



TUDOMÁNYOS DIÁKKÖRI DOLGOZAT

**ÖNSZERVEZŐ TÉRKÉP (SOM)
HASZNÁLATA GAZDASÁGI
ELEMZÉSEKBEN**

Szerző(k):

Majdán Márk

műszaki menedzser BSc. szak, II. évf.

Nagy Balázs

műszaki menedzser BSc. szak, II. évf.

Konzulens(ek):

Dr. Nagyné Dr. Hajnal Éva

egyetemi docens

Dr. Varga Valéria

főiskolai docens

Székesfehérvár, 2013.

Tartalomjegyzék

1. Bevezetés	1
2. A modell bemutatása	1
2.1. A neurális hálók fogalma, fajtái	1
2.2. A Kohonen-féle önszervező térképek és működésük	3
2.3. A SOM implementációja a MatLab programozási nyelvben	7
2.4. Önszervező térképek használata előrejelzésre	8
2.5. A tárgyalt probléma gazdasági háttere	9
2.5.1. A kutatás-fejlesztési szektor nemzetgazdasági jelentősége	9
2.5.2. A felsőoktatás, mint a K+F szektor utánpótlási bázisa	11
3. Adatok és módszerek	13
3.1. Adatok kiválasztása, feldolgozása	13
3.2. A modell felállításának lépései	14
4. Eredmények és értékelésük	16
5. Összefoglalás	27
6. Irodalomjegyzék	28
7. Függelék	30
7.1. A 2001-2012-ig rendelkezésre álló adatokból készült SOM programkódja	30
7.2. A 2001-2012-ig rendelkezésre álló adatokból készült SOM programkódja valamint a becsléshez szükséges számítások	31
7.3. A súlypontok és BMU-k felrajzolásának programkódja	34
7.4. A már kigyűjtött, a becslések elvégzéséhez rendelkezésre álló adatok	35
7.5. Becsült adatok	36
7.6. Első forgatókönyv	38
7.7. Második forgatókönyv	39

1. Bevezetés

A kutatás-fejlesztési szektor a nemzetgazdaságban betöltött szerepe miatt figyelemre méltó gazdasági tényező, amely meghatározza egy adott ország versenyképességét. Az ágazat sok mérőszámmal jellemezhető, ezek között érdemes a felsőoktatás arányainak alakulását vizsgálni, hiszen a kutatói utánpótlást a felsőoktatási intézmények biztosítják. Ezen megközelítés további előnye, hogy a felsőoktatási statisztikák könnyen hozzáférhetők és sokféle bontásban szolgáltatnak adatokat.

Az egyetemi (és főiskolai) hallgatói létszámok alakulását befolyásolják az oktatáspolitikai átalakítások, és a demográfiai mutatók. Magyarországon a csökkenő demográfiai tendenciák csökkenő hallgatói létszámokat eredményeznek. Ennek pontosabb becslésére azonban megfelelő módszert kell alkalmaznunk.

Az általunk használt önszervező térkép alkalmas az adatok közötti kapcsolat elemzésére, ám jellemzően nem előrejelzésre és becslésre használatos. Jelen dolgozatban szeretnénk bemutatni ezt az alkalmazási területet is, egy olyan statisztikai modell felállításával, amely képes jó közelítéssel megbecsülni a felsőoktatási hallgatói létszámok alakulását a demográfiai mutatókból kiindulva.

2. A modell bemutatása

Dolgozatunkban a mesterséges neurális hálózatok egy változatát, a Kohonen-féle önszervező térképet választottuk statisztikai adatelemző eszköznak. A mesterséges neurális hálózatok olyan matematikai-informatikai modellek, amelyek az emberi agytanulási képességét modellezik (ez a megállapítás különösen a korai változatokra igaz). Történetük közel hatvan éves múltat tekint vissza. Az alábbiakban az alapfogalmak tisztázása végett – a teljesség igénye nélkül – áttekintjük főbb típusaikat.

2.1. A neurális hálók fogalma, fajtái

A legkorábbi mesterséges neurális hálózatok (pl. McCulloch-Pitts-modell) a lehető legpontosabban követték a biológia területéről kölcsönzött idegrendszer-analógiát, ezért az alapvető kifejezések többsége is innen származik, többek között maga a mesterséges

neurális hálózat elnevezés. Felépítésük röviden a következő: mesterséges neuronok (tkp. összetett függvények) egymással kapcsolatban álló sokaságából áll, amelyek hálózati topológiát alakítanak ki. A mesterséges neuronok közötti hozzárendelés (függvény) határozza meg, hogy a rendszer a bemeneti információt milyen kimenetűvé alakítja.

A mesterséges neurális hálózatok fő tulajdonsága a gépi tanulás, amelyet különböző tanulási paradigmák használatával valósítanak meg. Ezek alapján a következő típusokba sorolhatjuk őket:

1. Felügyelt tanulás: ebben a típusban ismert az adatfeldolgozás kimenete, a tanítás során azokat a függvényeket választjuk ki, amelyek a legjobban megközelítik a kívánt eredményt. Ezt a típust gyakran használják mintázat felismerésre és függvényapproximációra.
2. Nem felügyelt tanulás: az adatfeldolgozás kimenete nem ismert, a feladat az, hogy a neurális hálózat az adott mintát optimális módon dolgozza fel. Jellemzően klaszterezésre, statisztikai eloszlások becslésére, ill. adattömörítésre használatos.
3. Megerősítéssel tanulás: itt a bemeneti minta nem ismert. A hálózat többszöri próbálkozás és szelektív mintavételezés útján viszi be a megfelelő adatokat, a cél itt is az optimális feldolgozási algoritmus megtalálása.

A gépi tanulás alapját a tanulási algoritmusok jelentik, a tanulás során a hálózat típusának megfelelően ebből választ ki egy optimálisat (az optimális szót a lineáris programozás, más néven lineáris optimalizáció területén használatos értelemben használjuk).

A neurális hálózatok azonban egyéb tulajdonságokkal is bírnak, amelyek indokoltá tehetik a használatukat. A statisztikai adatfeldolgozó eljárások többségében a rendelkezésre álló adatokat csak két dimenzióban vizsgáljuk. Ez azt jelenti, hogy egy függő változónk van, amely egyetlen független változó különböző pontjaiban vesz fel értékeket, pl. egy gazdasági mutató változása az évek során. A két értéksort könnyen ábrázolhatjuk egy derékszögű koordináta-rendszer x és y tengelyén, így megállapíthatjuk a köztük lévő lineáris, vagy egyéb összefüggést.

Bizonyos esetekben előfordulhat, hogy szeretnénk egyszerre több változó közti komplex kapcsolatot ábrázolni. Más szóval, ekkor egy többdimenziós adathalmazt szeretnénk vizsgálni. Ennek a geometriai megjelenítése nehéz, hiszen három dimenziónál többet nem vagyunk képesek elképzelni.

Erre a problémára kínálnak megoldást a neurális hálózatok. Ezek olyan dimenziószám csökkentő modellek, amely képesek többdimenziós adatfeldolgozásra. E fontos jellemzőn kívül azonban talán ugyanolyan jelentős az, hogy képesek több adatsor közti nemlineáris összefüggések egyidejű vizsgálatára.

A mesterséges neurális hálózatok gyakorlati alkalmazása igen széleskörű. Használják többek között matematikai problémák megoldására, valamint statisztikai adatfeldolgozásra, különösen nehezen kezelhető, nagy mennyiségű adathalmazok esetén. A mintázatfelismerő neurális hálózatokat arc- és karakterfelismerő szoftverekben alkalmazzák [1] [2] [3], de segíti a robotok valós környezetben való tájékozódását is[4](útvonal- és akadályfelismerés). Az orvosi és élettudományokban az idegrendszerek működését modellezzik mesterséges neurális hálózatokkal.

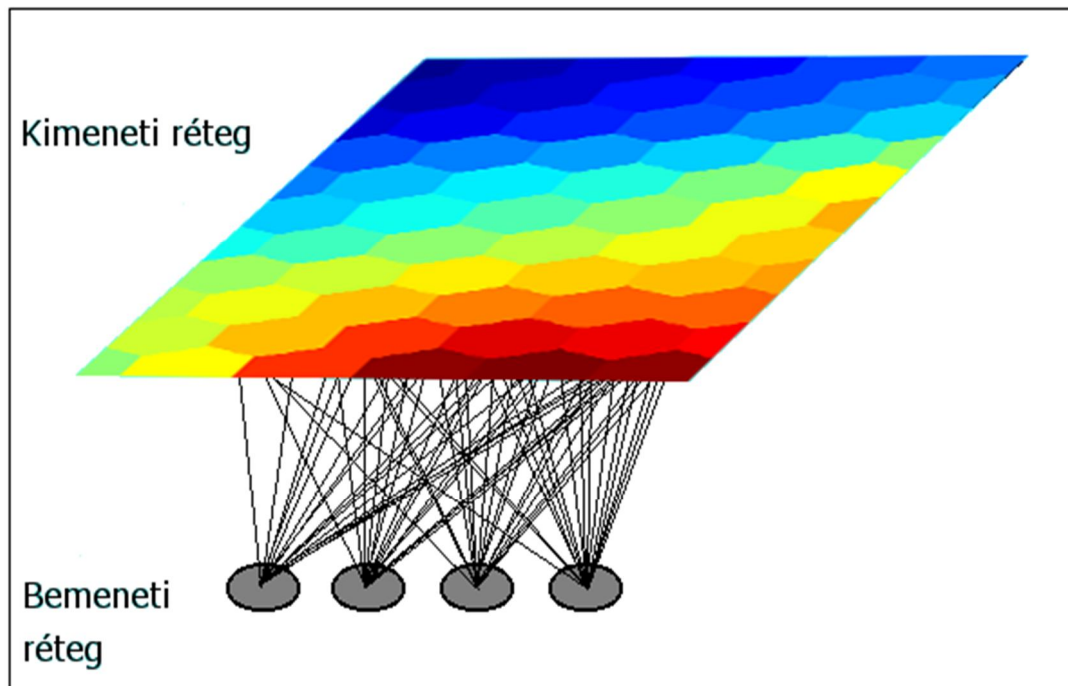
A közgazdaságban is elterjedt az alkalmazásuk, melynek alapja, hogy a gazdasági adatok közötti összefüggéseket felismerik.

2.2.A Kohonen-féle önszervező térképek és működésük

Az általunk választott mesterséges neurális hálózat a Kohonen-féle önszervező térkép (rövidítve gyakran SOM, self-organizing map), melynek megalkotása TeuvoKohonen nevéhez fűződik[5].

Ez a mesterséges neurális hálózat a nem felügyelt tanítású hálózatok közé tartozik, ez azt jelenti, hogy a hálózat a kimenetet illetően nem kap semmiféle megerősítést, csak a bemeneti minta adott. Minden mesterséges neurális hálózatban a hálózati topológiát tekintve a neuronok vezetékei legalább két rétegben helyezkednek el: egy bemeneti és egy kimeneti. Vannak olyan hálózatok is, amelyekben a bemeneti és a kimeneti réteg között egy harmadik, rejtett (köztes) réteg is van, azonban a Kohonen-hálók csak egy neuronréteg vezetékeinek a két alapvető rétegét tartalmazzák. A kimeneti réteg kétdimenziós, térképszerű elrendezésű, innen származik az elnevezése. A kimeneti réteg neuronjai

egymással kapcsolatban állnak, és jellemző elrendezésben helyezkednek el (ez lehet hexagonális, vagy négyzetes rács).



*1. ábra: az önszervező térképek felépítése. J.M. Marín
(halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema5dm.pdf) nyomán*

A bemeneti jelek a kimeneti réteg bizonyos területének neuronjait aktiválják, amennyiben mintázatuk hasonló. Ahhoz, hogy az aktiváció síkbeli kiterjedését optimalizálni lehessen, fontos ismerni a neuronok kapcsolódási mintázatát. Az optimalizációs folyamat a bemeneti jelek topológiai térképét készíti el, amelyben a bemeneti jelek közti hasonlóságokat a neuronok közti szomszédsági mutatóvá alakítja. Ez azt eredményezi, hogy a térkép képes elhanyagolni a bemeneti jelek közti zavaró tényezőket, mindössze a legjobban felismerhető hasonlóságokat (értsd: összefüggéseket) fogja figyelembe venni. Összefoglalva tehát a Kohonen-féle önszervező térkép a bemeneti jeleket (adatokat) hasonlítja össze, majd ezeket a számszerű összefüggéseket felhasználja a kimeneti neuronok egymáshoz képest való elhelyezkedésének kiszámításában.

A neuronok hálózatbeli elhelyezkedését a hozzájuk rendelt súlyvektor jellemzi. A valódi idegsejtekhez hasonlóan a mesterséges neuronok is szinaptikus kapcsolatban állnak egymással. Ahogyan az idegsejtek esetében is léteznek gátló (információáramlást

akadályozó) és serkentő (információáramlást engedélyező) szinaptikus kapcsolatok, így az önszervező térképeknél is megvan ez a hasonlóság.

A Kohonen-féle önszervező térképek működési elve lépésekre lebontva a következőképpen jellemezhető. Legyen $\mathbf{x}_k$ az n dimenziós tanítási minta. Jelöljük $\mathbf{w}_{ij}$ -vel a kiválasztandó neuront, amelynek elhelyezkedését a mintában i és j mutatja meg. A tanulási ráta értéke $0 \leq \alpha \leq 1$. A neuronok egymástól való távolságát a szomszédsági függvény, $h(\mathbf{w}_{ij}, \mathbf{w}_{mn})$ határozza meg. A szomszédsági függvény magas értékeket ad vissza, ha a neuronok közel vannak egymáshoz a kimeneti térben. A bemeneti mintából kiválasztott nyertes neuron pedig $\mathbf{w}_{nyertes}$ lesz (rövidítve BMU, best-matching unit).

Minden egyes bemeneti mintára a következő műveleteket hajtjuk végre:

1. Kiszámítjuk a minta és az összes neuron közti távolságot (egyenként).

$$d_{ij} = \|\mathbf{x}_k - \mathbf{w}_{ij}\|$$

2. A legközelebbi (a legkisebb távolságra eső) neuron lesz a nyertes neuron.

$$\mathbf{w}_{nyertes}(\mathbf{w}_{ij} : d_{ij} = \min(d_{mn}))$$

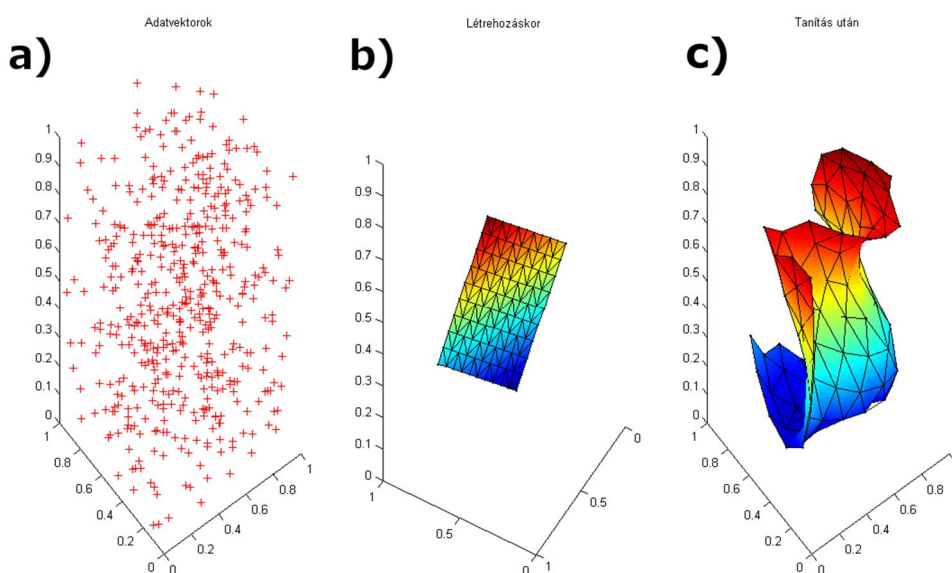
3. Miután kiválasztottuk a nyertes neuront, minden egyes neuron értékét frissítjük.

$$\mathbf{w}_{ij} = \mathbf{w}_{ij} + \alpha \cdot h(\mathbf{w}_{nyertes}, \mathbf{w}_{ij}) \cdot \|\mathbf{x}_k - \mathbf{w}_{ij}\|$$

4. A folyamat a 3. lépés után újraindul, és minden egyes bemeneti neuron esetén megtörténik a nyertes neuronnal (BMU-val) való összehasonlítás. A folyamat mindaddig ismétlődik, amíg el nem érünk egy előzetesen megadott feltételt. Ez lehet például az iterációk száma.

A súlyvektorok közötti távolság kiszámítására többféle matematikai módszer is létezik, azonban leggyakrabban az euklideszi távolságmetrikával találkozhatunk, számításaink során mi is ezt használtuk.

Amint látható, a SOM egyesével változtatja meg a súlyvektorok értékét, amíg azok közelebb nem kerülnek a BMU-hoz, ezt a típusú felépítést nevezik adaptívnek. A tanulási ráta (α) egy számérték, amely meghatározza, hogy az egyes tanítási lépésekben a neuronok mennyivel kerüljenek közelebb a nyertes neuronhoz. A hálózat tanítása lehet iteratív, amikor a fentebb vázolt módon, a teljes adathalmazt lépésenként (iteráció) építi be a térképbe, a folyamatot az egyes neuronok szintjére lebontva.



2. ábra: önszervező térkép tanítása demo adatsorral:

a) az adatvektorok elhelyezkedése a vektortérben, b) a létrehozás utáni helyzet, c) tanítás után.

Létezik egy másik megoldás, a köteget tanítás (batch training), amelyben a teljes adathalmazt a benne lévő hasonlóságok alapján Voronoi-cellákra bontja, amelyek egy-egy neuron köré rendeződnek. Az önszervező térképek többsége alapértelmezés szerint az iteratív tanítási módot használja.

A SOM tanítása után a bemeneti térben egymáshoz közel elhelyezkedő minták a kimeneti térben is egymáshoz közel helyezkednek el, és ez a kapcsolat fordítva is igaz. Ezért mondhatjuk, hogy a SOM a tanítás során megőrzi az eredeti topológiát (szomszédsági értékeket). A tanítási lépések számától függetlenül mindig van azonban eltérés a bemeneti minta és annak neuronhoz való rendelése között. Ezt nevezzük kvantálási hibának. Itt érdemes megjegyezni, hogy a SOM tanítási algoritmus nagyban

hasonlít a digitális jelfeldolgozás és adattömörítés területén használt vektorkvantáláshoz, amely szintén egy nem felügyelt típusú tanítás.

A kimeneti teret jellemző számérték a topológiai hiba. Ez az érték azt mutatja meg, hogy egy adott bemeneti minta második legközelebbi szomszédja hányszor nem kerül be a neuron kimeneti szomszédjai közé.

A kimeneti réteg ábrázolásán kívül a térkép megjeleníti még az U-mátrixot is, amely a szomszédos neuronok adatvektorainak euklideszi távolságát ábrázolja színárnyalatok segítségével. Leolvasásával a teljes adathalmazra vonatkozó információkat kaphatunk. Világos színek jelölik a klasztereket, a sötétek pedig a klaszterek közti határvonalat.

2.3.A SOM implementációja a MatLab programozási nyelvben

Az önszervező térképek informatikai megvalósítására több alkalmazás is akad. Az egyik legkedveltebb a természettudományokban és műszaki tudományokban gyakran használt MatLab szoftverhez a Helsinki Műszaki Egyetem által 2005-ben közreadott csomag (SOM ToolboxforMatLab). A SOM függvényei a MatLab rendszer saját. M fájlformátumaként kerülnek betöltésre. A SOM Toolbox függvényei egy saját adatformátumot, a SOM_PAK adatszerkezetet használnak, de lehetőség van bármely ASCII numerikus adat importálására külső forrásokból. A támogatott fájlformátumok között van a .DAT, de külön MatLab függvény segítségével közvetlenül Microsoft Excel (.XLS) táblázatok is beolvashatók. Mivel a függvények csak a numerikus adatokat dolgozzák fel, a szöveges adatok külön címke mezőben tárolódnak, a ki nem töltött táblázatcellák pedig NaN (Not-a-Number) értékként jelennek meg. A SOM Toolbox jellegzetessége, hogy az adatokat adatstruktúrákban (vagy adatszerkezetekben, datastruct) tárolja, amelyek különböző mezőkből állnak (programozási szempontból adatvektorok, ill. mátrixok). Például külön mezőben tároljuk magukat az adatokat és a hozzá rendelt címkéket.

Hogyan néz ki ez egy Excel táblázat esetében? A táblázat oszlopai a térkép változói lesznek, a sorok pedig ezen változók egyes értékei (pl. egy vizsgált objektum különböző

jellemzői, vagy mérések idősorai). Amennyiben a táblázat tartalmaz sztring adatokat, azok a címke mezőben fognak tárolódni, amely a térkép feliratait adják.

Az adatok importálása után az előfeldolgozás következik. Ennek során a rendezetlen adathalmazból eltávolítjuk a nem kívánt értékeket, majd azonos nagyságrendbe váltjuk át őket.

A térkép kétféle módon (random és lineáris) inicializálható, tanítási módja pedig (mint fentebb említettük) lehet iteratív és köteget. Alapbeállítás szerint a program a lineáris inicializálást és az iteratív tanítást használja.

Az elkészült térképet három fő módszerrel tudjuk megjeleníteni:

1. méhsejtszerű megjelenítés, amely a kimeneti rétegben neuronjainak elhelyezkedését mutatja,
2. grafikonos ábrázolás, amelyen minden egyes neuron helyén egy-egy diagram van,
3. rácsszerű megjelenítés.

2.4. Önszervező térképek használata előrejelzésre

A SOM, mint klaszterező eljárás elterjedt a gazdasági adatok feldolgozásában. Pablo-Martí cikkében[6] kimutatta a gazdasági tevékenységek regionális eloszlását a (spanyol) nemzetgazdaságban; Collan, Eklund és Back[7] pedig a világ országainak gazdasági fejlettség szerinti csoportosítására használták az önszervező térképet. Ezek a példák többnyire a kétdimenziós, térképszerű leképezés nyújtotta előnyöket aknázzák ki, hiszen ez a megjelenítés rávetíthető egy valós földrajzi térképre is.

Az önszervező térképek, mint statisztikai becselő módszerek használatára számos példát találhatunk a szakirodalomban[8]. Ezek igen szerteágazó területekre is jellemzőek.

Al-Shayea és munkatársai [9] 66 spanyol bank adataiból készítettek adatbecslő modellt, amely a bankok pénzügyi kimutatásaiból képes a csődkockázat előrejelzésére. Ultsch és Röske[10] historikus meteorológiai adatok alapján működő tengervízszint-előrejelző modellt állítottak fel önszervező térkép segítségével.

2.5.A tárgyalt probléma gazdasági háttere

2.5.1. A kutatás-fejlesztési szektor nemzetgazdasági jelentősége

E rész tárgyalásához elengedhetetlenül szükséges ismerni három alapfogalmat, nevezetesen a kutatás, a fejlesztés és az innováció közti különbséget. A Központi Statisztikai Hivatal szerint a kutatás és fejlesztés „...jellemzői: az alkotás eleme, az újdonság eleme, a tudományos módszerek alkalmazása és új ismeret létrehozása.”

A kutatás és fejlesztés fogalmak a mindennapi életben elterjedt jelentésük szerint abban különböznek, hogy a „kutatás” kifejezést az alapkutatás szinonimájaként használják, míg a „fejlesztés” alatt a már meglévő, a piacra bevezetett technológiák hatékonyabbá tételét, továbbfejlesztését értik.

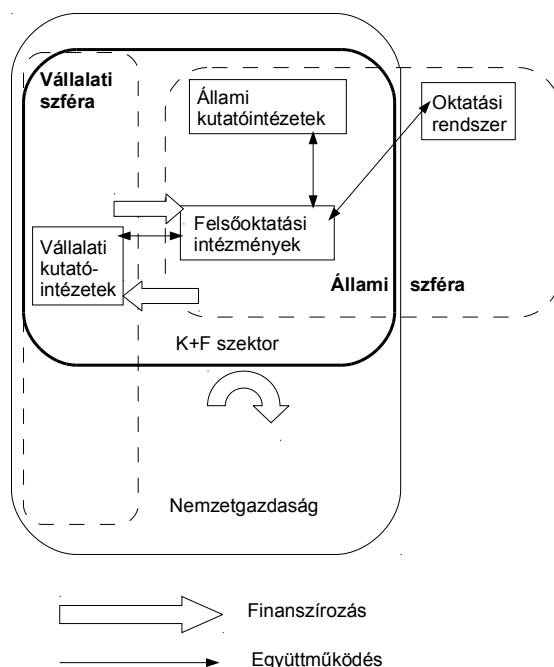
Ugyancsak a KSH szerint az innováció „új vagy jelentősen továbbfejlesztett termék (áru vagy szolgáltatás) vagy eljárás, új marketingmódszer, vagy új szervezési-szervezeti módszer bevezetése az üzleti gyakorlatba, munkahelyi szervezetbe vagy a külső kapcsolatokba.”

A kutatás-fejlesztési ágazatba (a továbbiakban K+F) e három fogalom tartozik bele a köznapi használat szerint.

Mindenki számára egyértelmű, hogy az elmúlt néhány évtized alatt a K+F szektor a világ legtöbb országában stratégiai ágazattá vált, amelytől nagyban függ az adott országok gazdasági versenyképessége. Röviden összefoglalva mondhatjuk, hogy a K+F igen hosszú távú befektetés, mivel több esetben először csak közvetett hatással van a tényleges gazdasági tevékenységre, megtérülése pedig évtizedekben mérhető.

Amilyen fontos a K+F, ugyanolyan nehezen mérhető számszerűen a nemzetgazdaságra kifejtett tényleges hatása. A K+F tevékenységek egyaránt jelen vannak az állami, ill. a versenyszférában. Fontos kiemelni a felsőoktatás szerepét, amellyel a következő fejezet is foglalkozik.

Finanszírozási források tekintetében ugyanilyen bonyolult a helyzet, hiszen a felsőoktatási intézmények is indítanak vállalatokkal közös kutatási projekteket, ugyanakkor a versenyszféra cégeinek is lehetőségük van pályázni állami támogatásokra a K+F tevékenységükhöz. Ez a kölcsönös viszony az alábbi ábrán szemléltethető:



3. ábra: a kutatás-fejlesztési szektor intézményrendszere és finanszírozási forrásai

A kutatás-fejlesztési tevékenységek gazdasági hatásait többen is próbálták elemezni. Mojtahedzadeh és Abedi[11] cikkükben statisztikai hipotézisvizsgálattal igazolták a cégek gazdasági adatainak és K+F tevékenységüknek a kapcsolatát. Négy hipotézist állítottak fel, amelyek alapján arra jutottak, hogy statisztikailag szignifikáns kapcsolat áll fenn a K+F ráfordítások, valamint: a cégek eladási statisztikái, üzemi költségei és befektetési kockázati értékelésük között.

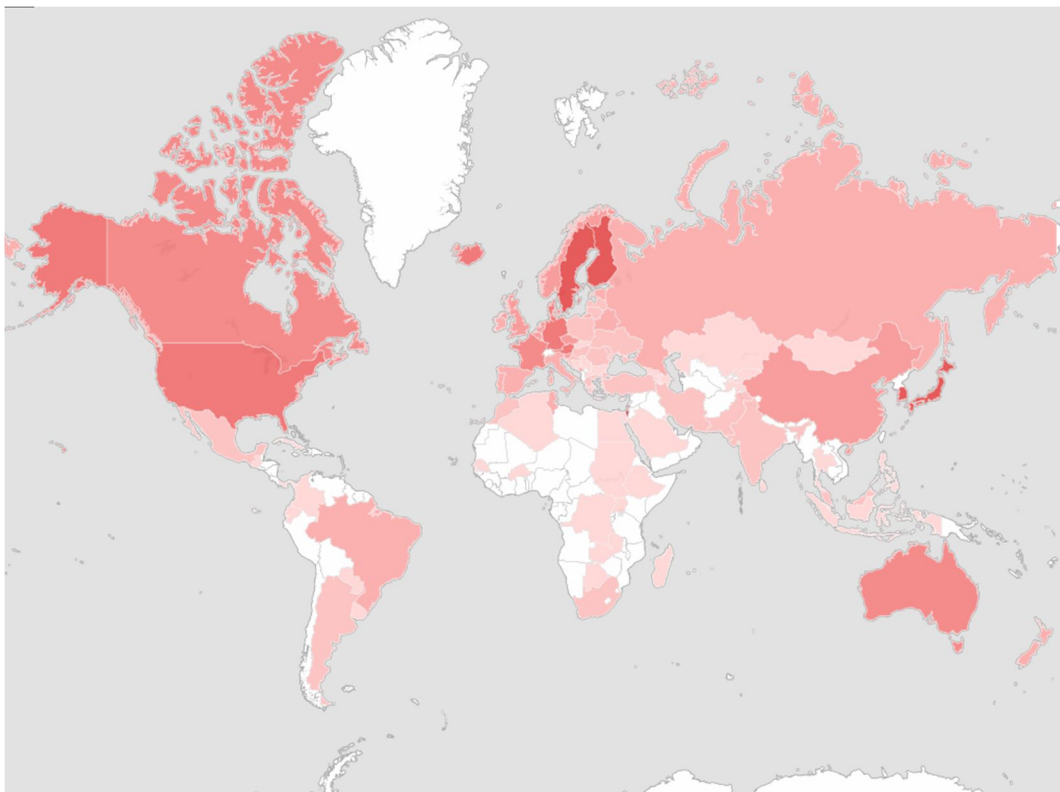
Hall, Mairesse és Mohnen[12] megállapították, hogy a K+F befektetéseknek egyetlen megtérülési rátája nincsen, mert az országonként és iparáganként is eltér.

A sikeresség egyik feltétele, hogy közvetlen együttműködés jöjjön létre az ipari innovációt és az alap- vagy alkalmazott kutatásokat végzők között[13], ezért az elmúlt években megfigyelhető tendencia, hogy egyre szorosabb az együttműködés a versenyszféra és a kutatás-fejlesztés szereplői között.

Egy 2000-ben született OECD-tanulmány[14] megállapítja, hogy a kormányzat szerepe kettős a K+F tevékenységeknek kedvező környezet kialakításában: a) közvetlen

támogatással kell fejlesztenie a felsőoktatási és állami kutatóintézeteket, b) kormányzati ösztönző programokkal, adókedvezményekkel kell támogatnia az versenyszféra kutatási tevékenységét.

Az alábbi adatok érzékeltetik, hogy hazánk hol áll a nemzetközi összehasonlításban. 2010-ben Magyarország GDP-je 1,16%-át költötte K+F beruházásokra, a listavezetők Izrael (4,4%), Finnország (3,88%), Dél-Korea (3,74%).



4. ábra: a világ országainak K+F kiadásai 2008-2013 között a GDP %-ban, legvilágosabb: 0 %, legsötétebb: 5 %, fehér: nincs adat. Forrás: <http://www.worldbank.org>

2.5.2. A felsőoktatás, mint a K+F szektor utánpótlási bázisa

A felsőoktatási intézmények oktatási feladataik ellátásán túl jelentős kutató-fejlesztő tevékenységet fejtenek ki. Nemcsak diplomát adó képzéseként vannak jelen, hanem elmosódik a határvonal köztük és a kutatóintézetek között. A PhD-fokozat elérése magában foglalja bizonyos mennyiségű kutatómunka elvégzését, amelyből – még ha az illető nem is tudományos munkatársként dolgozik tovább – kísérleti eredmények, publikációk születnek.

Az elmúlt húsz évben vált elterjedtté az ún. tudásközpontok felépítése, ahol cégek képviselői, állami, vagy privát kutatóintézetek, ill. egyetemek kihelyezett intézetei együtt, egy helyen dolgoznak, közös projekteken. Megjegyzendő, hogy ez a szisztéma Magyarországon még nem eléggé kiforrott.

A hazai felsőoktatás intézményrendszerét és minőségi követelményeit alapvetően meghatározzák az oktatáspolitikai reformok, úgymint a kétszintű érettségi vizsgarendszer bevezetése, amely felváltotta az egyetemi felvételik rendszerét.

Emellett a bolognai folyamat, mely Magyarországon 2006-ban lépett életbe, jelentős hatással volt a képzés szerkezetére és követelményeire. A régebben ötéves egyetemi képzéseket – kevés kivételtől eltekintve – egy három- (BSc / BA), ill. egy kétéves (MSc / MA) képzésből álló szerkezet váltotta fel. Ugyancsak szűkült a szakok választéka is: pl. a 2005-ös 400 szak helyett 2006-ban már csak 108 alapszakból választhattak a jelentkezők.

A 2000-2010 közötti évtizedre a felsőoktatási létszámok nagyarányú bővülése volt jellemző. A teljes lakosság arányához viszonyítva sokkal többen iratkoztak be felsőoktatási intézménybe és szereztek diplomát, mint bármikor azelőtt. Ennek elemzésekor azonban számolni kell néhány, az adatokat torzító tényezővel. A statisztikákban például a felvételizők létszámánál 2006 után a mesterképzésre jelentkezőket is figyelembe kell venni.

Mindazonáltal a legjelentősebb hatást a felsőoktatási hallgatói létszámokra az általános demográfiai tendenciák gyakorolják. A hosszú távú idősor elemzésénél egyértelműen látható, hogy a csökkenő népességszámhoz természetesen csökkenő hallgatói létszámok tartoznak – és itt a reformok okozta változások csak kisebb mértékű ingadozásokat eredményeznek. Ahogyan a demográfiában a népességet kohorszokra osztva végigkövethető az egyes korosztályok „életútja”, ugyanígy viszonylagos biztonsággal megbecsülhető az is, hogy egy adott korosztály hány százaléka jut majd el a felsőoktatásig.[15][16]

3. Adatok és módszerek

3.1. Adatok kiválasztása, feldolgozása

A dolgozat megírásához szükséges adatok különböző helyekről származnak. A főbb források a Központi Statisztikai Hivatal, a Felvi.hu valamint az Oktatási Hivatal internetes portáljai voltak. Összesen 152 db hivatalos adatot, valamint az ezekből származtatott adatokat használtuk a becsléshez.

A KSH statisztikái közül a Népeség, népmozgalom[17], valamint az Oktatás[18]idősoros adatait használtuk fel. A Népeség, népmozgalom adatai közül a számunkra fontos adatok az élve születések száma (1982-2012, 31 adat), a népesség száma (2000-2012, 13 adat), valamint a csecsemőhalandóságok száma (1982-2012, 31 adat), voltak.

Az oktatási statisztikából a középiskolában tanulók (2000-2012, 13adat), a felsőoktatásban tanulók (2000-2012, 13 adat), az érettségi vizsgát tettek (2000-2012, 13 adat), valamint a felsőfokú oklevelet szerzettek száma (2000-2012, 13 adat) származik.

A Felvi.hu valamint az Oktatási hivatal portáljáról a felsőoktatással kapcsolatos további adatok álltak a rendelkezésünkre. A Felvi.hu statisztikái közül az elmúlt évek statisztikáit használtuk fel[19], ezen belül is a jelentkező és felvett diákok számát mutató idősoros táblázatot, amelyből a felvett hallgatók számát (2001-2012, 12 adat) szereztük meg.

Az Oktatási Hivatal portálján megtalálható felsőoktatási statisztikákat[20]. felhasználva megkaptuk a kezdő évfolyamos hallgatók adatait (2000-2012, 13 adat).

3.2.A modell felállításának lépései

A rendelkezésekre álló adatokból becsléseket végeztük. Ehhez a Microsoft Office 2010-es Excel programját használtuk. Ennek segítségével megkaptuk a 18 éves korosztály létszámát az élve születettek számának eltolásával (2000-2030, 31 adat), a hallgatók-felvettek számát (2000-2012, 13 adat), valamint a felsőoktatásban eltöltött évek számát is (2001-2012, 12 adat). Ezt úgy kaptuk meg, hogy a hallgatók-felvettek számát elosztottuk a felsőfokú oklevelet szerzettek számával. [18] [19] [20] [21]

A becslések nagy részét is a Microsoft Excel programmal végeztük. A végső becsléshez azonban a MATLAB nevű programot és az abba implementált SOM Toolbox-ot használtuk. Az Excel becslés segítségével (2013-2030) a felsőfokú oklevelet szerzettek számát (18 adat), az eltöltött évek számát 2 módon, egyszer exponenciálisan másként pedig az átlagos évek számából, (36 adat), hallgatók- felvettek számát (18 adat), hallgatók * évek valamint a hallgatók létszámát kaptuk meg.

A MATLAB-os becsléshez 60 adatot használtunk fel.(ld. Függelék 8.4) Ezek a hallgatók*eltöltött évek száma, a 18 éves korosztály száma, eltöltött évek száma, diplomát szerzők száma valamint az érettségi vizsgát tettek száma volt. Az adatok közti kapcsolat feltárásához a SOM Toolbox-ot használtuk, amely a klaszterek és a gradiens kirajzolásával teszi lehetővé az összefüggések megállapítását.

A Toolbox-ot ingyenesen a <http://www.cis.hut.fi/> [22] weboldalról töltöttük le majd implementáltuk a MATLAB- fejlesztői környezetébe.

A SOM térképet betanításkor különböző beállításokkal hoztuk létre (ld. Függelék 8.1). A minta nevű mátrixba (12x5) helyeztük el a rendelkezésre álló 60 adatot. A programkód első részében a D(=Data,adat) nevű változóba importáltuk az adatstruktúrát, majd ezt követően normalizáltuk az adatokat, így minden tényező azonos súllyal számít a vizsgálatban. Az adatok normalizálását lineáris módon végeztük ezen belül is range típust használtunk. A range típus azt jelenti, hogy az adatok egy 0-1-ig tartó tartományon belül veszik fel az értékeket az adatsor legnagyobb értékéhez arányítva. Ezt követően elkezdtük betanítani a normalizált adatokból a SOM hálót, amihez az Euklideszi távolság mérést használtuk, mert a szakirodalmi tapasztalatok szerint ez a legrobosztusabb módszer. Ennek a módszernek az alapja a Pitagorasz-tétel. Jelen esetben két pont távolsága, a pontok koordinátáinak, különbségnégyzetének gyökével egyezik meg:

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2}$$

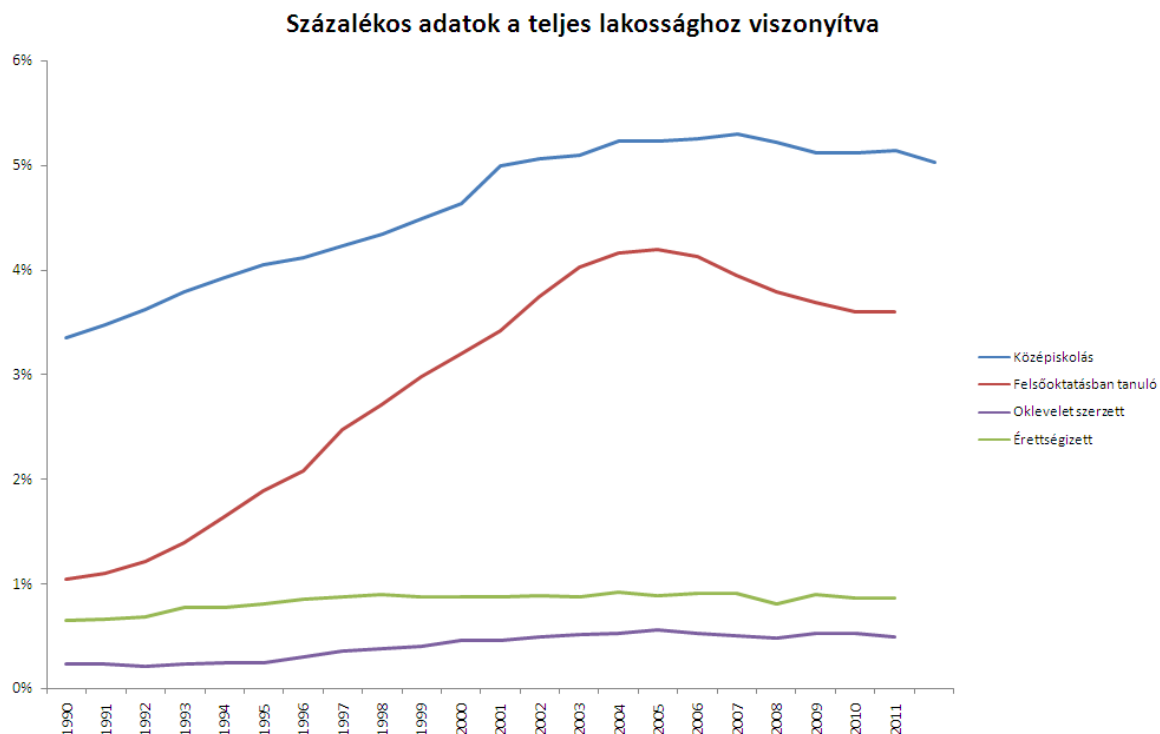
Az ötödik sorban meghatároztuk a térkép nevét ezután pedig megjelenítettük a már betanított térképet.

A modellezéshez szintén önszervező térképet készítettünk az elsőhöz hasonlóan, csak itt már a 2001-2030ig terjedő adatokkal. A modellezést két scenárió alapján is elkészítettük (ld. Függelék 8.2). Az egyikben a felsőoktatásban eltöltött évek számát exponenciális görbével becsültük (első scenárió), a másikban pedig a felsőoktatásban töltött időt állandónak feltételeztük (második scenárió). A térkép elkészítése után a becslés elvégzéséhez kimentettük az egyes súlyvektorokhoz legközelebb eső 4 neuron adatait (BMU-t). Ezt követően egy index nevű változóba elhelyeztük az egyes sorokhoz tartozó BMU-kat. Ez a változó segítette megtalálni a súlymátrixban az indexhez tartozó értékeket. Utolsó előtti lépésként létrehoztunk egy mátrixot, amelyben a súlyok átlagát helyeztük el. Az utolsó lépés az adatok denormalizálása volt, amelynek eredményeként megkaptuk a kérdéses adatokat.

4. Eredmények és értékelésük

Az első vizsgálatban a neurális háló elkészítéséhez a rendelkezésre álló adatsorokat összeegyeztettük.

A teljes lakossághoz viszonyítva vizsgáltuk meg a középiskolások, a felsőoktatásban tanulók, az érettségis tett valamint az oklevelet szerzettek számát. (5. ábra)



5. ábra: Középiskolások, az érettségizett, a felsőoktatásban tanulók valamint az oklevelet szerzett diákok aránya a teljes lakossághoz viszonyítva

A diagramon jól látszik, hogy a különböző adatok értékei az évek során jelentősen változtak. Mind a középiskolások, az érettségizők, mind a felsőoktatásban tanulók és az oklevelet szerzettek számának aránya az 1990-es évek elejétől egészen 2006-ig nagymértékben nőtt.

Az érettségizők, középiskolások, valamint az oklevelet szerzettek száma a 2006-ot követő és már az azt megelőző időszakban 2000-től kezdődően stagnálni kezdett, egészen 2008-igamikor is egy nagyobb mértékű visszaesés tapasztalható. Azonban a 2008-at követő években a számuk visszatért a visszaesés előtti állapothoz, és azóta is kisebb eltérésekkel állandó.

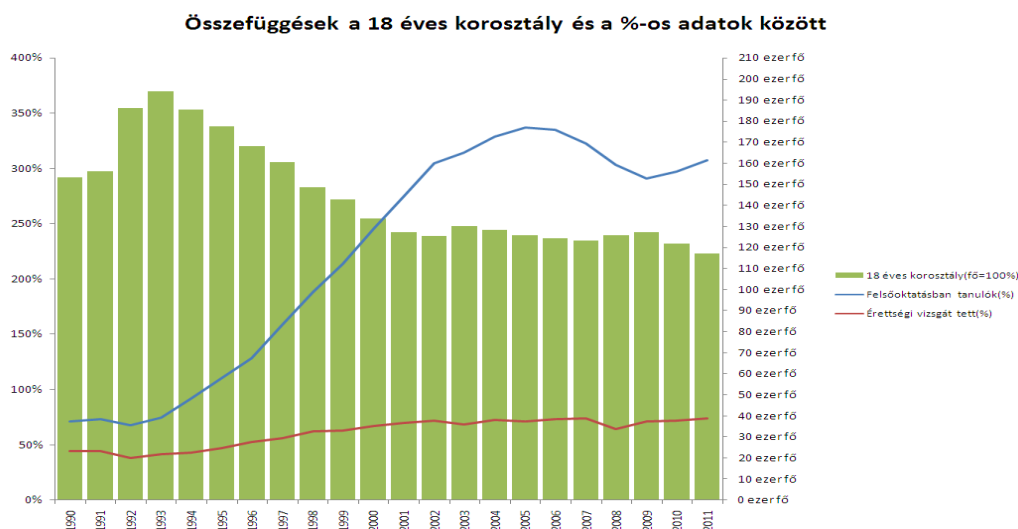
A felsőoktatásban tanulók számának változása, azonban máshogy alakult. A növekedés egyik kiváltója az 1985-ös felsőoktatási törvény illetve annak az 1990-es módosításának kiadása volt, hiszen ennek hatására megindult az oktatás tömegessé válása. Ezt számos olyan törvény követte ami, tovább növelte a hallgatók számát, többek között olyan intézkedésekkel, mint a költségtérítéses képzés valamint a felsőfokú szakképzés bevezetése.

Ez a növekedés 2006-ban érte el a csúcspontját, amikor új felsőoktatási törvényt vezettek be. Ennek következményeként Magyarország is áttért a Bolognai-rendszerre. Ez azt eredményezte, hogy elkezdett csökkenni a felsőoktatásban hallgatók létszáma. Egyrészt, azért mivel korábban a plusz előadásokért nem kellett fizetni, azonban 2006-tól egy bizonyos százalék után a hallgatónak már fizetniük kell „extra” tanulmányaikért. Másrészt 2006 előtt csak alapfokú nyelvvizsga bizonyítványhoz kötötték a diploma megszerzését, az ezt követő időszakban pedig már a középfokú nyelvvizsga lett a követelmény. Jelentős szerepe volt még a csökkenésben, annak a ténynek is, hogy bevezették a felvételi pontminimumot.

A jelentős csökkenés azonban 2008-tól figyelhető meg. Ez feltételezhetőleg egyrészt a gazdasági válság hatásaként másrészt a növekvő tandíj hatásaként értelmezhető.

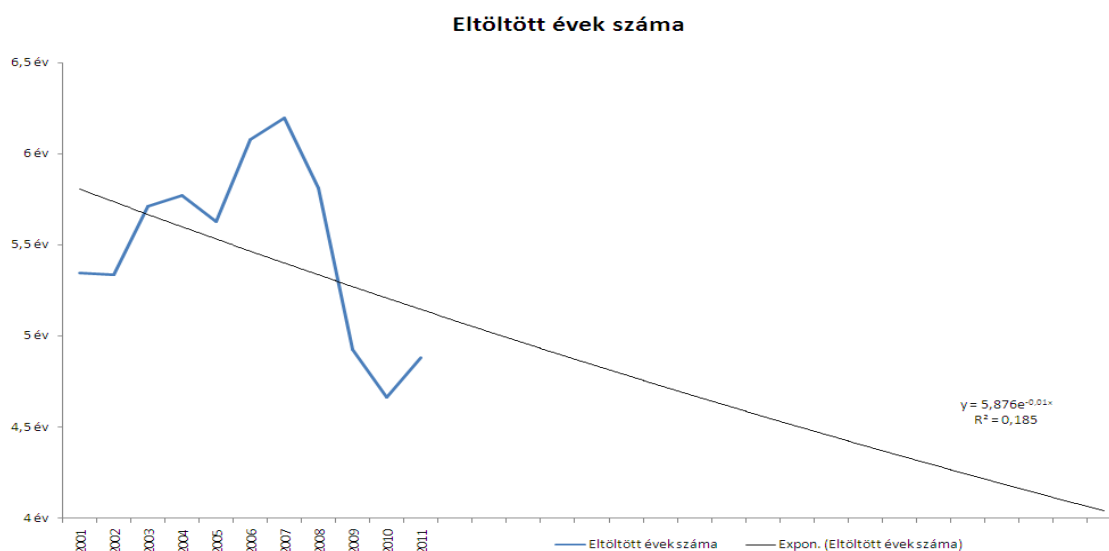
Ezen tények ismeretében a célunk a felsőoktatásban hallgatók számának megbecslése a következő húsz évre. Ehhez elsőként a 18 éves korosztály számát határoztuk meg úgy, hogy az élve születettek számát eltoltuk 18 évvel, így megkaptuk az adott évben lévő létszámukat. Mivel a csecsemő-, és gyermekhalandóság olyan kismértékű, hogy a becslésünket számottevően nem módosítaná, így eltekintettünk tőle.

Mivel már tudtuk a 18 éves korosztály létszámát, megállapítottuk, hogy a korosztály hány százaléka tett érettségit és hogy hány százaléka járt az adott évben felsőoktatásba. (6. ábra). A százalékos adatokat megvizsgálva észrevettük, hogy az utóbbi tíz évben az érettségizők számaránya stagnált, azaz átlagosan 71,13% volt. Ennek ismeretében már megbecsülhettük a számukat a következő évekre, így a 18 éves korosztály számát felszoroztuk az átlagos értékkel.



6. ábra: Összefüggések a 18 éves korosztály és a felsőoktatásban tanulók valamint az érettségi vizsgát tett diákok között

Ezt követően meghatároztuk a már minimum egy féléve a felsőoktatásban hallgatók számát úgy, hogy az összes hallgatói létszámából kivontuk a felvett hallgatók számát. Az évek számának megállapításakor nem csak a BSc. hallgatók számát, hanem az összes képzési szinten lévő hallgatók számát vettük figyelembe. A kapott értékeket elosztottuk a felsőfokú oklevelet szerzettek számával, ezzel megbecsülve az egyetemen eltöltött éveket, ami 5,5 évre jött ki átlagosan. Ezeket az adatokat diagramon ábrázoltuk. Az első szcenárióként exponenciális trendvonal segítségével a következő húsz periódusra (év) becsültük az értékeket. (7. ábra).



7. ábra: Eltöltött évek számának becslése exponenciális trenddel, vagyis az első szcenárió. A becsléshez 2001-2011-ig használtuk fel az adatokat. Ezeket az adatpontokat köti össze a kék színű egyenes. A trendvonal fekete színnel van jelölve. Az ábrán jól látszik az ehhez a vonalhoz tartozó egyenlet is, amelybe behelyettesítve megkaptuk az értékeket a következő húsz periódusra.

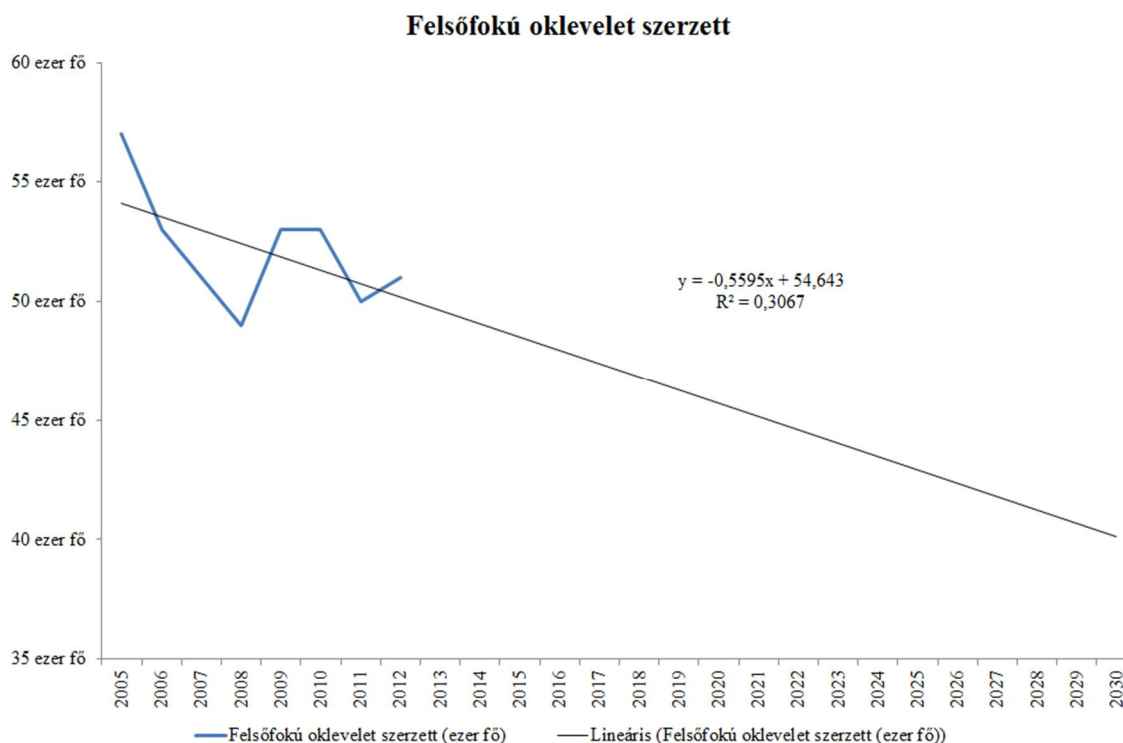
A diagramon megfigyelhető, hogy az eltöltött évek 2001-2010 között 4,5 és 6,5 év között ingadoztak. 2001-től kisebb ingadozással 2008-ig nőtt az évek száma, majd ezt követően jelentősen visszaesett. Az exponenciális trenddel való becslés azt mutatja meg, hogyha az értékük ilyen mértékben csökken, akkor 2030-ra várhatólag az eltöltött évek száma átlagosan 4-4,5 év lesz.

Feltételezve, hogy az évek száma nem csökken és megmarad az átlagos 5,5 év körül, egy másik scenáriót is végigkövettünk. (8. ábra).



8. ábra: Az eltöltött évek számának előrejelzése az átlagos eltöltött évek számával, vagyis a második scenárió. A becsléshez 2001-2011-ig használtuk fel az adatokat. Ezeket az adatpontokat köti össze a vörös színű egyenes. A kék egyenes a következő húsz évet prezentálja, abban az esetben, ha az eltöltött évek száma megmarad az átlagos szinten. Az átlagos szint a meglévő adatokból lett megbecsülve.

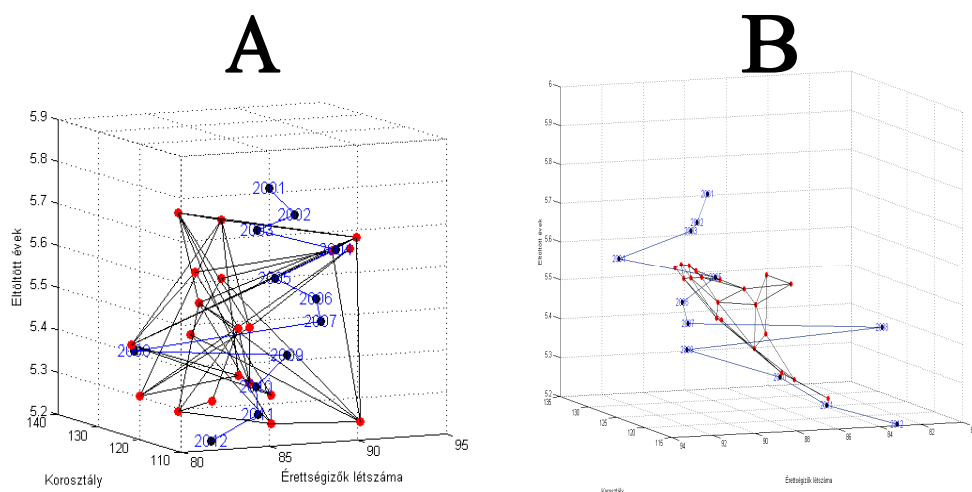
Már szinte az összes adat rendelkezésünkre állt a hallgatók számának megbecsléséhez. Ennek utolsó lépéseként megállapítottuk a felsőfokú oklevelet szerzettek számát és a már minimum egy felére a felsőoktatásban hallgatók számát is előre. Ehhez az adatokat szintén grafikonon ábráztuk, majd lineáris trendvonal segítségével szintén 20 periódusra megállapítottuk az oklevelet szerzettek értékeit. A 9. ábrán jól megfigyelhető az oklevelet szerzettek számának változása. A 2005-től 2008-ig tartó időszakban az oklevelesek száma csökkent, majd az ezt követő években nőni kezdett egészen 2011-igamikor egy újabb csökkenés figyelhető meg. Ha megmarad ez a csökkenő tendencia, akkor 2030-ra az átlagos 51,8 ezer fő körülbelül 40 ezer főre fog csökkeni. Ez természetesen egy nagyon durva becslés csak. Ezt a képet kell az önszervező térkép módszerrel finomítani.



9. ábra: Felsőfokú oklevelet szerzettek száma. A diagram elkészítéséhez 2005-től 2012-ig használtuk fel az adatokat. Ezeket az adatpontokat a kék egyenesek kötik össze. A fekete egyenes a lineáris trendvonalat jelenti. A kapott egyenletbe visszahelyettesítve kaphatjuk meg a következő 20 évre a keresett adatokat.

Most már minden szükséges adat megvolt a Self-Organizing Map (SOM) elkészítéséhez. A neurális háló megalkotásához a hallgatók*évek számát, a 18 éves korosztály számát, az eltöltött évek számát, a diplomát szerzők számát valamint az érettségi vizsgát tettek számát használtuk fel 2001-2011-ig.

A térképet 55 db adatból készítettük el, amit importáltunk a MATLAB nevű programba. Az adatok rendeződését a 10. ábrán láthatjuk. Az "A" ábrát a már normalizált adatokból hoztuk létre. Ez a kiindulási állapotot tükrözi. A „B” ábrán szintén ugyanabban a dimenziós térben elhelyezkedő adatpontokat láthatjuk, de itt már rendezett formában, köszönhetően a betanításnak. (ld. Függelék 8.3) Az ábrákon egyaránt láthatóak az importált adatok, a BMU-k valamint a szomszédos adatpontok között lévő kapcsolatok.



10. ábra : A tanulás előtt, B, tanulás után. A piros pontok a bevitt adatokat jelölik, míg a kék pontok pedig az egyes BMU-kat jelölik. A B ábrán megfigyelhető, hogy az A-val ellentétben, már kevesebb adatpont látszik. Ez a módszer dimenziócsökkentő tulajdonságával magyarázható. A fekete vonalak a szomszédos adatpontokat kötik össze. Az ábrák elkészítéséhez 33 adatot használtunk fel a rendelkezésre álló adatok közül. Az adatpontok egy 3D-s ábrán lettek ábrázolva. Ez azt jelenti hogy három tengely van és ezek egymástól különbözőek. A három tengely az Érettségizők létszáma, a Korosztály létszáma, valamint az Eltöltött évek száma.

A háló betanítása után elkészítettük a térképet. A 11. ábrán ez a térkép látható. Mivel 2001-2011-ig tartó intervallum adatait használtuk, még nem kell különbséget tenni az első illetve a második scenárió között hisz ebben az időszakban ismertek voltak az adatok. Az ábrán jól látható hogy az U-mátrixban két csoport jól elkülönül. A térképen az adatokat visszakeresve arra a megállapításra jutottunk, hogy a két csoport választóvonal a 2008-as év.



Felsőoktatási minta

11. ábra: Az elkészült önszervező térkép. A SOM elkészítéséhez 2001-2011-ig használtuk fel az adatokat. Jól látható, hogy 5 különböző változót vittünk be a térkép létrehozásakor. A mintázat felismerés segítségével az adatok közti kapcsolat könnyen megállapítható. Az U-matrixban két jól elkülönülő cluster figyelhető meg. Az adatokat visszakeresve könnyen megállapítható, hogy a két csoport választóvonal a 2008-as év.

A számításvégén két minőségjelző adatot kaptunk vissza:

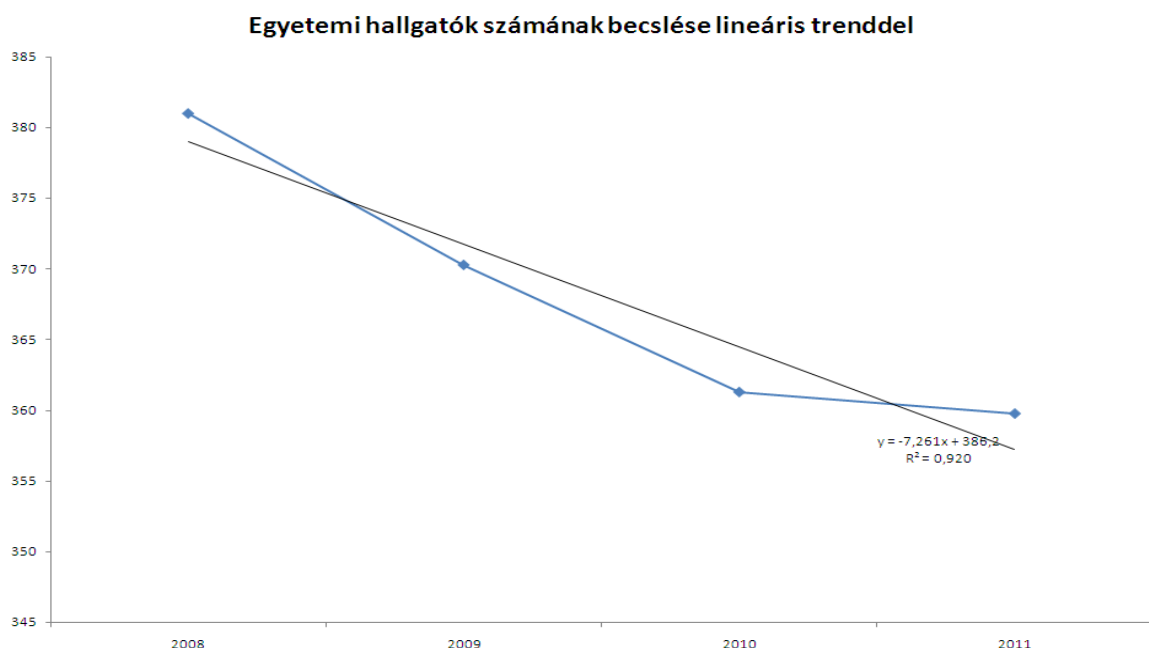
Final quantization error: 0.292

Final topographic error: 0.000

A kvantálási hiba az egyes adatok és a hozzájuk tartozó BMU-k közti átlagos távolságot mutatja meg. A topográfiai hiba pedig azt mutatja meg, hogy a különböző adatpontoknál mekkora százalékkal fordul elő, hogy a hozzájuk tartozó egyes BMU-k nem egymás szomszédságában találhatók. Mint láthatjuk az elkészült térkép esetében ezek a hibák kis százalékban jelennek meg. A kvantálási hiba 29%-os, ami esetünkben azt jelenti, hogy a bemeneti minta pontossága 71%-nak felel meg. A topográfiai hiba esetünkben 0%, ami azt jelenti, hogy az egyes BMU-k egymás szomszédságába esnek.

Ezeknek tudatában 2011-től lineáris trenddel megbeccsültük a hallgatók számát 2030-ig.

A hallgatók számának becslését két módszerrel is elvégeztük. Az első módszert az Excel lineáris trenddel való becslési módszerével készítettük el. (12.ábra)

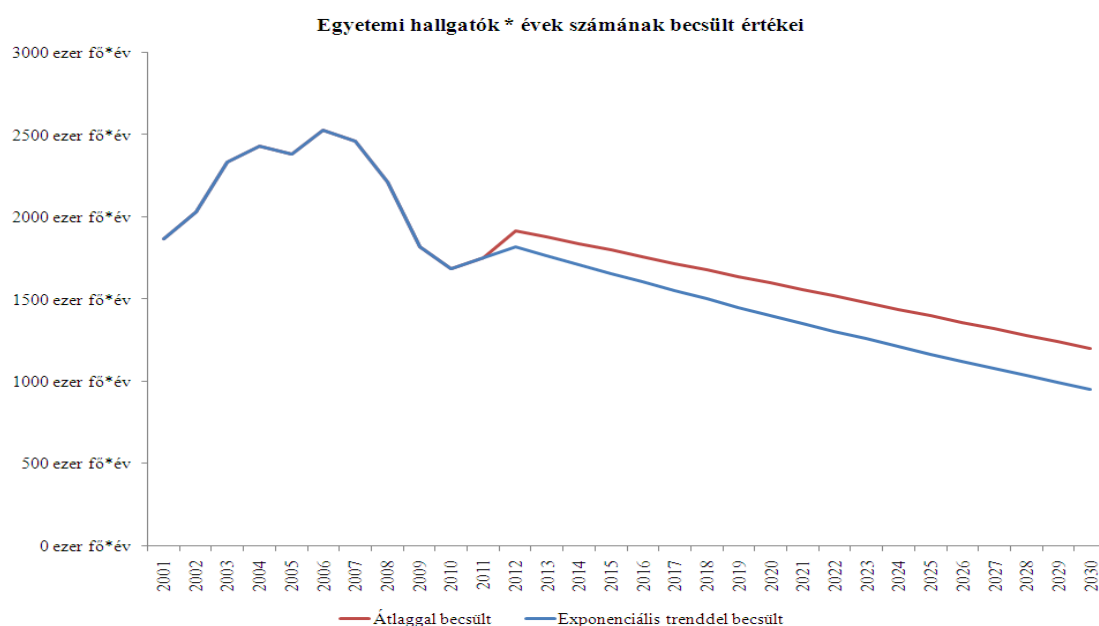


12. ábra: Egyetemi hallgatók számának becslése lineáris trenddel. A becsléshez a 2008-2011 közötti adatokat használtuk fel, mint az jól látszik a diagram vízszintes tengelyén. Az egyes adatpontokat, amelyek kék rombuszsal vannak jelölve, a kék színű vonalak kötik össze. Az erre a kék szakaszokból álló görbére lett ráillesztve a fekete vonallal jelölt lineáris trendvonal. A lineáris trend vonal és az általa szolgáltatott egyenlet segítségével meg tudtuk állapítani a hallgatók létszámát a következő időre.

A módszerrel kapott egyenlet segítségével meghatároztuk a számukat 2030-ig. Itt nagymértékű csökkenés figyelhető meg. A kapott értékek megdöbbentő eredményt mutatnak. A lineáris becslés eredménye szerint az egyetemi hallgatók száma 2030-ra a maihoz képest körülbelül a felére fog csökkenni. Azonban meg kell jegyeznünk, hogy a

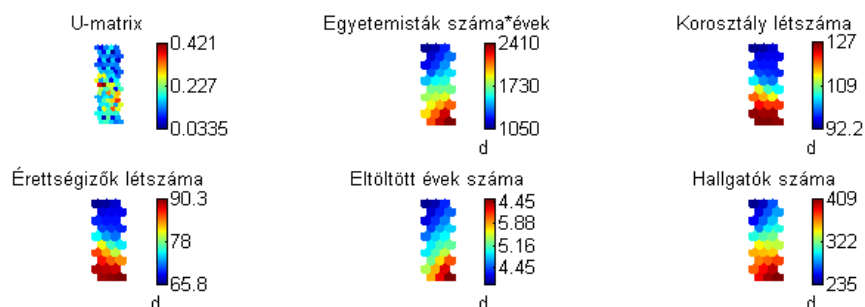
lineáris trend nem illeszthető pontosan az adatokra, mert egy bizonyos idő után az értékek mínuszra váltanak.

A lineáris becslés pontosítása végett elvégeztük a SOM-mal történő becslést. Itt első lépésben a lineáris trenddel kapott értékeket szoroztuk meg elsőként az átlaggal megállapított eltöltött évek számával (első scenárió), majd az exponenciális trenddel becsült évek számával (második scenárió). A két variáció, ahogy a 13. ábrán is látható nagyon hasonló egymáshoz, tehát kevés eltéréssel, de közel megegyező eredményeket várunk.



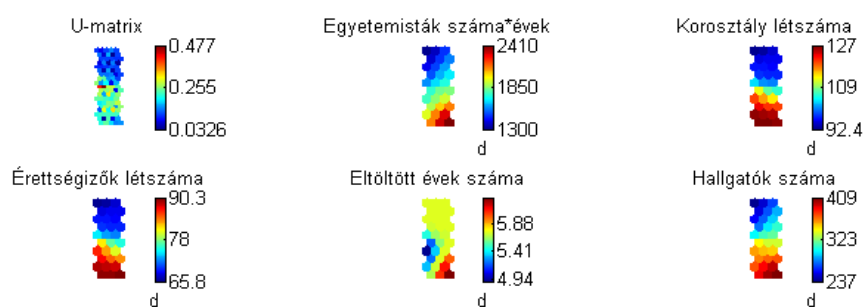
13. ábra: Egyetemi hallgatók * évek számának becsült értékei az első illetve a második scenárió alapján. Az első scenáriót vagyis az exponenciális trenddel valóbecslést a kék egyenes mutatja meg. A piros vonal az átlaggal becsült, vagyis a második scenárió értékeit mutatja meg. Az első és a második scenárió adatai 2001-2011-ig megegyeznek majd ezt követően szinte hasonló mértékben csökkeni, kezdenek. Az átlaggal való becslés mint láthatjuk kisebb csökkenést mutat meg mint az exponenciális trenddel való becslés.

Most már hogy ezek az adatok is a rendelkezésünkre álltak elvégezhetjük mindkét variációval a SOM-os becslést. Ennek első lépéseként, ahogy a 2001-2011ig tartó adatoknál is, importáltuk a 2001-2030-ig rendelkezésre álló adatokat a MATLAB környezetbe. Itt az adatok normalizálásán estek át, majd következett a tanítási fázis, aminek eredményeként rendeződtek az adatok. Az ezt követő lépésként elkészítettük a térképek ábráját. A 14. ábrán az első scenárió, míg a 15. ábrán az második scenárió látható.



Felsőoktatási minta

14. ábra: Az első scenárió, vagyis az exponenciális trenddel becsült adatok alapján elkészült önszervező térkép. A SOM elkészítéséhez 2001-től 2030-ig használtuk fel az adatokat. A térkép elkészítéséhez 5 változót használtunk fel. Ezeknek a változóknak a hexagonális térképei láthatók az ábrán- Az U-mátrixban jól megfigyelhetők a clasterek.



Felsőoktatási minta

15. ábra: A második scenárió, vagyis az átlaggal becsült adatok alapján elkészült önszervező térkép. A SOM elkészítéséhez 2001-től 2030-ig használtuk fel az adatokat. A térkép elkészítéséhez 5 változót használtunk fel.

A 14. ábra szinte mindegyik hexagonális térképén jól látszanak az elkülönülő adatok. Az eltöltött évek számánál valamint az érettségit tett diákok számánál egyértelműen megfigyelhető ezen az ábrán is a nagymértékű csökkenés. A 15. ábra U-mátrixán is nagyon jól kivehetők az egyes klaszterek, ugyanis egymáshoz közeli neuronok két csoportja látszik, közöttük viszonylag nagy távolsággal.

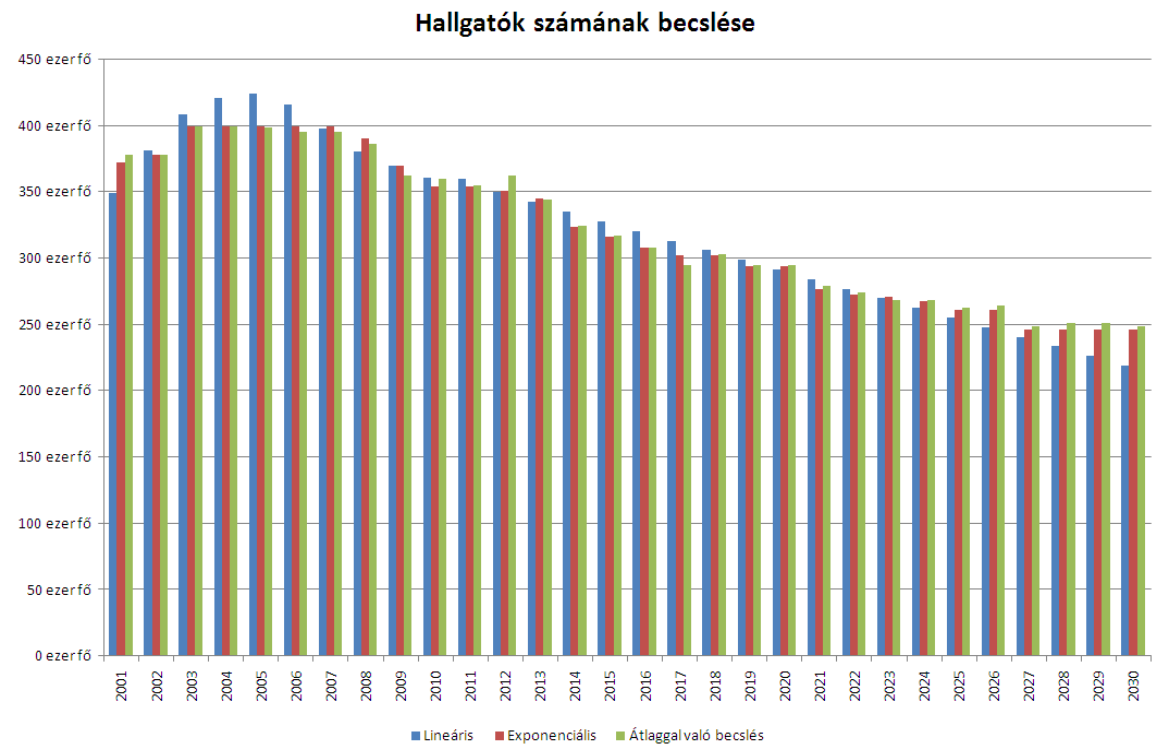
A térképek elkészülte után ismét kaptunk minőség jelzőket

	14. ábránál	15 ábránál
Final quantization error:	0.094	0.101
Final topographic error	0.033	0.000

Látható, hogy itt már vannak olyan adatok, amelyeknél az egyes BMU-k nem egymás szomszédságába esnek. A topográfiai hiba esetén minél nagyobb az érték annál komplexebb kimeneti teret kapunk. Ez azt jelenti, hogy felléphet, csoportosítási probléma valamint bonyolódhat a topológia. Nagyon magas érték esetén pedig arra következtethetünk, hogy nem volt megfelelő a tanulási folyamat. Azonban esetünkben mivel a minőségjelzők értékei alacsonyak, a kvantálási hiba 9 és 10%, a topográfiai hiba pedig 0 és 3%, ezért mondhatjuk, hogy a modell megfelelően működik.

A térkép elkészítése után a becsléshez kimentettük az egyes súlyvektorokhoz tartozó négy legközelebbi neuron adatait, amit arra használtunk fel, hogy a súlymátrixban megtaláljuk a hozzájuk tartozó értékeket. A megtalált adatokat átlagolva pedig megkaptuk a becsülni kívánt értékeket. A módszer jól működött, igaz kaptunk vissza ismétlődő adatokat, mivel a módszer szélsőérték esetén nem érzékeny. (ld. Függelék 8.6 és 8.7) Ez azt jelenti, hogy néhány esetben ugyanazokat a neuronokat találja a legközelebbinek, így ugyanazt az értéket is adja vissza.

A SOM-mal történő becslés eredményeként, az első scenárió alapján közel 30%-os csökkenés, míg a második scenárió alapján 35%-os csökkenés mutatkozott meg a 2011-es adatokhoz képest. Megfigyelhető tehát, hogy míg a lineáris becslés meghökkentő eredményt ad, addig a SOM-mal becsült eredmény már valamennyivel enyhébb változást jósol. Ez látható a 16. ábrán. A kapott eredmények csak egy lehetséges változatra mutatnak rá. Sok egyéb tényező is befolyásolhatja a becslés mértékét. Ilyenek például a külföldön tanuló, dolgozó hallgatók is, akiknek a létszáma lehet, hogy pontosítaná a becslésünket. Meg kell említeni ezen kívül, hogy sajnos a meglévő adatmennyiség nem volt elegendő, ahhoz hogy visszamenőleg is kipróbáljuk a módszert. Azonban jövőbeli terveink között szerepel a módszer ellenőrzése.



16. ábra: Hallgatók számának becslése lineáris trenddel és a SOM módszerrel kétféle scenárió esetén. A diagramon a kék színnel jelölt oszlopok mutatják az eredeti illetve lineáris trenddel becsült adatokat. Jól megfigyelhető, hogy az érték 2011-től jelentős mértékben csökken. A piros színű oszlopok az első scenárió alapján elkészült, míg a zöld a második scenárió alapján elkészült SOM-os becslés értékeit mutatják meg. Megfigyelhető, hogy az eredeti adatok esetén nemcsak a piros, de a zöld oszlopoknál is az önszervező térképes becslés kevesebb értéket mutat. Ez annak köszönhető, hogy nem csak a 2011 utáni adatokra néztük meg a legközelebbi neuronokat, hanem visszamenőleg is. A mintázat felismerés segítségével, így körülbelül 5%-os hibával sikerült visszakapnunk az eredeti értékeket.

5. Összefoglalás

Ebben a dolgozatban egy esettanulmány elkészítésén keresztül próbáltuk bemutatni, miként használható a Kohonen-féle önszervező térkép a közgazdasági adatok elemzésében. A SOM módszer segítségével a statisztikai adatok alapján becsültük meg a felsőoktatásban hallgatók számát 20 éves időtávlatban. A térképet jellemző számértékek (kvantálási hiba, topológiai hiba) arra utaltak, hogy a modell jól közelíti a valós helyzetet. A jelenlegi adatmennyiség nem elegendő a módszer ellenőrzéséhez, a meglevő adatokra való extrapoláláshoz, viszont megfelelő bontásban rendelkezésre álló historikus adatok esetén alkalmazható lesz a módszer.

A továbbiakban többféle forgatókönyv alapján szeretnénk továbbfejleszteni a becslési modellt. Érdekes lesz megvizsgálni, hogy a hallgatói létszám jelenlegi csökkenési üteme hogyan fogja csökkenteni a kutatói létszámot, valamint a publikációk és a hazai bejegyzésű szabadalmak számát. Egy másik forgatókönyv szerint pedig azt is meg lehetne becsülni, hogy a K+F tevékenység jelenlegi szintjének finanszírozására mekkora összeget kellene fordítani, csökkenő kutatói létszám mellett.

6. Irodalomjegyzék

- [1] J. L. Alba, A. Pujol, and J. J. Villanueva, "Separating geometry from texture to improve face analysis," in *IEEE International Conference on Image Processing*, 2001, vol. 2, pp. 673–676.
- [2] C. H. Lee, D. S. Seong, and K. H. Park, "Face recognition using self-organizing map," *J. Korea Inf. Sci. Soc.*, vol. 20, no. 11, pp. 1730–1738, Nov. 1993.
- [3] Z. Q. Liu, "Adaptive subspace self-organizing map and its applications in face recognition.," *International Journal of Image and Graphics Vol. 2, P. A. ; Garrido Frenich, A. ; Torres, J. A. ; Pulido Bosch, A.*, no. 4, pp. 519–540, 2002.
- [4] I. K., N. S., and U. T., "A self-organizing map based navigation system for an underwater robot," in *2004 IEEE International Conference on Robotics and Automation IEEE Vol. 5*, 2004, p. 5726.
- [5] T. Kohonen, "Self-organizing formation of topologically correct feature maps," *Biol. Cyb.*, vol. 43, no. 1, pp. 59–69, 1982.
- [6] J. M. Pablo-Martí, F.; Arauzo-Carod, "Spatial distribution of economic activities : an empirical approach using self-organizing maps," in *Rethinking the economic region. New possibilities of regional analysis from data at small scale*, E. F.-V. F. Rubiera-Morollón, Ed. Springer, pp. 1–41.
- [7] M. Collan, T. Eklund, and B. Back, "Using the Self-Organizing Map to Visualize and Explore Socio-Economic Development," *EBS Rev. 22(1), 6-15*, vol. 22, no. 22, 2007.
- [8] C. Science, M. Po, T. Honkela, and T. Kohonen, *Bibliography of self-organizing map (som) papers: 2002–2005 addendum*. 2009, pp. 2002–2005.
- [9] Q. K. Al-shayea, G. A. El-refae, and S. F. El-itter, "Neural Networks in Bank Insolvency Prediction," *IJCSNS International Journal of Computer Sciences and Network Security VOL.10 No.5*, vol. 10, no. 5, pp. 240–245, 2010.
- [10] A. Ultsch and F. Röske, "Self-organizing feature maps predicting sea levels," *Self-Organizing Feature Maps Predicting Sea Levels, Inf. Sci. 144/Elsevier*, pp 91 - 125, vol. 144, pp. 91–125, 2002.
- [11] V. Mojtahedzadeh and Z. Abedi, "The Effect of Research and Development(R & D) Expenditures on Firms Value," *International Review of Business Research Papers*, vol. 6, no. 6, pp. 187–200, 2010.
- [12] B. H. Hall, J. Mairesse, and P. Mohnen, "Measuring the Returns to R&D," *National Bureau of Economic Research Working Paper Series*, no. 15622, 2009.

- [13] D. Guellec and B. van P. de la Potterie, “The Impact of Public R&D Expenditure on Business R&D,” *OECD Science, Technology and Industry Working Papers*, vol. 2000/04, p. 27, 2000.
- [14] OECD, “Science, Technology and Innovation in the New Economy,” *OECD Policy Brief*, 2000.
- [15] I. Polónyi, “A hazai felsőoktatás demográfiai összefüggései a 21. század elején,” *Felsőoktatási Kutatóintézet, Kutatás Közben (Educatio Füzetek)*, no. 255, pp. 219–228, 2004.
- [16] J. Polónyi, István; Timár, “Oktatáspolitikai és demográfia,” *Felsőoktatási Kutatóintézet, Kutatás Közben (Educatio Füzetek)*, no. 274, pp. 219–228, 2006.
- [17] K. S. Hivatal, “Népesség, népmozgalom 1949-” 2013. [Online].
https://www.ksh.hu/docs/hun/xstadat/xstadat_hosszu/h_wdsd001a.html.
- [18] K. S. Hivatal, “Oktatás (1960-),” 2013. [Online].
https://www.ksh.hu/docs/hun/xstadat/xstadat_hosszu/h_wdsi001b.html.
- [19] Felvi.hu, “Elmúlt évek statisztikái.” 2013. [Online].
http://www.felvi.hu/felveteli/ponthatarok_rangsorok/elmult_evek.
- [20] Oktatási Hivatal, “Kezdő évfolyamos hallgatók száma képzési szintenként,” 2013. [Online].
http://www.oktatas.hu/felsooktatas/felsooktatasi_statisztikak/!DARI_FelsooktStat/fir_fir_stat2010/stat2010_113.xls.
- [21] Oktatási Hivatal, Felsőoktatási statisztikák 2013.
http://www.oktatas.hu/felsooktatas/felsooktatasi_statisztikak
- [22] SOM Toolbox
<http://www.cis.hut.fi/research/som-research/som.shtml>

7. Függelék

7.1.A 2001-2012-ig rendelkezésre álló adatokból készült SOM programkódja

```
%%Az Excel-táblázat felső, kitöltött része.
minta =
[
1867,4226 127,258      89  5,34617 349,301      2001
2036,5204 125,359      90  5,337353 381,56      2002
2336,3508  130,2       89  5,711302 409,075     2003
2432,8651 128,204      93  5,771648 421,52      2004
2387,1855  125,84      89  5,628018 424,161     2005
2531,1287 124,296      91  6,079358 416,348     2006
2464,7277 123,304      91  6,197392 397,704     2007
2213,9028 125,679      81  5,810265 381,033     2008
1823,6706 127,207      90  4,924434 370,331     2009
1685,9156 121,724      87  4,665642 361,347     2010
1755,8188 117,033      86  4,87966 359,824     2011
]
;
%%Adatstruktúra létrehozása.
D = som_data_struct(minta(:,1:5));
D = som_normalize(D, 'range');
D.labels= minta(:,end);

M = som_make(D);
M.name = 'Felsőoktatási minta';

M.comp_names{1,1} = 'Egyetemisták száma*évek';
M.comp_names{2,1} = 'Korosztály létszáma';
M.comp_names{3,1} = 'Érettségizők létszáma';
M.comp_names{4,1} = 'Eltöltött évek száma';
M.comp_names{5,1} = 'Hallgatók száma';

som_show(M, 'umat', 'all', 'comp', 1:5, 'empty', 'Labels', 'norm', 'd');
som_show_add('label', M, 'subplot', 5);
```

7.2.A 2001-2012-ig rendelkezésre álló adatokból készült SOM programkódja valamint a becsléshez szükséges számítások

a, Első scenárió alapján

%%Az Excel-táblázat felső, kitöltött része.

```
minta =[
1867,4226    127,258           89    5,34617    349,301    2001
2036,5204    125,359           90    5,337353   381,56     2002
2336,3508     130,2           89    5,711302   409,075    2003
2432,8651    128,204           93    5,771648   421,52     2004
2387,1855     125,84           89    5,628018   424,161    2005
2531,1287    124,296           91    6,079358   416,348    2006
2464,7277    123,304           91    6,197392   397,704    2007
2213,9028    125,679           81    5,810265   381,033    2008
1823,6706    127,207           90    4,924434   370,331    2009
1685,9156    121,724           87    4,665642   361,347    2010
1755,8188    117,033           86    4,87966    359,824    2011
1919,6907    115,598           83    5,486477   349,895    2012
1879,8534    112,054       79,87396   5,486477   342,634    2013
1840,0161    105,272       75,03964   5,486477   335,373    2014
1800,1788     100,35       71,53115   5,486477   328,112    2015
1760,3415     97,301       69,35777   5,486477   320,851    2016
1720,5042     94,645       67,46453   5,486477   313,59     2017
1680,6669     97,597       69,56877   5,486477   306,329    2018
1640,8296     97,047       69,17672   5,486477   299,068    2019
1600,9923     96,804       69,0035    5,486477   291,807    2020
1561,155      94,647       67,46596   5,486477   284,546    2021
1521,3177     95,137       67,81524   5,486477   277,285    2022
1481,4803     97,496       69,49677   5,486477   270,024    2023
1441,643      99,871       71,18971   5,486477   262,763    2024
1401,8057     97,613       69,58017   5,486477   255,502    2025
1361,9684     99,149       70,67506   5,486477   248,241    2026
1322,1311     96,442       68,74546   5,486477   240,98     2027
1282,2938     90,335       64,39229   5,486477   233,719    2028
1242,4565     88,049       62,76279   5,486477   226,458    2029
1202,6192     90,269       64,34525   5,486477   219,197    2030
]
```

%%Adatstruktúra létrehozása.

```
D = som_data_struct(minta(:,1:5));
```

```
D = som_normalize(D, 'range');
```

```
D.labels=minta(:,end);
```

```
% adat.labels{1,1} = '2001'
```

```
% adat.labels{2,1} = '2002'
```

```
% adat.labels{3,1} = '2003'
```

```
% adat.labels{4,1} = '2004'
```

```
% adat.labels{5,1} = '2005'
```

```
% adat.labels{6,1} = '2006'
```

```
% adat.labels{7,1} = '2007'
```

```

% adat.labels{8,1} = '2008'
% adat.labels{9,1} = '2009'
% adat.labels{10,1} = '2010'
% adat.labels{11,1} = '2011'
% adat.labels{12,1} = '2012'
M = som_make(D);
M.name = 'Felsőoktatási minta';
% label=['Egyetemisták létszáma*eltöltött_évek'
%       'Korosztály létszáma'
%       'Érettségizők létszáma'
%       'Eltöltött_évek száma'
%       'Egyetemi_hallgatók száma'];

M.comp_names{1,1} = 'Egyetemisták száma*évek';
M.comp_names{2,1} = 'Korosztály létszáma';
M.comp_names{3,1} = 'Érettségizők létszáma';
M.comp_names{4,1} = 'Eltöltött évek száma';

M.comp_names{5,1} = 'Hallgatók száma';
som_show(M,'umat','all','comp',1:5,'empty','Labels','norm','d');
som_show_add('label',M,'subplot',5);
[bmus,qerrs]=som_bmus(M,D,[1,2,3,4]);
for i=1:30
index=bmus(i,1:4);
sulyok=M.codebook(index,:);
ats(i,:)=mean(sulyok);
end
eredmeny=som_denormalize(ats,M.comp_norm);

```

b, Második szcenárió alapján

%%Az Excel-táblázat felső, kitöltött része.

```

minta = [
1867,423      127,258      89      5,34617      349,301      2001
2036,52       125,359      90      5,337353      381,56      2002
2336,351       130,2       89      5,711302      409,075      2003
2432,865       128,204      93      5,771648      421,52      2004
2387,186       125,84      89      5,628018      424,161      2005
2531,129       124,296      91      6,079358      416,348      2006
2464,728       123,304      91      6,197392      397,704      2007
2213,903       125,679      81      5,810265      381,033      2008
1823,671       127,207      90      4,924434      370,331      2009
1685,916       121,724      87      4,665642      361,347      2010
1755,819       117,033      86      4,87966      359,824      2011
1823,493       115,598      83      5,211544      349,895      2012
1767,885       112,054      79,87396      5,159689      342,634      2013
1713,202       105,272      75,03964      5,108349      335,373      2014
1659,433       100,35      71,53115      5,05752      328,112      2015
1606,564       97,301      69,35777      5,007197      320,851      2016
1554,583       94,645      67,46453      4,957374      313,59      2017
1503,477       97,597      69,56877      4,908048      306,329      2018
1453,235       97,047      69,17672      4,859212      299,068      2019
1403,843       96,804      69,0035      4,810862      291,807      2020
1355,291       94,647      67,46596      4,762993      284,546      2021

```

1307,565	95,137	67,81524	4,7156	277,285	2022
1260,655	97,496	69,49677	4,668679	270,024	2023
1214,55	99,871	71,18971	4,622225	262,763	2024
1169,237	97,613	69,58017	4,576233	255,502	2025
1124,705	99,149	70,67506	4,530699	248,241	2026
1080,944	96,442	68,74546	4,485618	240,98	2027
1037,943	90,335	64,39229	4,440985	233,719	2028
995,6898	88,049	62,76279	4,396797	226,458	2029
954,175	90,269	64,34525	4,353048	219,197	2030

]

```
%%Adatstruktúra létrehozása.
D = som_data_struct(minta(:,1:5));
D = som_normalize(D, 'range');
D.labels=minta(:,end);
% adat.labels{1,1} = '2001'
% adat.labels{2,1} = '2002'
% adat.labels{3,1} = '2003'
% adat.labels{4,1} = '2004'
% adat.labels{5,1} = '2005'
% adat.labels{6,1} = '2006'
% adat.labels{7,1} = '2007'
% adat.labels{8,1} = '2008'
% adat.labels{9,1} = '2009'
% adat.labels{10,1} = '2010'
% adat.labels{11,1} = '2011'
% adat.labels{12,1} = '2012'
M = som_make(D);
M.name = 'Felsőoktatási minta';
% label=['Egyetemisták létszáma*eltöltött_évek'
%       'Korosztály létszáma'
%       'Érettségizők létszáma'
%       'Eltöltött_évek száma'
%       'Egyetemi_hallgatók száma'];

M.comp_names{1,1} = 'Egyetemisták száma*évek';
M.comp_names{2,1} = 'Korosztály létszáma';
M.comp_names{3,1} = 'Érettségizők létszáma';
M.comp_names{4,1} = 'Eltöltött évek száma';

M.comp_names{5,1} = 'Hallgatók száma';
som_show(M, 'umat', 'all', 'comp', 1:5, 'empty', 'Labels', 'norm', 'd');
som_show_add('label', M, 'subplot', 5);
[bmus, qerrs]=som_bmus(M, D, [1,2,3,4]);
for i=1:30
    index=bmus(i,1:4);
    sulyok=M.codebook(index,:);
    ats(i,:)=mean(sulyok);
end
eredmeny=som_denormalize(ats,M.comp_norm);
```

7.3.A súlypontok és BMU-k felrajzolásának programkódja

```
%Az Excel-táblázat felső, kitöltött része.
%Az Excel-táblázat felső, kitöltött része.
minta = [
1867,4226    127,258          89      5,34617      349,301
2036,5204    125,359          90      5,337353     381,56
2336,3508     130,2           89      5,711302     409,075
2432,8651    128,204          93      5,771648      421,52
2387,1855     125,84          89      5,628018     424,161
2531,1287    124,296          91      6,079358     416,348
2464,7277    123,304          91      6,197392     397,704
2213,9028    125,679          81      5,810265     381,033
1823,6706    127,207          90      4,924434     370,331
1685,9156    121,724          87      4,665642     361,347
1755,8188    117,033          86      4,87966      359,824
]
%%Adatstruktúra létrehozása.
adat = som_data_struct(minta);
adat = som_normalize(adat);
-----
a,Ha a tanítás előtti adatpontokra vagyunk kíváncsiak
terkep = som_randinit(adat);

b,Ha a tanítás utáni adatpontokra vagyunk kíváncsiak
terkep=som_make(adat);
-----
terkep.name = 'Felsőoktatási minta';
M = som_denormalize(terkep.codebook,terkep);
%%Vektorok felrajzolása.
som_grid(terkep,'coord',M(:,2:4),'markercolor',[1 0 0],'linecolor',[0 0
0])
holdon, gridon
D = som_denormalize(adat.data,adat);
%%Rács megjelenítése.
cimke = ['2001'
'2002'
'2003'
'2004'
'2005'
'2006'
'2007'
'2008'
'2009'
'2010'
'2011'
'2012']
som_grid('rect',[12 1],'coord',D(:,2:4),'label',cimke,'labelcolor','b')
xlabel('Korosztály');
ylabel('Érettségizők létszáma');
zlabel('Eltöltött évek');
```

7.4.A már kigyűjtött, a becslések elvégzéséhez rendelkezésre álló adatok

Év	18 éves korosztály (ezer fő)	Felsőoktatásban hallgató (ezer fő)	Felsőfokú oklevelet szerzett (ezer fő)	Kezdő évfolyamos hallgatók száma összesen (ezer fő)	Felvettek száma összesen (ezer fő)
2001	127,258	349,301	47	128,237	98,031
2002	125,359	381,56	51	138,642	109,355
2003	130,2	409,075	53	144,998	106,376
2004	128,204	421,52	54	144,176	109,851
2005	125,84	424,161	57	137,781	103,364
2006	124,296	416,348	53	123,565	94,142
2007	123,304	397,704	51	107,606	81,637
2008	125,679	381,033	49	106,514	96,33
2009	127,207	370,331	53	116,793	109,336
2010	121,724	361,347	53	126,055	114,068
2011	117,033	359,824	50		115,841

7.5. Becsült adatok

Év	18 éves korosztály(ezer fő)	Felsőoktatásban hallgató (ezer fő)	Felsőfokú oklevelet szerzett (ezer fő)
2012	115,598	349,895	51
2013	112,054	342,634	47,3695
2014	105,272	335,373	46,81
2015	100,35	328,112	46,2505
2016	97,301	320,851	45,691
2017	94,645	313,59	45,1315
2018	97,597	306,329	44,572
2019	97,047	299,068	44,0125
2020	96,804	291,807	43,453
2021	94,647	284,546	42,8935
2022	95,137	277,285	42,334
2023	97,496	270,024	41,7745
2024	99,871	262,763	41,215
2025	97,613	255,502	40,6555
2026	99,149	248,241	40,096
2027	96,442	240,98	39,5365
2028	90,335	233,719	38,977
2029	88,049	226,458	38,4175
2030	90,269	219,197	37,858

Év	Eltöltött évek száma(állandó)	Eltöltött évek száma(becsült)	Hallgatók- felvettek(becsült)
2012	5,486477	5,211544	265,788769
2013	5,486477	5,159689	244,411876
2014	5,486477	5,108349	239,121816
2015	5,486477	5,05752	233,912832
2016	5,486477	5,007197	228,783834
2017	5,486477	4,957374	223,733746
2018	5,486477	4,908048	218,761505
2019	5,486477	4,859212	213,866062
2020	5,486477	4,810862	209,046382
2021	5,486477	4,762993	204,301442
2022	5,486477	4,7156	199,63023
2023	5,486477	4,668679	195,03175
2024	5,486477	4,622225	190,505016
2025	5,486477	4,576233	186,049057
2026	5,486477	4,530699	181,662912
2027	5,486477	4,485618	177,345632
2028	5,486477	4,440985	173,096283
2029	5,486477	4,396797	168,913938
2030	5,486477	4,353048	164,797686

Év	Érettségizett	Hallgatók*évek (Exponenciális)	Hallgatók*évek (Átlaggal)
2012	83	1823,493	1919,6907

2013	79,87395871	1767,885	1879,8534
2014	75,03963608	1713,202	1840,0161
2015	71,53115245	1659,433	1800,1788
2016	69,35777443	1606,564	1760,3415
2017	67,46453337	1554,583	1720,5042
2018	69,56876817	1503,477	1680,6669
2019	69,176719	1453,235	1640,8296
2020	69,00350455	1403,843	1600,9923
2021	67,465959	1355,291	1561,155
2022	67,81523917	1307,565	1521,3177
2023	69,49677368	1260,655	1481,4803
2024	71,18971327	1214,55	1441,643
2025	69,58017323	1169,237	1401,8057
2026	70,67505963	1124,705	1361,9684
2027	68,74546492	1080,944	1322,1311
2028	64,39229354	1037,943	1282,2938
2029	62,76279464	995,6898	1242,4565
2030	64,34524764	954,175	1202,6192

7.6. Első forgatókönyv

Egyetemisták létszáma*eltöltött évek száma	Korosztály létszáma	Érettségizők létszáma	Eltöltött évek száma	Egyetemi hallgatók száma
1634.215	109.0574	78.1512	5.115789	319.4453408
1526.605	102.6369	73.27104	5.071436	301.020318
1487.671	100.5285	71.70699	5.064126	293.7666783
1443.121	98.64039	70.32858	5.05982	285.2118728
1364.221	97.05002	69.17957	5.058103	269.710039
1364.221	97.05002	69.17957	5.058103	269.710039
1364.221	97.05002	69.17957	5.058103	269.710039
1268.776	96.67033	68.90824	5.058029	250.8439613
1268.776	96.67033	68.90824	5.058029	250.8439613
1165.766	96.51508	68.79756	5.058026	230.4783995
1165.766	96.51508	68.79756	5.058026	230.4783995
1117.207	96.58031	68.84406	5.058026	220.8779895
1061.661	96.08676	68.49224	5.058026	209.8963622
1061.661	96.08676	68.49224	5.058026	209.8963622
969.6043	94.53602	67.38685	5.058026	191.6961831
969.6043	94.53602	67.38685	5.058026	191.6961831
969.6043	94.53602	67.38685	5.058026	191.6961831
969.6043	94.53602	67.38685	5.058026	191.6961831

7.7. Második forgatókönyv

Egyetemisták létszáma*eltöltött évek száma	Korosztály létszáma	Érettségizők létszáma	Eltöltött évek száma	Egyetemi hallgatók száma
1681.594542	113.6401062	80.7917164	5.164216921	325.6243043
1540.824202	107.1812483	75.90926993	5.062469518	304.3621688
1467.946136	101.0107704	71.80544007	5.012206793	292.8742162
1425.168607	99.54044936	70.82889594	4.980975264	286.1224021
1378.255203	97.85348839	69.73074046	4.946780894	278.6165857
1378.255203	97.85348839	69.73074046	4.946780894	278.6165857
1324.918551	97.59804903	69.54349042	4.906970577	270.0074375
1261.803055	96.87517419	69.04022033	4.859705225	259.6460066
1183.097238	96.75923254	68.95533567	4.800146153	246.4710867
1128.528902	97.41620126	69.42418114	4.758202905	237.1754473
1056.14297	97.38105216	69.41889614	4.702913473	224.5720608
1001.075873	97.25091479	69.32725223	4.660472598	214.8013644
974.4747843	95.94824772	68.3991542	4.64032999	210.0011823
973.9541578	96.60062303	68.86666908	4.639559632	209.9238365
915.9327663	94.85588199	67.62355348	4.59498026	199.3333408
915.9327663	94.85588199	67.62355348	4.59498026	199.3333408
915.9327663	94.85588199	67.62355348	4.59498026	199.3333408
915.9327663	94.85588199	67.62355348	4.59498026	199.3333408