big-data, randomized data asset

# Előzmények

A közlekedési döntések racionalitásának GINOP-projekt keretében történt komplex modellezése és a kapcsolódó további 12 tanulmány világosan rámutat arra, hogy már véletlenszám-alapú adatvagyonokon is racionális modellezési aktivitások végezhetők: 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_ertekelse.docx>   
10. <http://miau.gau.hu/miau/241/only_one_engine.docx>   
11. <https://miau.my-x.hu/miau/242/onvezeto_parkolas_v1.docx>   
12. <http://miau.my-x.hu/miau/242/egyenszilard_adatvagyon_v1.docx>   
13. (<https://miau.my-x.hu/miau/243/meres_vs_rnd.docx>)

# Bevezetés

Tételezzük fel, hogy egy információs rendszer kísérleti jelleggel egy járműben elhelyezett okostelefon applikációjának segítségével méri egy-egy jármű GPS-koordinátáját például 1 másodperces ritmusban, mely koordináták alapján a jármű maga elhelyezhető a térképen, s így az utcahálózatban, vagyis pl. levezethető térképen megadott közlekedési lámpákhoz mért távolsága is az időtengely mentén. A GPS koordináták mellett mérésre kerül ismét csak pl. másodpercenként a jármű sebessége (m/s), mely verifikálható az utcahálózat és a GPS-koordináta-változások alapján (vö. vice versa). Valamint mérésre kerül a jármű gyorsulása (m/s2) is a fentiekkel azonos idő-egységenként, s így a sebesség-verifikálás egy újabb inputja áll elő (mely adat rel. ritkán mérve és pl. vészfékezést feltételezve már zavarokat okozhat a fizikai valóság reprodukálásakor). A szintén másodpercenként mért iránytű-adatok (0-360 fok - É=0) alapján a jármű utcahálózaton belüli mozgása két GPS-koordináta között is értelmezhető, ami ismét egy új ellenőrzési lehetőség a jármű-adatok konzisztenciáját illetően.

A jármű mellett mérésre kerülhet a közlekedési rendszerben elhelyezett érzékelőkkel a levegő hőmérséklete (vö. mérőeszközben mért külső hőmérséklet - Celsius fokban, mely mérések gyakorisága: 5-10 perc/mérés. Hasonlóképpen mérésre kerül a légnyomás (hPa) és a páratartalom (%), valamint a NOX (ppm) értéke is a fenti 5-10 perces gyakorisággal.

Ahhoz, hogy egy projekt a virtuális valóságban (pl. egy szimulátor-fejlesztés kapcsán) racionálisnak ható viselkedéseket tudjon felmutatni nem feltétlenül szükségesek valós adatok: elegendők lehetnek az ismert értelmezési intervallumokat kielégítő véletlen számok (esetlegesen az egyes attribútumok arányaira vonatkozó max-min intervallumokkal kiegészítve). A véletlenszámokból előálló adatvagyon hermeneutikája kapcsán általában véve nagyobb a lefedendő spektrum, vagyis quasi minden véletlen-alapú értelmező szabályrendszer képes a valóságot elfogadhatóan kezelni, míg ez fordítva nem kell, hogy igaz legyen, lévén a valóság adott mérési időszakban elképzelhető, hogy csak nagyon szűk spektrumát mutatja fel egy végtelen hosszú mérési időszak által ténylegesen lefedhető kombinatorikai térnek.

Amint azonban valós adatok rendelkezésre állnak, úgy utólag mindenképpen vizsgálandó a szóban forgó rendszer tesztelésének speciális alakzataként, vajon a valós és a szimulált adatok által leírt univerzumok mennyiben térnek el egymástól. Ideális esetben a két valóság elemeinek átlagára igaz, hogy minden valóság lehet másként egyforma!

# Véletlenszámok és valós adatminták

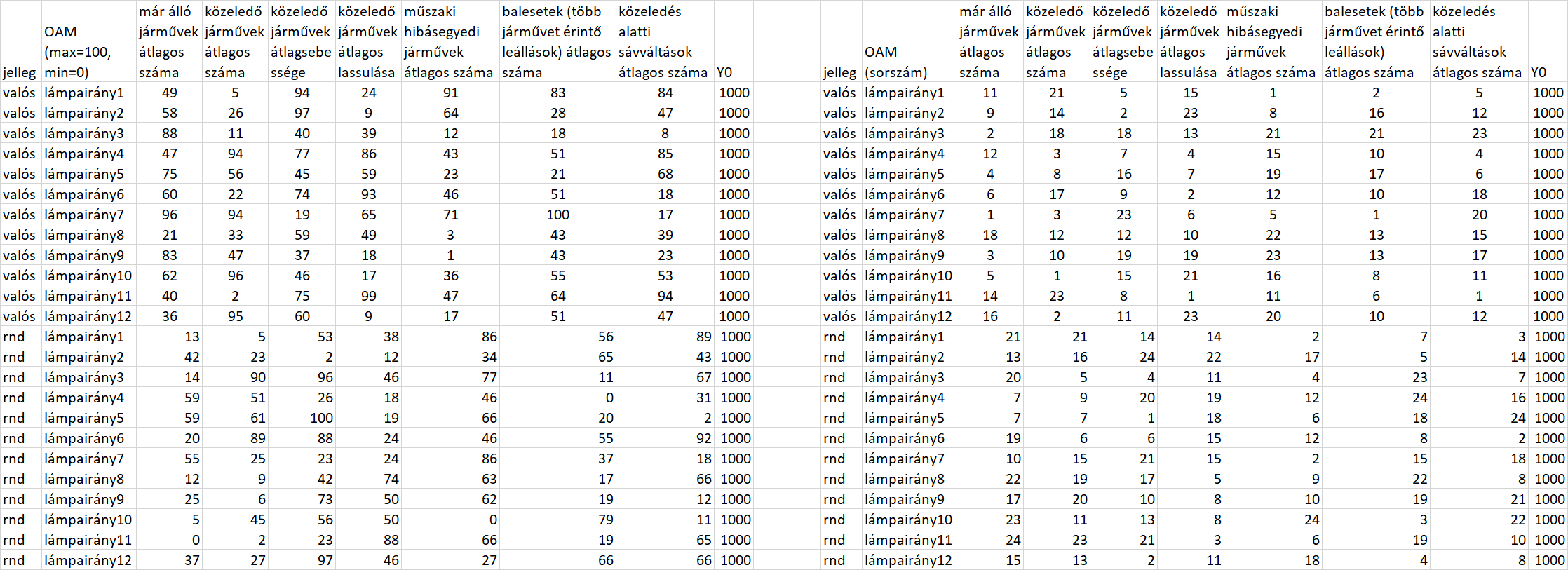
Amennyiben egy projekt keretében a nyers (akár véletlenszám-alapú, akár valós adatvagyonok) nem önmagukban fontosak, hanem csak ezek elemzésre méltó nézetei, akkor az adatforrások összevetését is ezen nézetek mentén kell végrehajtani. Ilyen nézet a szóban forgó GINOP-projektben pl. az OAM, vagyis az objektum-attribútum mátrix, ahol pl. objektumként lehetséges lámpairányokat/sávokat értelmezni, attribútumként pedig pl. egy adott lámpás kereszteződés minden sávjára/lámpairányára vonatkozóan az ezen a sávon álló, ilyen-olyan sebességgel ilyen-olyan távolságból adott lámpához adott sávon közeledő járművek karakterisztikáit (pl. álló és mozgó járművek darabszáma, átlagos, maximális, minimális állásidő, átlagsebesség, átlaggyorsulás, egyedi sebességek szórása, egyedi gyorsulások szórása, stb.) lehet kiszámítani. Maguk a nyers adatok (vö. GPS koordináták, iránytű-adatok, stb.) tehát nem önmagukban fontosak, csak a lámpairányokhoz, sávokhoz képesti mibenlétükön keresztül.

Következésképpen az így kontextusba helyezett nyersadatokból kialakított attribútumok esetében ezek értelmezési intervalluma is levezethető: pl. egy konkrét lámpás kereszteződésben a legközelebbi szomszédos lámpa távolsága legyen pl. 500 méter. A legkisebb létezőnek vélt jármű (pl. egy motorkerékpár) hossza legyen 2 méter, két jármű között vélelmezzünk 0.5 méter átlagos biztonsági távolságot álló járművek esetén, így a minden irányban 500 méterre limitált vizsgálati zónában az álló járművek maximális száma legyen 500/(2+0.5)=200 db, vagyis az adott időablakban (pl. egy lámpaprogram körbeérésének ideje alatt) az álló járművek maximális értéke 0-és 200 között lehet bármely sávon/lámpairány esetén. A járművek egyedi sebessége egy lakott területen lévő lámpás kereszteződés környezetében elvileg 0 és 50 km/h lehet, de a szabálytisztelet vélelmezhetően nem teljeskörűsége okán a felső határ legyen pl. 70 km/h. Az ennél nagyobb egyedi sebesség olyan ritka és az átlagra olyan kis hatást gyakorol, hogy a felső határ további növelése nem racionális. Vagyis, ha minden egyes OAM-attribútumra képesek vagyon értelmezési intervallumként a maximum és a minimum értékét megadni, akkor az ezen belüli eloszlás ismerete nélkül is a véletlenszámok quasi egyenletes eloszlással le fogják fedni a kombinatorikailag kijelölt mozgásterüket. A valós ehhez képest lehet pl. normál-eloszlású, mert a közeledő járművek egyedi sebességei a nagyon alacsony (=rácsorgó jelleg) és a nagyon gyors (=vészfékező) objektumok nagyon kis előfordulásával szemben zömmel egymást racionális követő tömegeket enged felismerni. Vagyis ideális esetben, amikor nem csak a maximális és a minimális értékekre lehet racionális becslést adni, hanem az ezen az intervallumon belüli előfordulási gyakoriságokra is, akkor tulajdonképpen a véletlenszámok már a valóság véletlenszerűségét képezik le, vagyis per definitionem nem illik, hogy különbözzenek a helyesen becsült paraméterek miatt a valóság által ezen paramétereket megerősítő történésektől.

Jelen cikk érdekében egy tipikus, konkrét lámpás kereszteződés került figyelembe vétel, ahol minden lámpának önálló sávja is van. A járművek nyers mérési adataiból az 1. ábra fejlécében szereplő attribútumok kerültek leképezésre az adott kereszteződés és sok-sok időszak átlagában kerültek kiszámításra ezen attribútumok értékei valós és generált adatok esetében, ahol a generálás teljesen véletlen szerű volt.

Az elsődleges elvárás az, hogy a két kezelés (valósság és generáltság) átlagos hasonlósági indexe az adatok együttállására vonatkozóan nem különbözik egymástól, vagyis minden objektum (sor) önmagában is közelíti a minden sáv lehet másként egyforma elvet, de ezek kezelésenkénti átlaga még inkább.

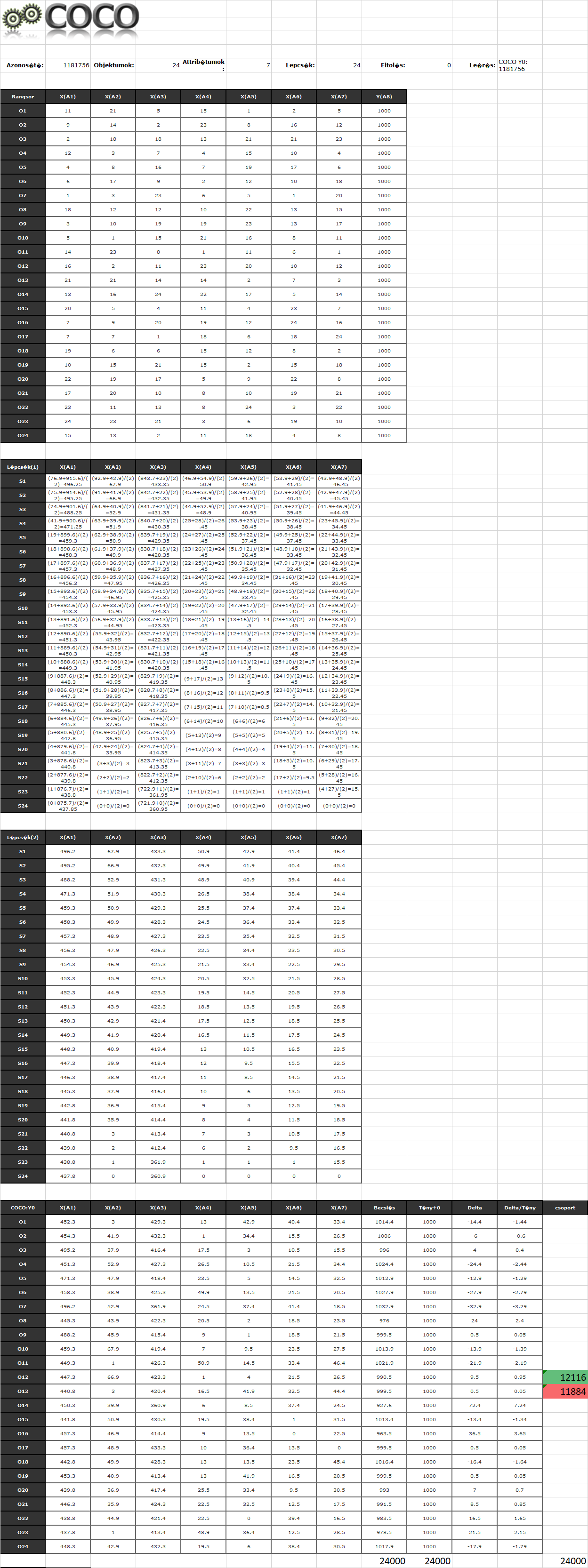
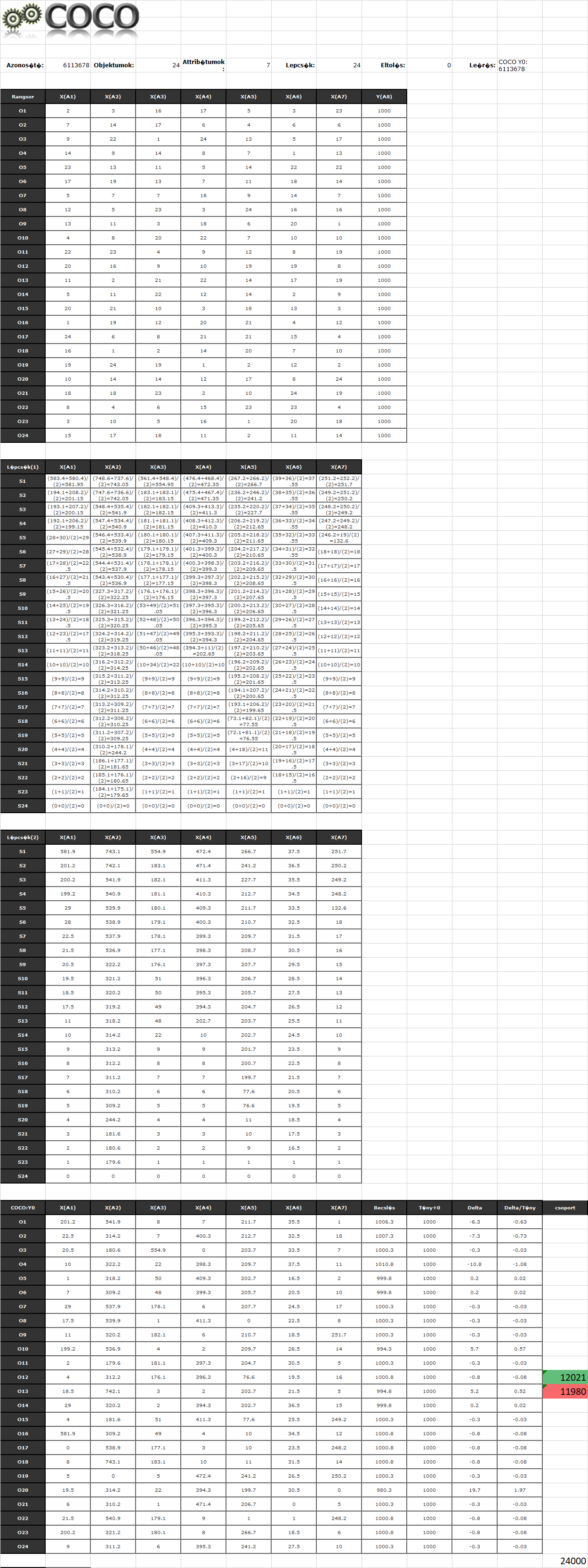
Az eltérő kereszteződés karakterisztikái teljesen eltérők lehetnek, így kereszteződésenként is más-más lehet a viszonya a valós és a generált adatoknak.



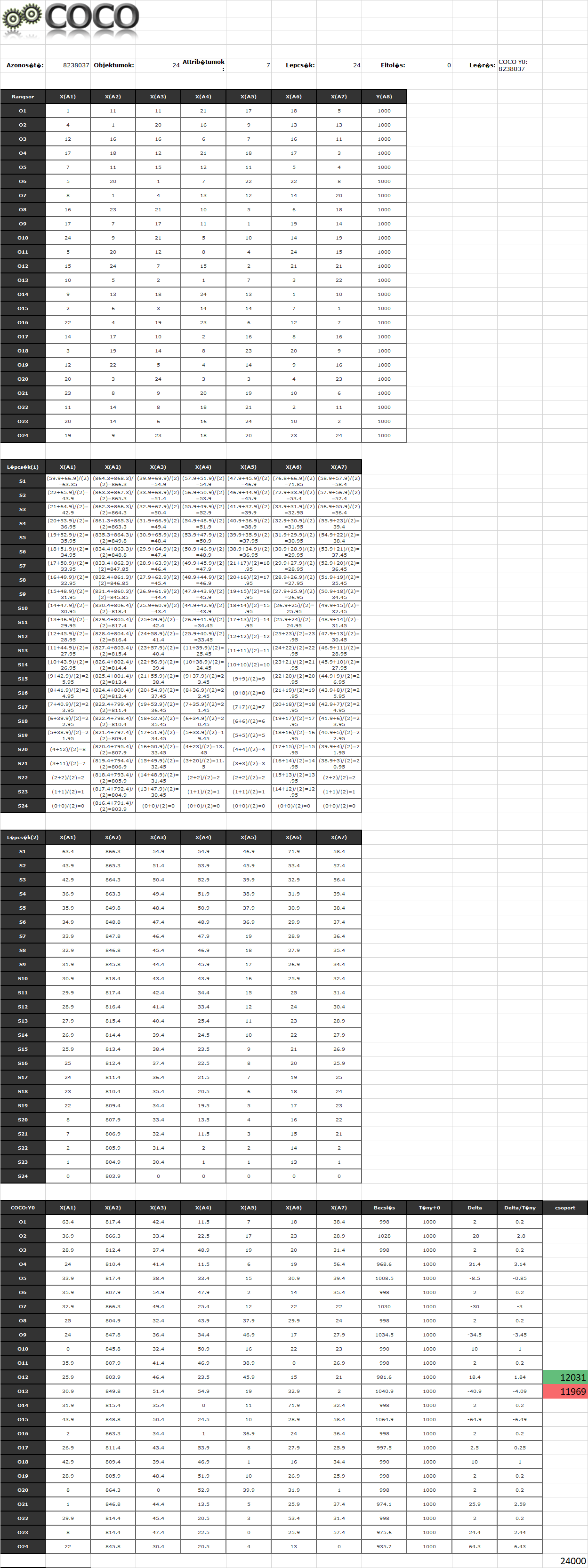
1. ábra: Véletlen és valós minta a mindenkori értelmezési intervallumon, azaz 0-100%-on belüli értékekkel egy adott kereszteződésre átlagosan (forrás: saját számítások)

Az alábbi 4 ábra eltérő kereszteződések egy-mérőszámos összehasonlítását mutatja be:

* a 2. ábra egy, a legkevésbé homogén és egy lényegesen homogénebb adatvagyonra mutat be példát
* a 3. ábra közepes eltéréseket felmutató kereszteződéseket demonstrál

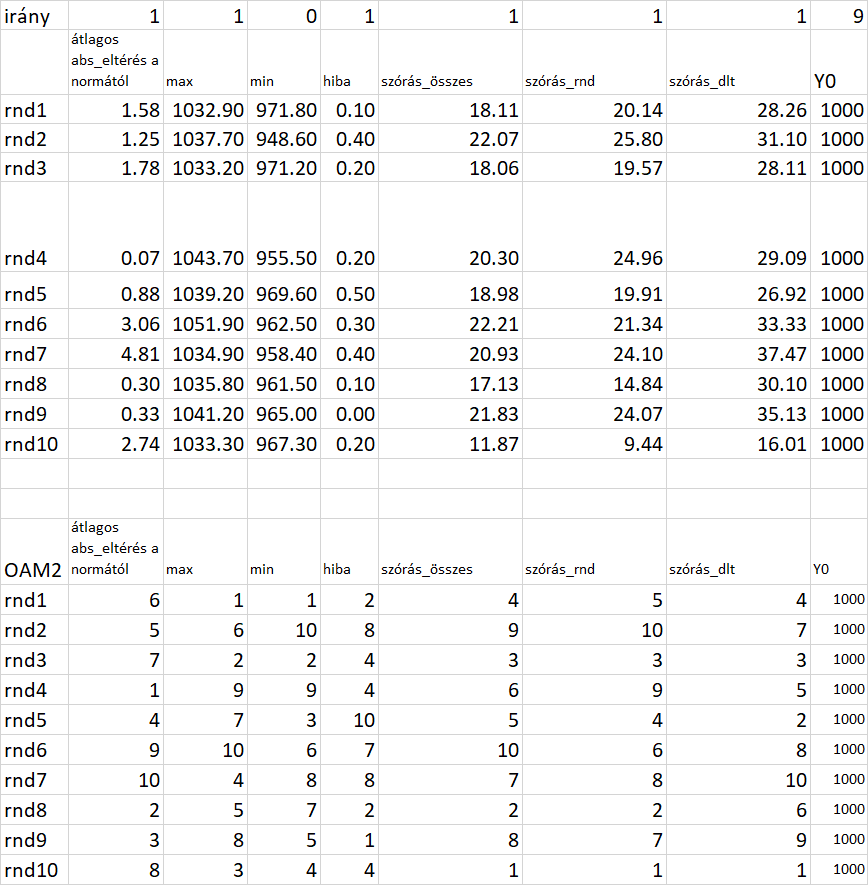
1. ábra: Egy jelentős és egy kisebb eltérést mutató kereszteződés (forrás: saját számítások)

1. ábra: Közepes eltéréseket mutató kereszteződések (forrás: saját számítások)

Amennyiben azonos valóság mellé/helyett számos generált adatvagyon kerül kialakításra (vö. érzékenység-vizsgálat, modellezési kockázatok elemzése), akkor az alaphipotézis, miszerint minél kisebb a valós és a generált csoport hasonlósági indexeinek átlaga közötti különbség, annál jobb, az alaphipotézis másik feltételével mindenképpen kiegészítendő, vagyis azzal, hogy a minden objektum minél közelebb áll a minden másként egyforma elvhez, annál inkább beszélhetünk a valóság és a generált adatvagyon hasonlóságáról:

Amennyiben egy adott valóság mellé 10 generált adatvagyon kerül kialakításra és minden egyes generált adatvagyon okozta hasonlósági (becslési) hatásokat több attribútumra bontva vizsgáljuk (vö. 4. ábra), akkor kimutatható, hogy az átlagos eltérés abszolút mértéke a normához (1000) viszonyítva és a komplex hasonlósági mutató közötti korreláció nem közelíti tetszőlegesen az ideális -1-es értéket (vö. 5. ábra: korreláció = 0.4669), ahol a negatív előjel a mérések logikája adja (vö. hasonlósági érték minél nagyobb, annál jobb, míg a normától való eltérés minél kisebb, annál jobb):

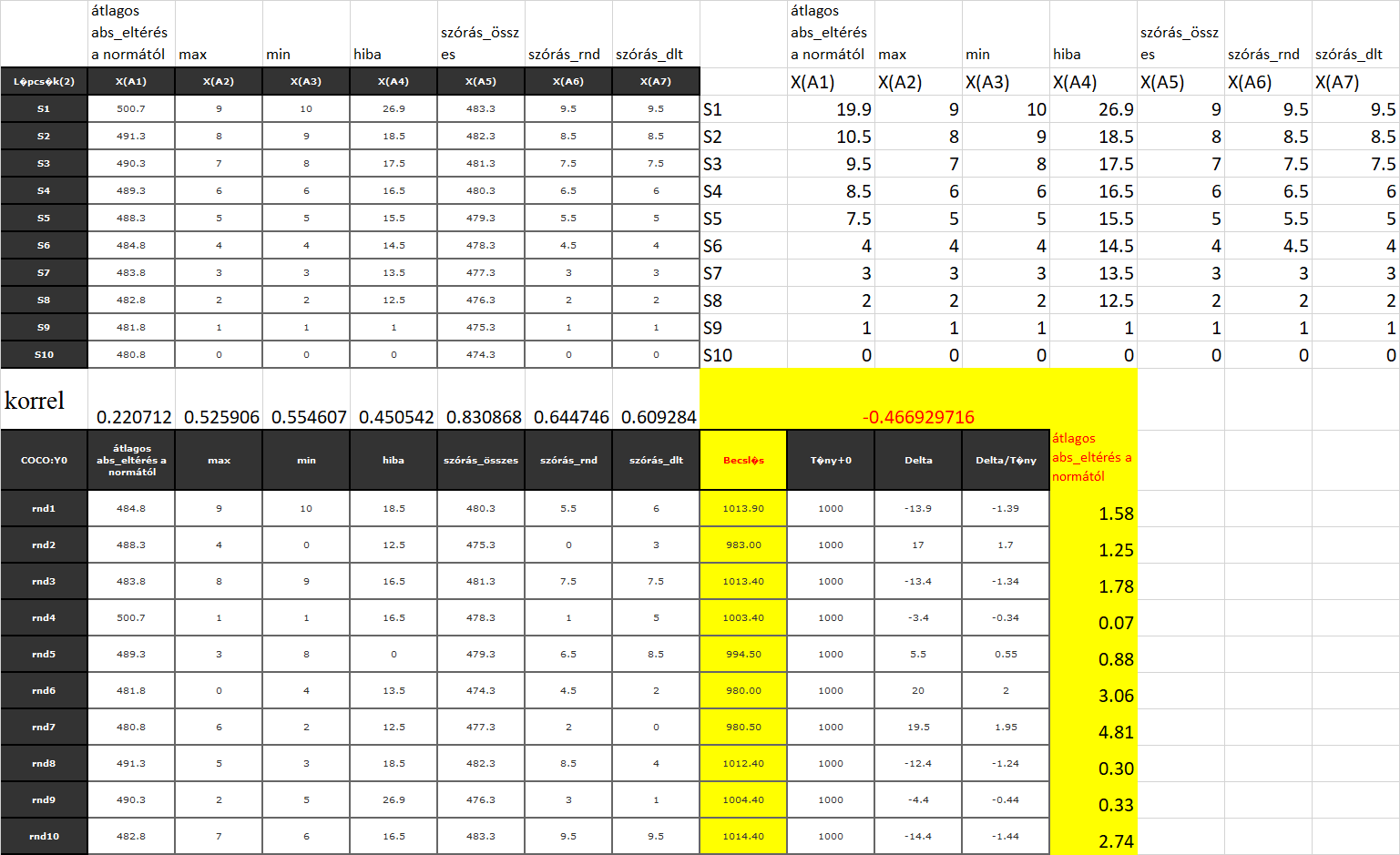


1. ábra: A generált valóságok hatása a valós és generált objektumok hasonlóságára (forrás: saját számítások)

A 4. ábra további attribútumai:

* max: a becslések maximuma
* min: a becslések minimuma
* hiba: a modell normától való eltérése
* szórás\_összes: az összes becslés szórása
* szórás\_rnd: csak a generált objektumokhoz tartozó becslések szórása
* szórás\_dlt: a becslési hibák szórása csak a generált objektumok esetén

Az attribútumok a minimum érték kivételével a minél kisebb annál jobb irányt követik, hiszen így írható le az azonosság ideálja a mutatószámokra visszavezetve. A minimum érték minél nagyobb annál jobb elvet követ ugyanezen okokból (vö. a max és a min eltérését a normától lehetett volna egységes irány=1, azaz minél kisebb annál jobb alapon is leírni).



1. ábra: Az alternatív generált adatvagyonok komplex értéke (forrás: saját számítások)

A hasonlósági transzformáció nyomán kialakult korreláció az átlagos normától való eltérés attribútuma esetén eltér a nyers adatok és a becslés közötti (0.4669-es) korrelációtól (vö. 0.220712). Ennek oka, hogy a sorrend-alapú standardizálás rugalmassága nagy.

Az alternatív generált világok hasonlóságainak komplex leírását az összes (valós és generált) objektum szórása közelíti a legjobban (vö. korrel=0.830868).

A legnagyobb hasonlósági erőtere a modellhibának van (vö. 26.9 vs. 0), ill. ezt követi a normától való átlagos eltérés maga (19.9 vs 0). A szórás\_összes értelemszerűen nem hat a hasonlósági erőtéren keresztül, lévén a magas korreláció a nyersadatokon keresztül adott a komplex becslés irányába.

Az attribútumok és a becslés közötti korrelációk kényszerűen pozitív értékek, hiszen ezek az adatok már átestek az irány-alapú standardizáláson – szemben a nyersadatokkal.

A normától való átlagos eltérés korrelációjának nyers és hasonlóságelemzés nyomán előálló nézete között a csökkenés a nyersadatoktól a feldolgozott adatok irányába is azt erősíti, hogy az átlagos eltérés a normától olyan komplex konstellációk keretében állhat elő, melyek ennek az attribútumnak az információértékét csökkentik.

# Hasonlóságelemzési vizsgálatok konklúziói

Jelen cikk alapvetése abból indul ki, hogy lehetőség szerint minden elemzést hasonlósági alapon kell végrehajtani, így a szignifikancia vizsgálatok is kiváltható/kiváltandók hasonlóságelemzésekkel.

A primer hasonlóság-vizsgálat két OAM összevonása alapján az egyes rekordok egymáshoz mért hasonlóságán keresztül a két adatforrás átlagos hasonlóságát képes feltárni, ahol elsődlegesen nem nyer értelmet az, mely rekord melyiknek a párja. Csak a komplex hasonlóság számításakor kerül olyan mutató értelmezésre, ahol ugyanazon objektum valós és generált objektumára vonatkozó becslések delta (dlt) értékeinek szórása kerül minimalizálásra.

A szekunder vizsgálat keretében pl. az összevont OAM rekordcsoportjainak hasonlósági indexei alapján új OAM-ot képes, ahol egy-egy (értelemszerűen mindenkor csak azonos forrásból származó rekordokat tartalmazó) rekordcsoport hasonlósági indexeinek maximuma, minimuma, szórása, stb. kerül levezetésre, majd immár ezen karakterisztikák kapcsán tevődik fel a kérdés: lehet-e minden csoport másként egyforma, ill. a csoport-hasonlóságok átlagos hasonlósága a két forrás esetén azonos?

A kereszteződésenkénti valóságok és a szimulátor által generált pl. forgalmi adatk homogenitása jelentősen eltérhet egymástól. A generált (alternatív) adatvagyonok homogenitásra gyakorolt hatása lehet versengő és akár azonos (vö. holtverseny esélye az 5. ábra sárgával kiemelt becslés-oszlopában).

Így a szignifikancia fekete-fehérként ható fogalma helyett egy fuzzy-jellegű hasonlósági értelmezés keretében az alternatívák relatív távolságának versengő alakzatai tárhatók fel. A modellezés során nem az a kérdés, vajon szignifikáns-e két halmaz hasonlósága, hanem az, hogy lehet-e, s ha igen, milyen mértékben javítani a generált és a valós adatvagyonok hasonlóságán? S ha lehet, akkor vajon az OAM-logika és a sorrend-alapú standardizálás komplex hatásain keresztül képesek-e kockázatok felmerülni a generált adatvagyonokra alapozott modellekben?

Summa summarum: érdemi kockázatok nem azonosíthatók be a generált adatvagyonok és a valós adatvagyonok eltérései révén, ugyanis a komplex hasonlósági nézetben az azonosságon túli konstellációk egy masszív fennsíkot alkotnak, ahol egyetlen egy attribútum sem képes a többi elnyomni, s az eltérő véletlen-konstellációk eredő hatása egymástól nem képes érdemben különbözni. Nem mellesleg: maga a szimulátor-logika valós adatvagyonon eleve a valóságnak megfelelő paraméterekkel áll majd fel a jövőbeli alkalmazásokban, így a generált valóságoknak „csak” a szimulátorok előállításához szükséges inkubátor-környezetet kell garantálni tudni, ami az értelmezési intervallumokon keresztül olyan nagy mértékben adott, hogy a véletlen-konstellációknak már nincs érdemi kockázata egymáshoz és a valósághoz képest.

# Mellékletek

…hivatkozások a szövegközben…

Az ábrák nyers verziói és a reprodukálhatóság garanciái:

<https://miau.my-x.hu/miau/243/valos_vs_rnd.xlsx>