DPR-adatbázis, mint big-data értelmezési kísérletei

(Experimental hermeneutics based on the DPR-database as a big-data issue)

Pitlik László (My-X team)

Kivonat: Az egyes szakok értékét a Hallgatók aktív döntései és a munkaerő-piaci hatások közösen alakítják, ahol az érték fogalma alapvetően a versengő intézmény-időszak objektumok versenyén keresztül kaphat értelmet mindennemű szubjektivitás kizárása mellett. A szak értékének dinamikája, ill. a legjobb intézmény az itt bemutatott számítások alapján a pályaválasztási tanácsadás részévé tehető, ill. az intézményi marketing-akciók alapja lehet.

Kulcsszavak: hasonlóságelemzés, anti-diszkriminatív modellezés, pályaválasztási tanácsadás

Abstract: The value of professional fields is influenced through active decisions of Students and/or reaction of the markets. The value as such can only be interpreted in the competition of institutional units in diverse time periods in an objective way. The value has always a dynamical view, and also the best institution can be derived. Both information can be part of career-consultations.

Keywords: similarity analysis, antidiscriminative modelling, career planning

# Bevezetés

Az országos, anonim DPR-adatbázis kiadásához/átvételéhez elsőként tisztázni kell a jogi kereteket (vö. <http://miau.gau.hu/tki/dpr_adatkezelesi_nyilatkozat_2016.docx>). Ezt követően jelenleg a 2011-2012-2013-2014-es felmérési évek adatai kerülnek kiadásra, melyek a 2007-2013-as időszakot érintik. Az adatvagyon egyes fontos mezői (pl. mely évet is érinti a felmérés) üresek. Ennek mértéke (0-8%) időben csökkenő. Ezen hiányos adatsorok értelmezése csak speciális célokból lehetséges. A rekordok száma felmérési évenként:

* 20458 (2011)
* 24890 (2012)
* 24233 (2013)
* 21164 (2014)
* A 2015-ös adatok kiadása hamarosan várható.
* A 2016-os felmérések majd csak 2017 őszén válnak publikussá a kutatók számára.

Az adatmennyiség nem mutat érdemi trendek az abszolút számok alapján. Vagyis a válaszadási hajlandóság fejlődéséről vélelmezhetően nem lehet beszélni (még az összes Hallgatói létszámmal relativálva sem).

Ezen cikk célja a válaszadási hajlandóság növelése azáltal, hogy mindenki számára világossá válik, milyen fontos a tény-alapú szakpolizitálás az egyedek (vö. pályaválasztási tanácsadás) és a szervezetek (vö. stratégiai tervezés, minőségbiztosítás) szintjén is. A cikk másik célja példaként szolgálni az adatbányászati témakörök oktatásához.

Az adatokkal együtt a következő utalások kerülnek kiadásra:

A kérdőív és a módszertant is tartalmazó kutatási záró-tanulmány az alábbi linkeken érhető el:

2014:

* <https://www.felvi.hu/felsooktatasimuhely/dpr/online_kerdoivek/online_kerdoivek_2014tavasz>
* <https://www.felvi.hu/pub_bin/dload/DPR_tanulmanyok/frissdiplomasok_2014_zarotanulmany.pdf>

2013:

* <https://www.felvi.hu/felsooktatasimuhely/dpr/online_kerdoivek/online_kerdoivek_2013_tavasz>
* <https://www.felvi.hu/pub_bin/dload/DPR_tanulmanyok/Diplomas_palyakovetesi_adatok_2013_Intezmenyi_adatfelvetelek_VEGLEGES_web.pdf>

2012:

* <https://www.felvi.hu/felsooktatasimuhely/dpr/online_kerdoivek/online_kerdoivek_2012_tavaszi_felmeres>
* <https://www.felvi.hu/pub_bin/dload/DPR_zarokotetek/zarokotet_2012.pdf>

2011:

* <https://www.felvi.hu/felsooktatasimuhely/dpr/online_kerdoivek/kerdoivek_2011tavaszi_felmeres>
* <https://www.felvi.hu/pub_bin/dload/DPR_zarokotetek/zarokotet_2011.pdf>

Mint a tanulmányok alapján látható, a kiértékelés az elsődleges statisztikai értelmezésekre fókuszál. Ebben a cikkben az adatbányászat oktatásának támogatása módszertani kérdéseket helyez a középpontba. Ezen belül is az antidiszkriminatív modellezést, mely hermeneutikája triviális döntéstámogató karakterrel bír.

# Az antidiszkriminatív modellezés lehetőségeiről

A következőkben két, egymásra épülő modell kerül bemutatásra. A számítások reprodukálását támogató háttér XLS-állomány sajnos az nyers adatok védelme miatt nem adható ki. Így a cikkben csak egy fajta standardizálás utáni nézetektől lehet követni tételesen a modellezés lépéseit.

A kettős cél a következő:

* kiválasztva (véletlenszerűen) egy szakot (vö. gazdasági és vidékfejlesztési agrármérnök) megvizsgálásra kerül, vajon mely felsőoktatási intézményekhez évente milyen idealitás-indexek rendelhetők hozzá a diploma értékét befolyásolni képes kérdőív-kérdések (jelenségek) alapján annak érdekében, hogy magának a szaknak fejlődésének/idealitásának az időbeli dinamikáját meg lehessen határozni – s ez által a pályaválasztási tanácsadásba a szakról a végzettek által idősorosan alkotott véleményt orientáló jelleggel fel lehessen kínálni (hiszen ha egy szak piaci értékét sejtető idealitás-index idősorosan csökkenő tendenciát mutat, akkor annak választása nagyobb kockázat, mint egy felszálló ágat választani – azzal a kiegészítéssel, amit bárki pl. a sertés-ciklus megismerése kapcsán ezen értelmezéshez saját intuíciói alapján még hozzá tud tenni – vagyis a trendek potenciális hullámzását is figyelembe véve…)
* az idealitás-index és az ehhez kapcsolódó egyéb modell-robosztussági adatok alapján a legjobb intézmény levezetésével ismét csak operatív segítséget kap a pályaválasztás előtt álló diák (s nem mellesleg a felsőoktatási marketing és az országos stratégiai irányítás minden érintettje is)…

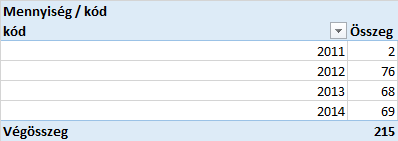
Az antidiszkriminációs modellezés alapját a hasonlóságelemzés adja (vö. <https://hu.wikipedia.org/wiki/Hasonl%C3%B3s%C3%A1gelemz%C3%A9s>). Az antidiszkriminációs modellek lényege, hogy egy optimalizálás keretében a kiindulási objektum-attribútum-mátrix minden egyes attribútuma kapcsán irányokat kell meghatároznia tudni az elemzőnek, vagyis vagy növeli egy adott jelenség nagyobb adata a diploma értékét (vö. egyenes arányosság), vagy nem (vö. fordított arányosság). Az irányról józan logikai megfontolások alapján lehet szakértői szinten dönteni. pl. annál nagyobb egy diploma értéke, minél többen választják az adott képzés más korábbi képzések után és minél kevesebben akarnak ezt követően vagy ezzel párhuzamosan tanulni – mert hisznek benne a Hallgatók… Az objektumok ilyen esetekben intézmények adott időszakai lehetnek.

Az antidiszkriminációs modellezés azon optimalizált súlyokat keresi, melyek mentén minden intézmény minden időszaka az összes attribútum alapján a minden objektum másként egyforma elvet leginkább képes megközelíteni.

Ha ez az másként azonosság nem bizonyítható, akkor a jobbak tényleg jobbak, mint a gyengébb idealitás-indexértékkel rendelkezők,. Ez és nem valamiféle misztikus és önkényes súlyozás és pontozás adja az erkölcsi és szakmai alapot a rangsorolás felvállalására. Ez az adatvezérelt modellezés ráadásul képes saját értelmezési intervallumának határait is felismerni a függvényszimmetriák kiaknázása révén.

# Az adatvagyonról

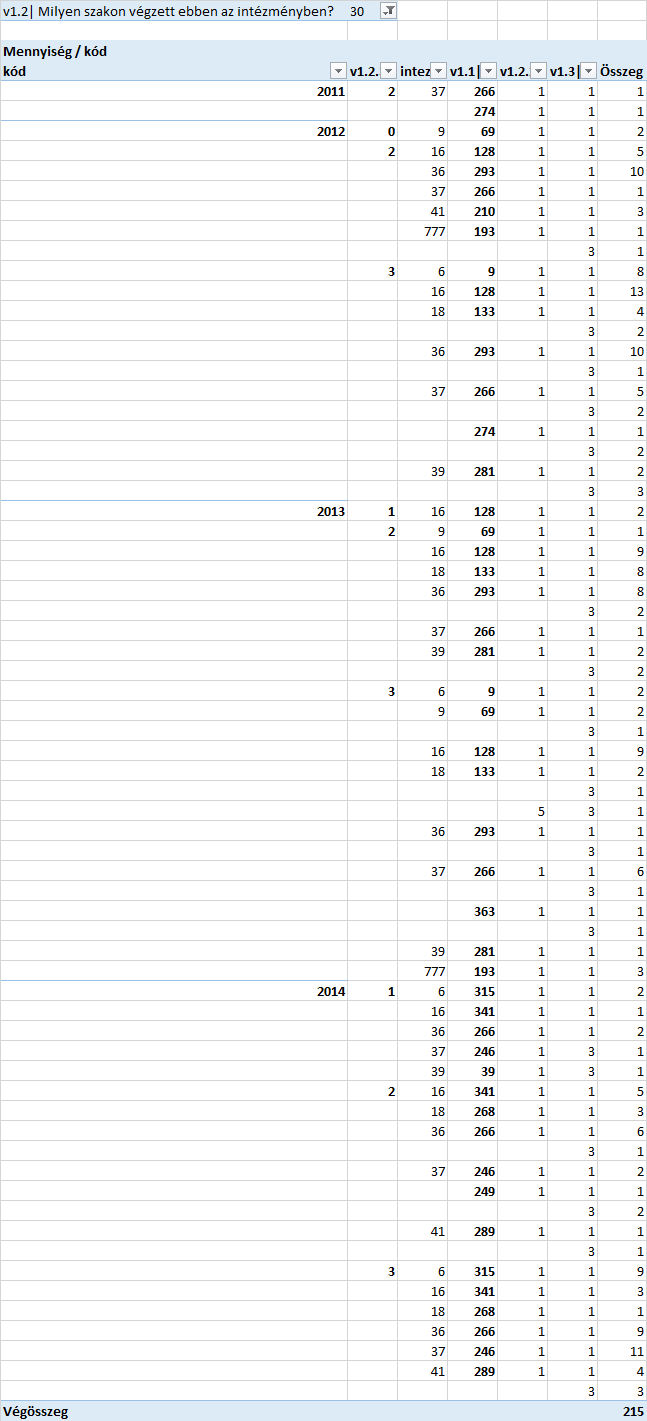
Az egyes vizsgálati években az alábbi rekordszám felelt meg a kiválasztott szak alapján való szűrésnek:

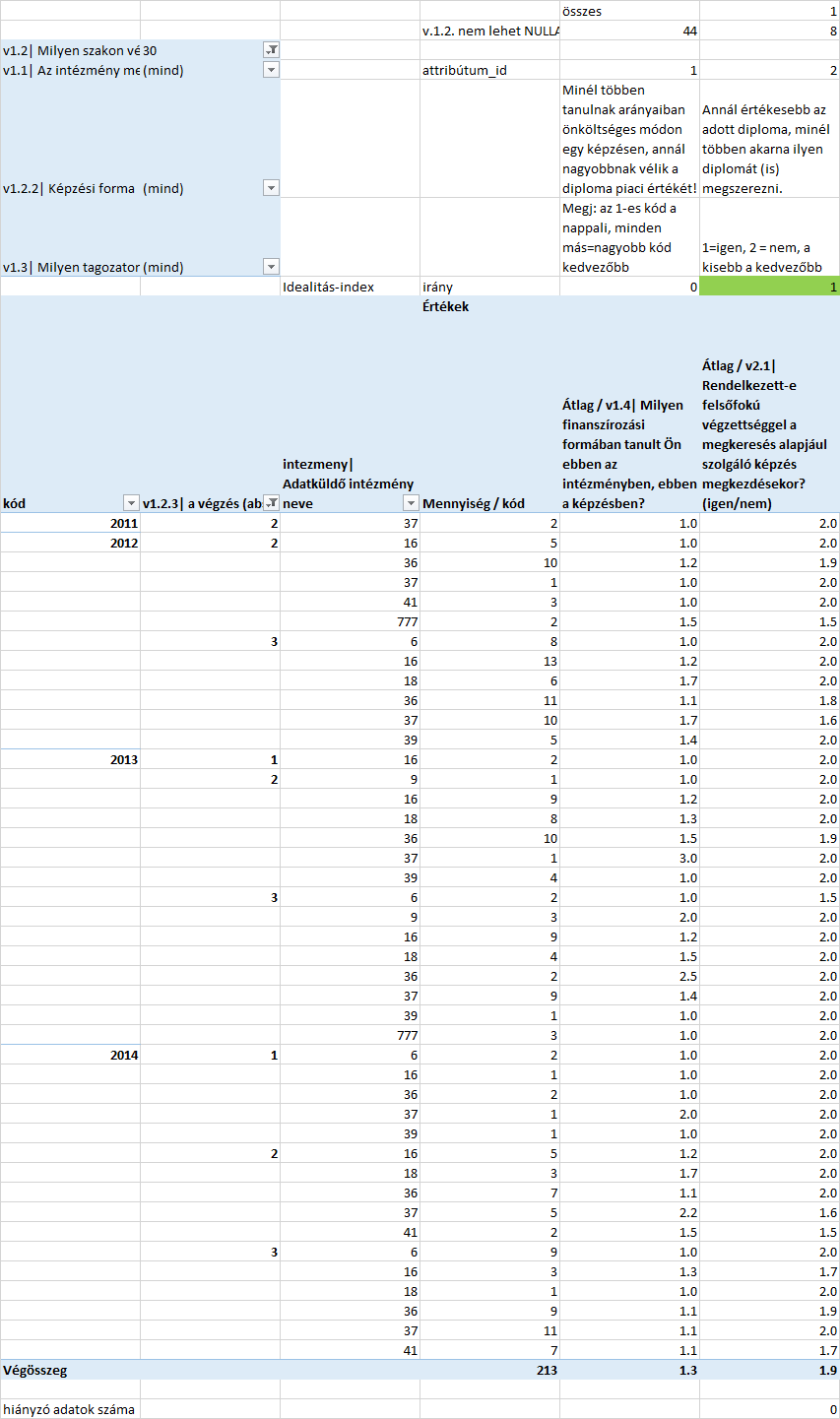


Mivel két rekord esetén éppen az idő-koordináták hiányoztak, így végül 213 rekordnyi adat került feldolgozásra. Ez természetesen nem sok, de jelen cikk célja a módszertan bemutatása, mely természetesen több adat alapján hitelesebb eredményekre vezethet. S ennek jeleit a függvényszimmetria vizsgálatok jelezik is az eredmények fejezetben.

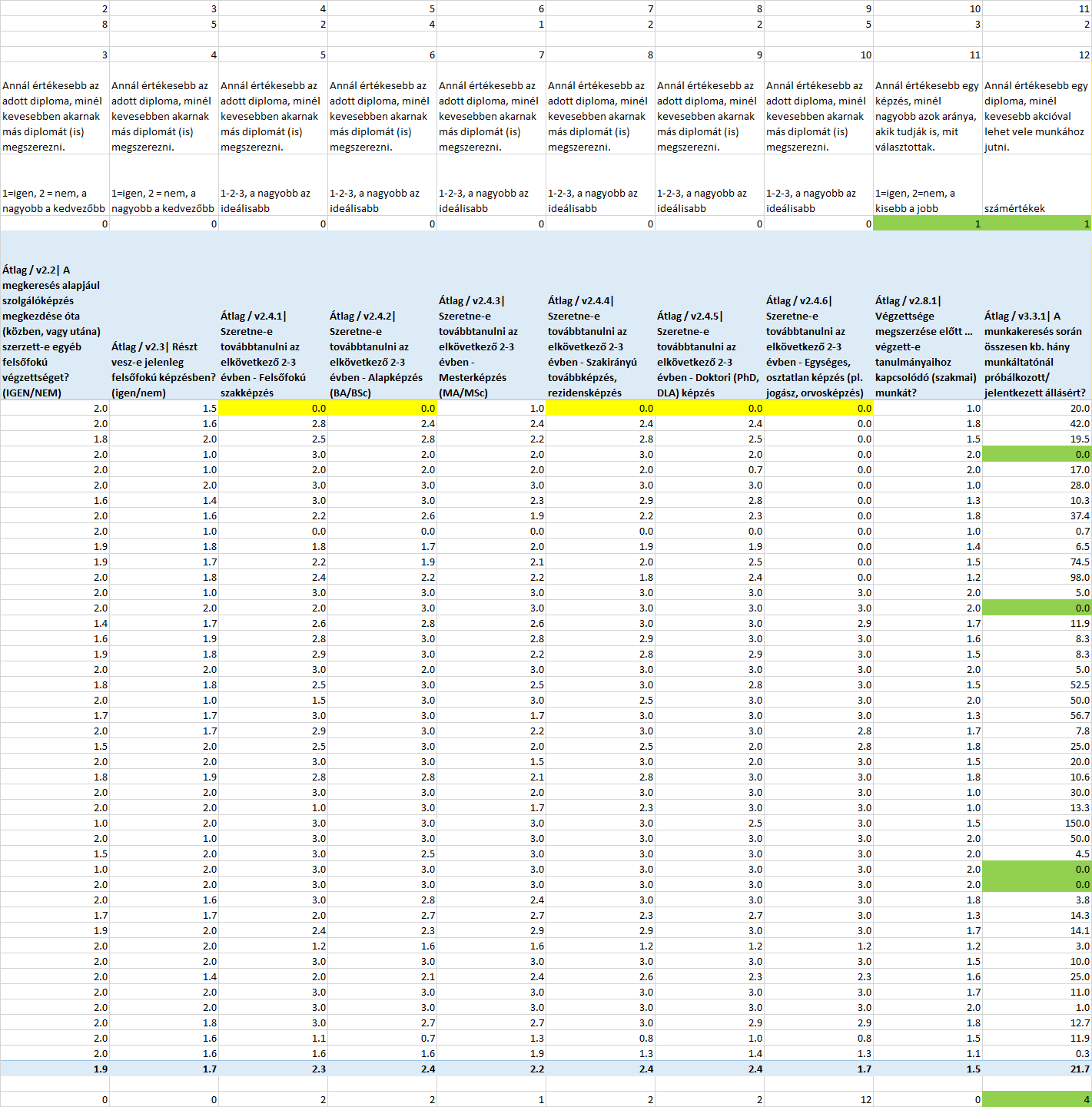
A következő pivot-riport a teljes kombinatorikai térre való felbontást jelzi, ahol az intézmény, a kar, a képzés jellege mind-mind új objektumot jelentenek évente.

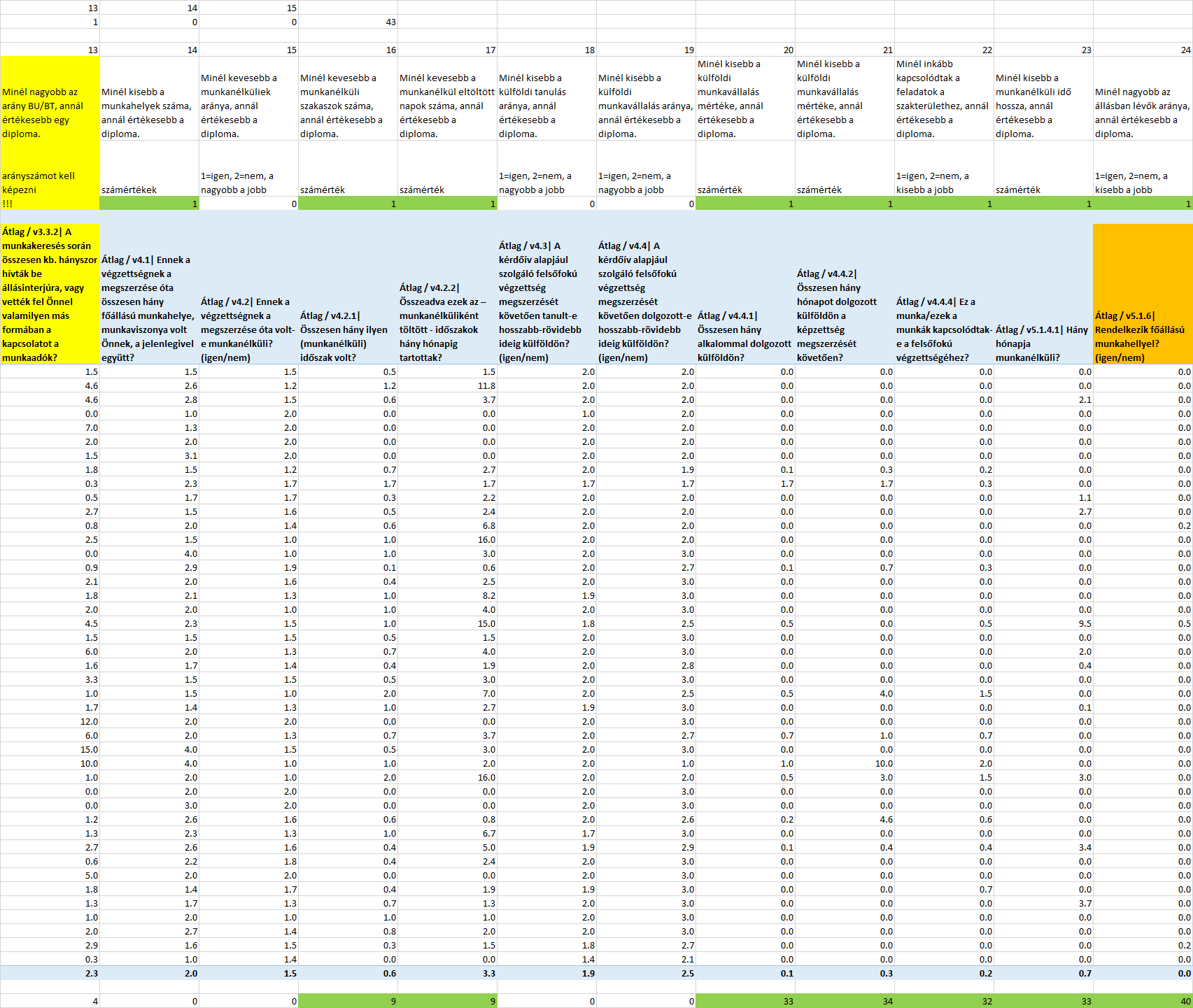
Ez a nézet sajnos túl sok 1-2 rekordos mintához vezetett, így végső soron csak az intézményi bontás került alkalmazásra. Ezáltal pl. a nappali és levelezős képzések között nem lehetett különbséget tenni a későbbiekben:





A 213 adatsor 43 intézmény-időszak objektumra bomlik fel. Az adathiányok azon esetekben jelentenek elemzési kockázatot, ha az adathiányt kódoló nulla érték irányába alakul ki preferencia, mert az, hogy egyes jelenségekről nincs tudatosan felvállalt adatközlés, még nem jelenti ezek tényleges hiányát. Szerencsére a fordított arányosság szerint értelmezendő változók jelentős része esetén nem volt érdemi a hiányzó adatok száma.

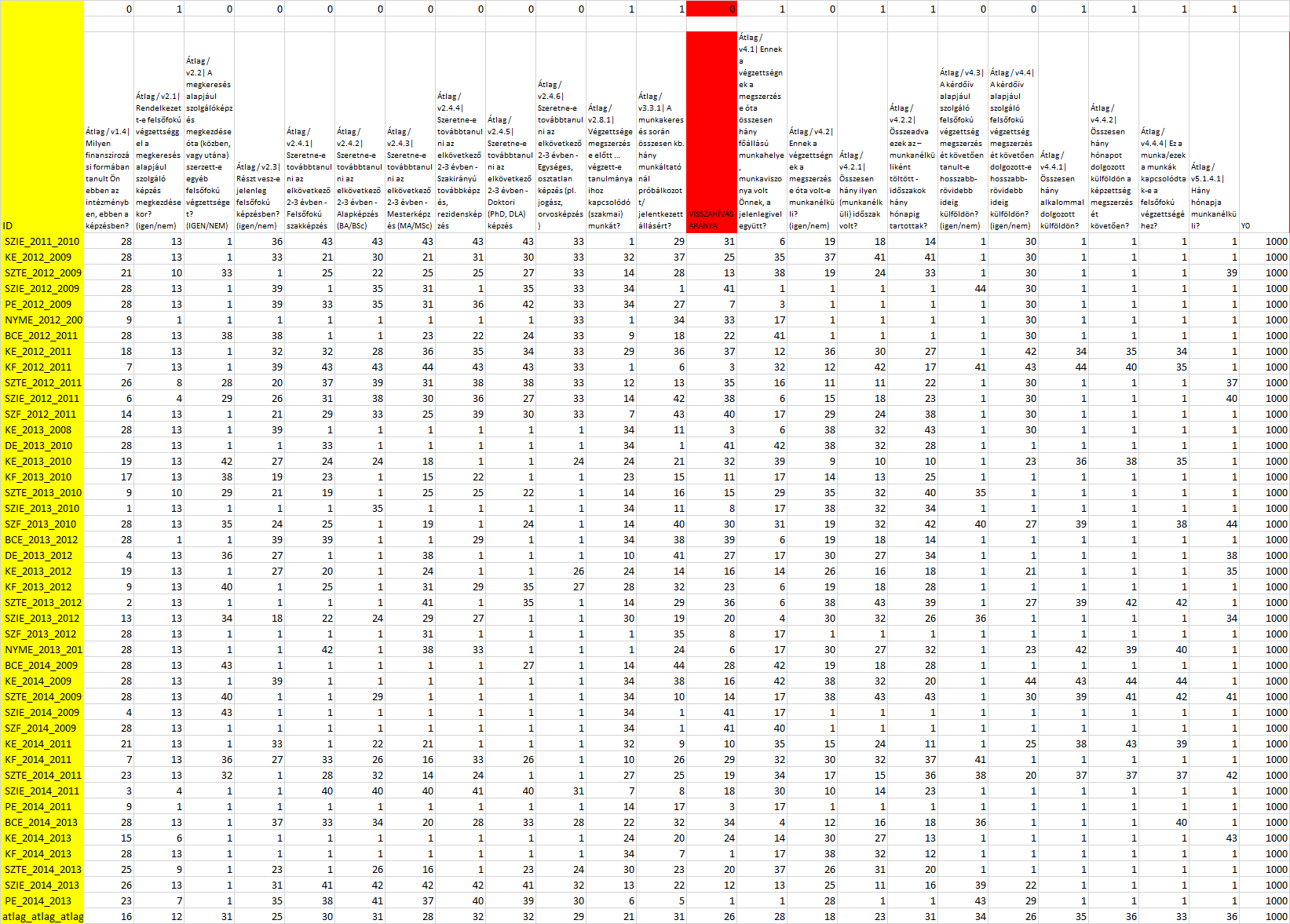




A citromsárga jel a nyers kérdőív adatok alapján képzendő mutatószámokra utal (vagyis jelen esetben a munkakeresések számához mért munkáltatói visszajelzések arányára). A zöld jel az adathiányokra mutat rá. Míg a narancssárga jel a masszív adathiány miatti jelenség-elhagyás kényszerét jelzi.

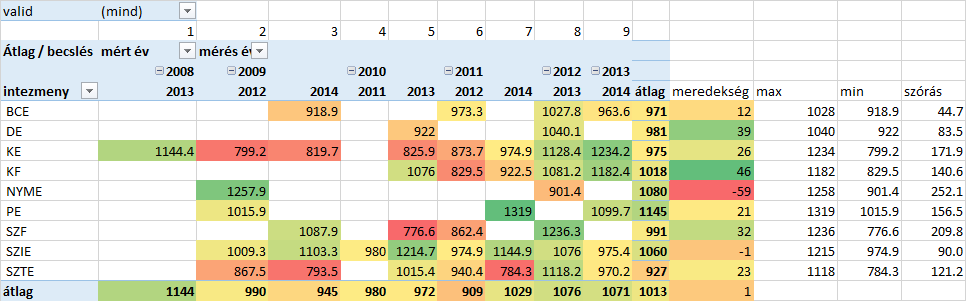
Mindösszesen tehát 23 mutatószám állt rendelkezésre, ebből 22 primer, 1 származtatott. Az irányok szöveges értelmezését a fenti táblázatok mutatják be.

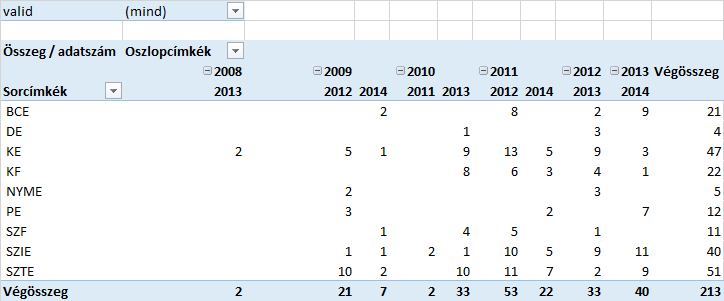
Az egy tizedesjegy pontossággal megadott aggregált adatok már nem esnek a nyersadatokra vonatkozó publikációs tilalom alá. De az elsődleges standardizálást az attribútumonkénti sorszámokból álló OAM jelenti, ahol a sorfejlécen látható a rekordok száma és az intézmény-időszak kód, ill. a származtatott attribútumra való (piros) utalás. A függő változó (Y0) egy fiktív konstans, mely értékeit kell a becsléssel minél jobban megközelíteni az optimalizáció keretében. Az idealitás index, vagyis a diplomákról alkotott komplex (többdimenziós) kép tehát számszerűen az 1000-es normaérték körül ingadozik majd.



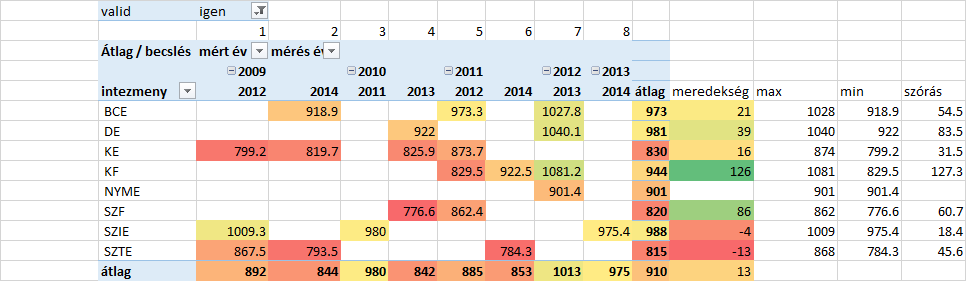
# Eredmények

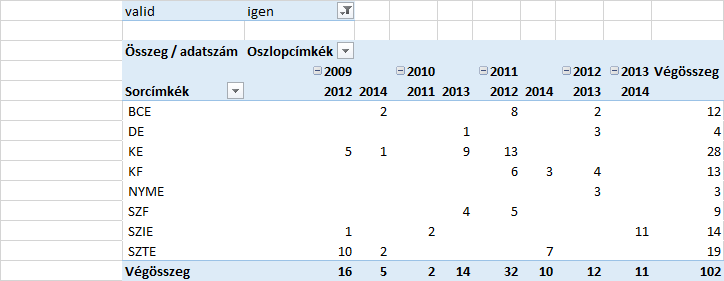
A modellezés eredményei tehát az 1000 jóságpont körül ingadozó idealitás-indexértékek, melyek intézményi és időbontás szerint az összes és a valid adatokra az alábbiak szerint alakulnak:



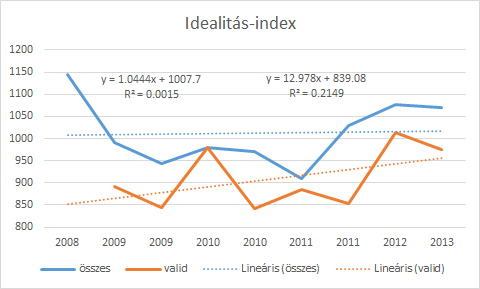


Valid egy becslés akkor, ha egy inverz irányokkal futtatott (tükörkép) modellben a becslési hibák előjele ellentétes a direkt becsléssel. A 43 objektumból 22 becslés invalid volt, ami a masszív adathiányokra, ill. a kevés adatrekordra utal félreérthetetlenül.

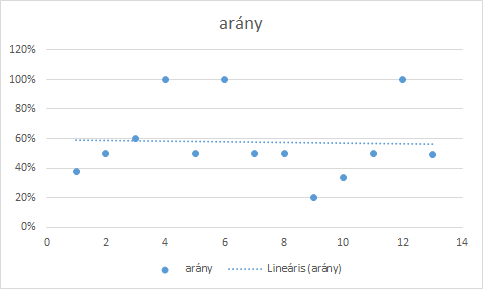
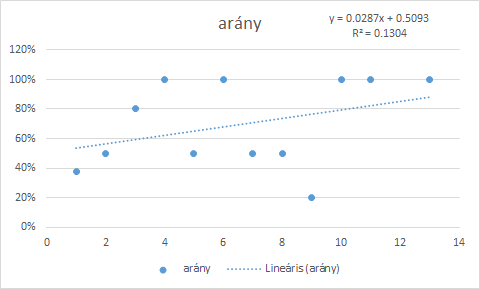




Az első modell összegző eredménye az idealitás-index idősoros nézete az összes becslés és a valid becslések alapján:

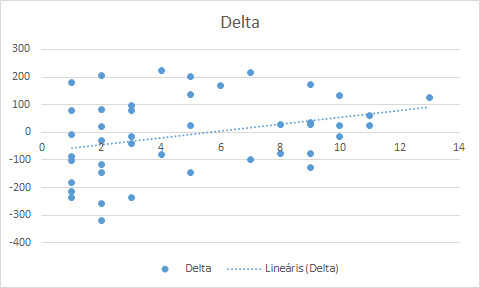
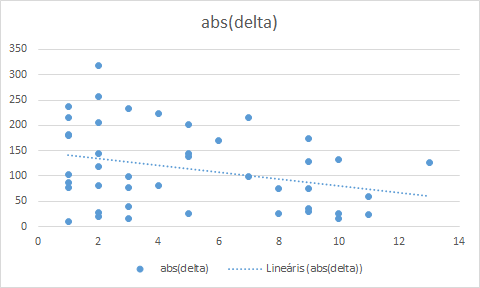


Mint látható, az összes adat alapján nincs az adott szaknak fejlődése, vagyis egy stagnáló értékű szakról beszélhetünk – nagy amlitúdók, alacsony R2 mellett. De a valid adatok már érzékelhető R2 mellett masszív növekedést jeleznek országos átlagban. Sajnos azonban a valid adatokhoz tartozó becslések egy kivételével alacsonyabbak, mint a mindösszesen minta becslései, sőt, éppen csak hogy átlépik a normát jelentő 1000 pontos küszöböt, vagyis pl. az egyedi, kismintás válaszok jobb képet festenek, mint a robosztusabb minták.

A csak a kerekítési hibát konszolidáló bal oldali és a minimális becslési hibát (<+/-10 egység) is konszolidáló jobb oldali ábra világosan jelzi, hogy az egyes objektumokhoz tartozó adatok számával (X-tengely) arányosan nő a validitás aránya (Y-tengely).

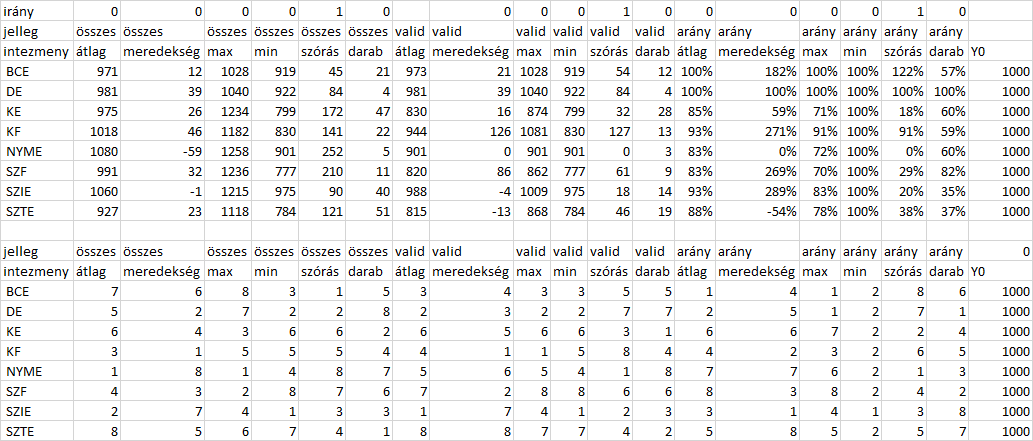
A modellhiba és az egyes objektumok mögötti megfigyelések számossága között az alábbi összefüggések tárhatók fel:

Mint látható, az előjeles tény-becslés=delta eltérések a nagyobb számosság kapcsán (X-tengely) gyenge alulbecslési hajlamot mutatnak (tény > becslés). Az abszolút (előjeltelenített) becslési hiba (Y-tengely) azonban a számosság növekedésével tendenciájában az elvárásoknak megfelelően csökken.

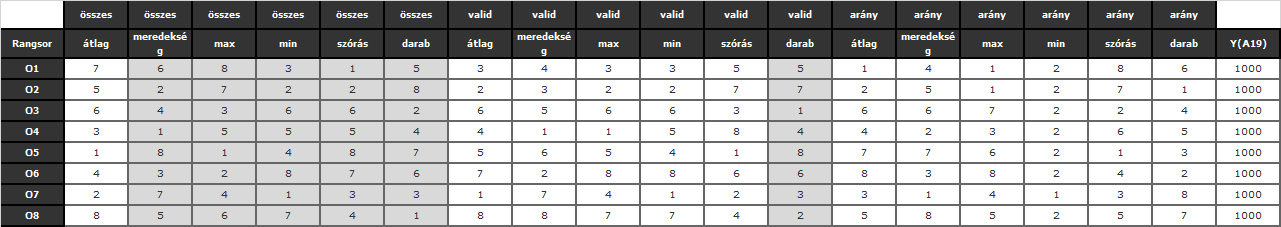
\*\*\*

Látszólag egy másik modell inputja, de formálisan az eredmények egy rétege a második modellezési célt támogató OAM:

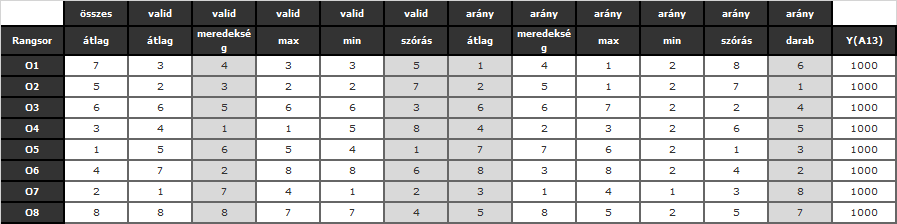


A PE-objektum validitás hiányában kiesett az értékelésből. A fenti OAM célja a legjobb intézmény levezetése antidiszkriminatív modellezés keretében, ahol annál ideálisabb egy intézmény, minél nagyobb idealitás-indexre vonatkozó becslésekkel rendelkezik a mindösszesen és a valid rétegekben egyaránt (ink. átlag, max, min, meredekség). S mely esetében az idősor szórása minél kisebb. A valid és a mindösszesen adatok arányát tekintve is az a jobb egy intézménynek, ha a valid adatai minél jobbak a mindösszesen halmazhoz mérten.

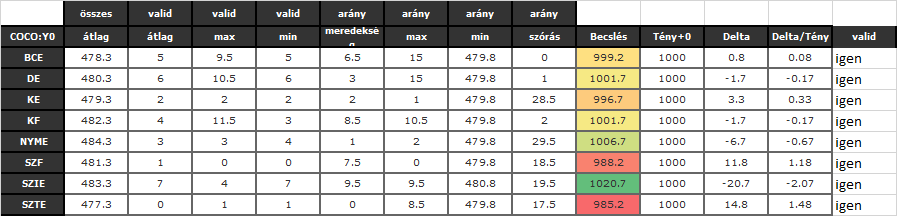
A fenti inputadatokból kiindulva két lépésben is igaz volt, hogy akár minden intézmény lehetne másként egyformán értékes az adott szak DPR-adatai alapján. Azonban ezek a modellek mindenkor csak az adatok egy részét használták fel:



1. lépés



1. lépés



1. lépés

A fennmaradt változók kapcsán fény derült arra, hogy a SZIE az az intézmény, melyet a pályaválasztási tanácsadás keretében jelen adatok alapján ajánlani illene egy független szakértőnek az érdeklődő diák számára a szak potenciális értékfejlődését fentebb már vizualizálva.

\*\*\*

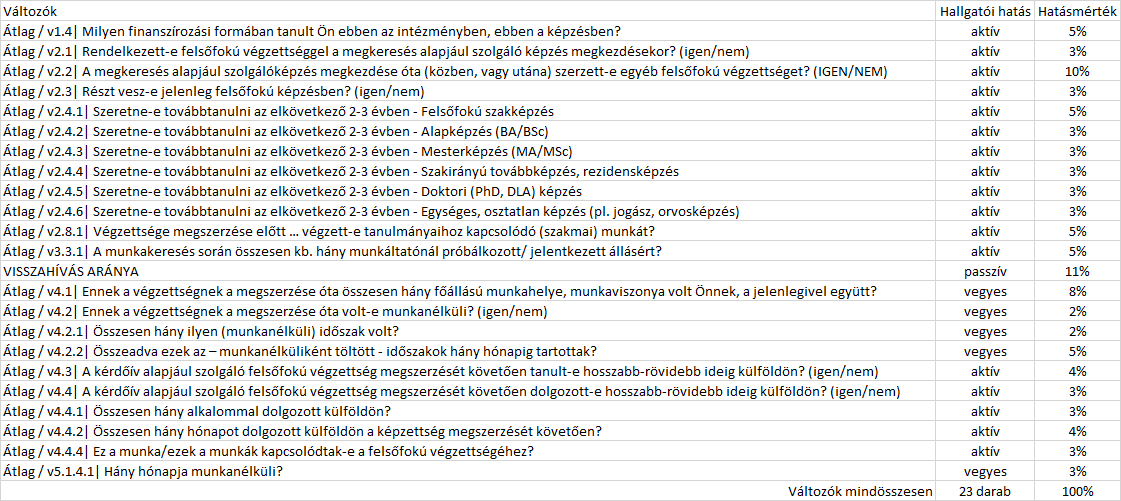
Az idealitás-index bemeneti jelei kettős természetűek:

* egyes változókról alakulásáról a Hallgató dönt (vö. korábbi, párhuzamos, későbbi tanulmányok)
* míg más változók esetében nem a Hallgatón múlik ezek alakulása (pl. visszajelzések aránya a munkáltatók által munkakeresés során),
* vagy a munkaerőpiac és a Hallgató közösen felelnek az adott jelenség alakulásáért (vö. munkanélküliségi adatok).

Mint az az alábbi táblázatból látható: a Hallgatói döntésen múló jelenségek túlsúlya érhető tetten a 23 változón belül. Ennek ellenére, vagyis a komplex döntési erőterek miatt a modell-hermeneutika is több rétegűvé válik:

* Ha csak a Hallgatói döntéssel befolyásolható (aktív) változók lennének egy modellben, akkor az idealitás-index értelmezése azt emelné ki, hogy mely intézmény mely időszakában viselkedtek a Hallgatók „tudatosan” úgy, hogy ezzel a diploma értékességét (f)elismerték.
* Ha csak a passzív (nem a Hallgató által befolyásolt) tényezők lennének egy modellben, akkor a külvilág (munkaerőpiac) viselkedése (értékítélete) lenne tetten érhető intézményenként és időszakonként az idealitás-index által.
* Ha ezek a hatások magukban a változókban (vö. vegyes hatásmechanizmus) és/vagy a passzív és aktív változók keveredésén keresztül válnak komplexé, akkor általában véve lehet arról beszélni, az adott szakot érintően térben és időben hol mutatkozik a diploma értékességét a háttérben tudatosan és/vagy spontán elismerő erőterek eredőjeként olyan hatás, mely alapján intézmények és/vagy időszakok bontásában az diplomát érintő érték fogalma nem homogén.

Az egyik legnagyobb hatásmértékű változó a minden intézmény\_időszak objektumok azonos értékűségét elvárva az optimalizálás során: éppen az egyetlen passzív változó (11%). De előkelő helyen van aktív (10%) és vegyes (8%) jellegű hatásmechanizmus is. A 100%-os becslési tér 23 változó esetén 4-5%-os átlagos hatásmértékét jelentene. A 8-11% ehhez képes minősül jelentősnek.



Egy antidiszkriminatív modell legnagyobb hatású változói azok, mely rendszerszintű felülsúlyozása kompenzálni igyekszik a többi karakterisztika által a differenciálódás irányába mutató erőtereket. S fordítva is igaz: a kis hatás-mértékek ezen változók elhagyása irányába mutatnak, mert ezek mentén azonos értékűségről beszélni lehetetlen lenne. Ha az azonos értékűséget lehetővé tevő és ezt kizáró erőterek eredőjeként nem áll be a minden objektum másként egyforma elvnek megfelelő állapot, akkor azok az objektumok, melyek több jóságpontot gyűjtenek, vitán felülállóan jobbak, ill. fordítva.

\*\*\*

Speciális részeredményként az átlagos szak karakterisztikája is feldolgozásra került a valós intézmény-időszak konstellációk mellett. Ezen átlagos objektumnak elvileg a norma (1000 pont) körül illik ingadoznia, kivéve, mint jelen esetben is, ha a vizsgált konstellációk inhomogének, vagyis olyan antagoizmus-gyanúk állnak fenn, melyek azt sejtetik, hogy egyes változók egyes tér-idő-egységekben nagyon jók, míg más változók szinte kényszerűen nagyon gyengék. Így az átlag objektum, mely semmiben nem elég jó, nem tud kellő jóságponthoz jutni és ezáltal a modell leggyengébb (737) idealitás-indexét produkálja. Ezzel is jelezve: ilyen élethelyzet nem ismert, ez a konstelláció kívül esik az ismert univerzumon. A modell genetikai potenciálja, vagyis annak a konstellációnak az idealitásértéke, mely minden változó kapcsán a legjobb állapotokat tükrözi, 1665 pontot mutatna, mely érték lényegesen magasabb, mint a modell legmagasabb idealitás-értéke (1319, ill. csak valid esetekben: 1081). Ez is annak jele, hogy az ideális állapot maximumát jelentő élethelyzet nem valószínű, hogy létezik. Ezek az instabilitási jelek a magas invalid arány mellett azt sejtetik, hogy ezen modellre még nem illik szimulációt (mi-lenne-ha-alapú döntéselőkészítést) alapozni.

# Következtetések

Az itt bemutatott módszertan választ adott a célként kitűzött elemzési feladatok kapcsán megfogalmazott kérdésekre. A módszertan bármely szak esetén (automatizálhatóan) beépíthető egy oktatási vezetői/diák/hallgatói információs rendszerbe. Az adatbázis számos további jelenségről is tartalmaz adatot, de az adatvagyon konszolidált állapota nem létezik, s viszonylag kevés transzformációval ezen adatok, s ezen belül is az itt iránnyal jellemzett adatok voltak azok, melyekre elemzést lehetett alapozni.

Az eddig, DPR-adatok felhasználásával készült modellek katalógusa nem ismert, de hasznos lenne ezeket az adatvagyon mellett nyilvántartani, hiszen a modellek iránti igény a DPR módszertan szerves részeként merült fel a korábban is (vö. <http://www.tka.hu/docs/palyazatok/dpr_kutatas.pdf>, <https://www.felvi.hu/pub_bin/dload/DPR_workshop_2013_0704/DPR_AVIR_20130704_fsz_vegleges.pptx>).

# Jövőkép

Ha valaki objektumként szakokat választana ki, akkor az átlagos objektumértékek kerülnének elsődlegesen összevetésre. Azok az átlagos értéksorok, melyekről a modellezés belátta, hogy ilyen formában nem léteznek konstellációk, ahogy az abszolút ideál sem tűnik létezőnek. Ennek ellenére, ha minden szak minden adata hasonló transzformáción esik keresztül, akkor ezen átlagos profilok egymáshoz képest hasonló modell-lépéseken keresztül racionális szak-értékelési rangsorokhoz fognak vezetni (ahol az értelmezés kapcsán a Hallgatói aktív, passzív és vegyes erőterek finomhangoló hatásáról ismét csak nem illik megfeledkezni)…

(Folytatás hamarosan: MIAÚ No. 219 – II. rész)

# Olvasói vélemények

Az MTMT- és a DPR-adatok elemzése egyes szakértők szerint az egyetemi vezetői információs rendszerek, minőségbiztosítási rendszerek szerves részeként kellene, hogy értelmezhető legyen – pl. az itt bemutatott módszertanok alapján is…