A logo of a bird

AI-generated content may be incorrect.

Budapesti Corvinus Egyetem

Tudományos Diákköri Konferencia

2025

**EU\_OKTATÁS**

EU\_EDUCATION

Molnár Marcell

Gazdaság -és pénzügy matematikai elemzés MSc

III. évfolyam

Témavezető: Madari Zoltán

Absztrakt

Tartalomjegyzék

[I. Bevezetés 5](#_Toc208229770)

[II. Szakirodalmi áttekintés 6](#_Toc208229771)

[III. Módszertan 6](#_Toc208229772)

[III. 1. Szigma és Béta konvergencia 6](#_Toc208229773)

[III. 2. Térbeli autokorreláció 7](#_Toc208229774)

[III. 2. 1. Globális Moran’s I 7](#_Toc208229775)

[III. 2. 2. Geary’s C 8](#_Toc208229776)

[III. 2. 3. LISA (Local Indicators of Spatial Association) 8](#_Toc208229777)

[III. 3. Lineáris panel modellek 9](#_Toc208229778)

[III. 3. 1. Pooled panel modell 9](#_Toc208229779)

[III. 3. 2. Fixhatású panel modell 9](#_Toc208229780)

[III. 4. Térbeli panel modellek 10](#_Toc208229781)

[III. 4. 1. Specifikációs tesztek 11](#_Toc208229782)

[III. 5. Egyéb machine learning modellek 12](#_Toc208229783)

[III. 5. 1. Oktatási profilok nem felügyelt klaszterezése 12](#_Toc208229784)

[III. 5. 2. Véletlen erdő 13](#_Toc208229785)

[III. 5. 3. Propensity Score Matching 15](#_Toc208229786)

[III. 5. 4. Vektor Autoregresszió 15](#_Toc208229787)

[IV. Adatok bemutatása 15](#_Toc208229788)

[V. Eredmények 17](#_Toc208229789)

[V. I. Szomszédsági viszonyok definiálása 17](#_Toc208229790)

[V. II. Modellek eredményei 18](#_Toc208229791)

[VI. Kitekintés, Magyarország helyzete 26](#_Toc208229792)

[VII. Korlátok 26](#_Toc208229793)

[VIII. Konklúzió 26](#_Toc208229794)

[Irodalomjegyzék 27](#_Toc208229795)

[Mellékletek 28](#_Toc208229796)

Táblázatok jegyzéke

[1. táblázat Local Moran's I besorolások 8](#_Toc207829511)

[2. táblázat Területi autokorreláció mérőszámok 22](#_Toc207829512)

Ábrajegyzék

[1. ábra Klaszterszám kiválasztása 12](#_Toc207829521)

[2. ábra Szomszédsági hálózat 17](#_Toc207829522)

[3. ábra Szigma konvergencia alakulása 18](#_Toc207829523)

[4. ábra Béta konvergencia a felsőfokú végzettséget szerzők között 19](#_Toc207829524)

[5. ábra Felsőoktatási végzettséggel rendelkezők aránya (2023) 21](#_Toc207829525)

[6. ábra Local Moran's I osztályozás 23](#_Toc207829526)

# I. Bevezetés

Regionális eltérések az oktatásban kimondottan fontos problémának mondható az Európai Unióban, mivel a képzett munkaerő hiánya nagyban befolyásolhatja egy régió gazdasági növekedésének sebességét, a munkaerőpiacának jellegét, illetve a társadalmi különbségekre is nagy hatása van. Az oktatás széles körökben elismert kulcsfontosságú eleme a fejlődésnek és hosszútávú növekedésnek. A klasszikus növekedéselméleti modellek, szintúgy empirikus kutatások kiemelik a magas szintű, jó minőségű oktatás elhanyagolhatatlan szerepét a fejlődésben. A felsőoktatás helyzet általában a gazdasági produktivitással, illetve jövedelmi szintekkel, míg a középszintű oktatás minősége a szociális és gazdasági faktorokra van hatással.

Ennek okán fontos megérteni az esetleges területi mintázatokat és trendeket az Unió területén, így olyan módon lehet az erősforrásokat allokálni, illetve különböző rendelkezéséket hozni, hogy az a felzárkózást és együttes növekedést segítse.

A kutatásom három fő oktatással kapcsolatos indikátort fogok figyelembe venni az Európai Unióban, NUTS 2-es szinten: a korai iskolaelhagyókat, a felsőfokú végzettség arányát, illetve az oktatásban/képzésben való részvételi arányt. Ezeket az indikátorokat férfi-női bontásban, illetve nemtől függetlenül is elemzem 2013-2023-ig terjedő időszakban.

A kutatásom a következő öt fő kérdésre keresi a választ:

1. Az Európai Unió régiói idővel elkezdenek-e egyre jobban egymásra hasonlítani?
2. Milyen a területi autokorreláció a regionális oktatási szinteken?
3. Milyen nagyok a nemi eltérések, és hol a legsúlyosabbak?
4. Megtalálhatóak egymástól távol eső régió klaszterek, ugyanolyan oktatási helyzettel?
5. Hogyan kapcsolódnak az oktatási indikátorok a regionális szociális és gazdasági dimenziókhoz, mint például a GDP, munkavállalás és innováció?

Ezen kérdések megválaszolására komplex módszertani keretrendszert alkalmazva keresek válaszokat.

…

# II. Szakirodalmi áttekintés

Kell még

# III. Módszertan

Ebben a pontban a választott módszertanokat részletezem, melyek egy, vagy több feltett kutatási kérdéshez kapcsolódnak. Konvergencia analízis, területi autokorreláció, ökonometriai modellek és gépi tanulási módszerek lesznek részletezve a következőkben.

## III. 1. Szigma és Béta konvergencia

A konvergencia megállapítására az oktatással kapcsolatos metrikáknál a szigma, szintúgy a béta konvergenciát figyelembe veszem. A szigma konvergencia azt mutatja meg, hogy egyes régiók az idő előrehaladtával hasonlóbbá válnak az adott változóra nézve. Ha egy adott indikátornak a szórása a régiók között kisebb az utolsó vizsgált időpontban, ebben az esetben 2023-ban, mint az elsőben, 2013-ban, akkor jelen van szigma konvergencia.

(1)

Fontos kiemelni, hogy a szigma konvergencia egy kimeneteli mérőszám, elmondja mi történik, de azt nem, hogy miért.

A béta konvergencia (a gyengébb régiók gyorsabban fejlődnek) szükséges, de nem elégséges feltétele a szigma konvergenciának, azaz a béta konvergencia jelen lehet szigma nélkül, ennek általában valamilyen előre nem látott sokk, exogán hatás lehet az oka. Ugyanakkor fontos eleme a kutatásnak, hogy vajon a gyengébb oktatással rendelkező régiók vajon felzárkóznak-e a fejletteb régiókhoz.

A béta konvergenciához a klasszikus Barro és Sala-i-Martin keresztmetszeti megközelítést használtam, amely a következőképpen néz ki:

(2)

ahol,

* Yi, t0  az i-edik régióban az adott oktatási indikátor értéke az első évben (2013)
* Yi, t1 ugyanazon indikátor régió pár értéke az utolsó évben (2023).
* alfa: konstans
* béta: konvergencia koefficiens
* ε: hibatag
* baloldal a relatív növekedés a vizsgált időszak alatt.

A regresszió bétáit a következőképpen kell értelmezni:

* β < 0: abszulút béta konvergencia, a kezdetben alacsonyabb értékkel rendelkező régiók gyorsabban nőnek, azaz utolérik a kezdetben magasabb értékkel rendelkezőket.
* β = 0: nincs konvergencia
* β > 0: divergencia

Fontos kérdés lehet a következő: ha az eredmény azt indikálja, hogy van szignifikáns negatív béta, azaz béta konvergencia, akkor az “utolérés” folyamata megkezdődött. Ugyanakkor, ha ezzel egyidejűleg nem figyelhető meg szigma konvergencia, akkor lehet, hogy relatív módon elkezdődött az utolérés, de az abszulút különbséget nem sikerült eliminálni. Ezt észben kell tartani a további eredmények kiértékelésénél.

## III. 2. Térbeli autokorreláció

A térbeli autokorreláció vizsgálata segít feltárni esetleges mintázatokat az Európai Unió oktatási helyzetéve kapcsolatban, legfőképpen azt, hogy a magas és alacsony értékek egy helyen csoportosulnak-e, vagy megfigyelhetőek-e kilógó régiók. Kétféle módon vizsgálom ezt, globális és lokális mérőszámokkal.

### III. 2. 1. Globális Moran’s I

Az első globális mérőszám a Moran’s I statisztika, ami a következő képlettel vizsgálja, hogy a szomszédos régiók mennyire egyeznek:

(3)

ahol,

* N: a régiók száma
* xi: az i-edik régióban vett változó értéke
* : a változó átlaga
* : a súlymátrix
* : a súlymátrix i-edik sorának j-edik eleme

A metrikát a következőképpen kell értelmezni:

* I > 0: pozitív területi autokorreláció
* I = 0: nincs területi autokorreláció
* I < 0: negatív területi autokorreláció

### III. 2. 2. Geary’s C

A második globális mérőszám a Geary’s C, amely annyiban különbözik a Moran’s I-től, hogy míg a korábbi a főátlaghoz képesti eltérést vizsgálja, addig az utóbbi sokkal szenzitívebb a szomszédos régiók kapcsolatát tekintve, ami a képletből is látszódik:

(4)

ahol minden jelölés ugyanaz, mint a Moran’s I-nél.

A metrika 0 és 2 között vehet fel értékeket, amit a következő módon kell értelmezni:

* C < 1: pozitív területi autokorreláció
* C = 1: nincs területi autokorreláció
* C > 1: negatív területi autokorreláció

### III. 2. 3. LISA (Local Indicators of Spatial Association)

Ugyan a Moran’s I és Geary’s C mutatók megmutatják, hogy milyen jellegű az autokorreláció régiók között, a pontos helyüket nem mutatja. Azért, hogy a lokális mintázatokat megfigyeljük, úgynevezett LISA-t használtam, pontosabb az Anselin által bevezetett Lokális Moran’s I-t. Ezen mutató minden i-edik régióra kiszámol egy értéket, ami megmondja, hogy az része-e egy nagyobb lokális klaszternek, vagy sem, mégpedig a következő képlettel:

(5)

ahol minden ugyanaz, mint a Globális Moran’s I-nél.

A kapott értékek alapján ezután a következő táblázat alapján 5 kategóriába sorolhatóak a régiók:

1. táblázat Local Moran's I besorolások

|  |  |  |  |
| --- | --- | --- | --- |
| Kategória | Leírás | Lokális Moran’s I értéke | Értelmezés |
| Magas - Magas | Magas, magasak mellett | Pozitív és szignifikáns | Hot-spot |
| Alacsony - Alacsony | Alacsony, alacsonyak mellett | Pozitív és szignifikáns | Cold-spot |
| Magas – Alacsony | Magas, alacsonyak mellett | Negatív és szignifikáns | Outlier |
| Alacsony – Magas | Alacsony, magasak mellett | Negatív és szignifikáns | Outlier |
| Nem szignifikáns | Nincs mintázat | Közel nulla | Nincs területi autokorreláció |

## III. 3. Lineáris panel modellek

Miután körbejártam a területi autokorreláció kérdéskörét szigorúan az oktatással kapcsolatban a következő lépése a kutatásnak, hogy megvizsgáljam az oktatási és a különböző szocio-ökonometriai változók kapcsolatát. A regressziós modelleket két különböző függő változóra is megépítettem: a GDP/fő-re és a foglalkoztatási rátára is.

### III. 3. 1. Pooled panel modell

Először egy egyszerű pooled panel modellt építettem, ami az idősoros adatokat úgy kezeli, mintha keresztmetszeti adatokkal dolgoznék, régió specifikus hatásoktól függetlenül. Ennél a modellnél, egy szimpla OLS regressziót futtattam minden fix, illetve randomhatást figyelmen kívül hagyva:

(6)

### III. 3. 2. Fixhatású panel modell

Az modellezés során fennállhat annak a lehetősége, hogy egy-egy kihagyott változó miatt torzul a becslés, amelyet a nem megfigyelt változókban és hatásokban fennálló heterogenitás okozhatja. A fixhatású modell képes ezt kezelni, amennyiben a nem megfigyelt hatások időben állandónak tekinthetőek. Ez annak köszönthető, mivel a pooled modellel ellentétben a becslés előtt alkalmaz egy transzformációt, ami eltávolítja ezen hatásokat, ezeket jelölje zi. Ezen fixhatás, vagy más néven within transzformációval egyidőben a mefigyelt, de időben állandó változók is kiesnek. Ez a következőképpen vezethető le matematikailag:

(7)

ezek időben kiátlagolva:

(8)

Most vegyük a két egyenlet különbségét, azaz vonjuk ki az eredeti értékekből az idősoros átlagokat:

(9)

ahol az új jelölések a különbségeket jelölik. Látható, hogy a zi tag eltűnt, így már lehet OLS elven becsülni. Ezen becslőket fixhatás, vagy within becslőnek nevezzük. „A szigorú exogenitási feltétel fennállása esetén a within módszer torzítatlan becslést eredményez” (Wooldridge, 2013, 485. old.). Általános esetben a következőképpen alakul a felírás:

(10)

ahol a korábbi levezetés alapján, kiesne a zi tag, de ez csökkentené a magyarázott varianciát, így a zi-t minden paraméterre meg szokták becsülni, tehát a tengelymetszet helyett, minden megfigyelési egységhez saját tengelymetszet tartozik, amelyek a köztük fellelhető különbséget hivatottak megmutatni. Ennek okán gyakran alkalmaznak fixhatás modelleket területi egységek bizonyos, több időszakban megfigyelt mutatóinak vizsgálatára. Tehát a modell formulája annyiban különbözik csupán a szimpla pooled panel struktúrától, hogy a konstans is i indexen fut:

(11)

## III. 4. Térbeli panel modellek

A fentebbi lineáris modellek azzal a feltevéssel működnek, hogy egy-egy régió vizsgált értéke csak a saját karakterisztikájának függvénye (és különböző fixhatásoké). Ugyanakkor korábbi, a feldolgozott szakirodalomban is bemutatott tanulmányok és empirikus kutatások arra engednek következtetni, hogy területi átgyűrűzések megfigyelhetőek. Például egyes régiók oktatási fejlődése hatással lehet akár a környező területi egységek gazdasására is. Ebben a dolgozatban három olyan modellt jártam körbe, amelyek ezt a régiók közötti hatást képesek lehetnek leírni.

Az első modell a térbeli autoregresszív (továbbiakban SAR) modell, ami a következőképpen írható fel:

(12)

ahol,

* a térbeli autokorrelációs együttható
* : a térbeli súlymátrix egy eleme
* : a vizsgált függő változó térben késleltetett értékei

A második a térbeli hibamodell (továbbiakban SEM), melyben a térbeli interakció a hibatagban található meg és a következőképpen írható fel, a korábbi jelölésekkel:

(13)

ahol,

(14)

A harmadik modell pedig a LeSage és Pace által leírt, térbeli Durbin modell (továbbiakban SDM) vezetem be. Ezen modell egyidőben tartalmazza a térben késleltetését a függőváltozónak és magyarázóváltozót is, így egyfajta kiterjesztése a korábbi SAR modellnek, amely így írható fel a korábbi jelöléseket követve:

(15)

ahol,

* a -hoz hasonló együtthatóvektor

### III. 4. 1. Specifikációs tesztek

A megfelelő modell kiválasztásához meg kell vizsgálni, hogy milyen típusú autokorreláció van jelen, a hibatagban jelenik meg a hatás, vagy a területi késleltetett értékek hatnak egymásra. Az erre alkalmas tesztek LM-teszten alapulnak, amely tesztek az autoregresszív esetben a ρ-ra, míg a hibatag autokorrelációjának vizsgálatakor a 𝜆 paraméterre irányulnak. A tesztek nullhipotézise, hogy nem áll fenn térbeli autokorreláció. Ezek próbákat Anselin határozta meg.

Az statisztika endogén térbeli változóként teszteli a késleltetett értékeket, vagyis azt vizsgálja, hogy a SAR struktúra szükséges-e? A próba feltételezi, hogy 𝜆 = 0, azaz a hibatagok között nincs területi autokorreláció. A statisztika Elhornst alapján:

(16)

ahol U a fixhatású panel hibatagjainak vektora, I az egységmátrix, a Kronecker-szorzat jele, Y az eredményváltozók vektora, W a súlymátrix és a hibavariancia.

J pedig a következő: amelyben , ahol tr a mátrix nyoma. A nullhipotézis elvetése a SAR modell választását indikálja.

Az statisztika pedig azt vizsgálja, hogy a hibákban jelentkezik-e térbeli autokorreláció, feltéve, hogy ρ = 0. Ha ezen tesztnél választjuk az alternatív hipotézist, akkor a SEM modellt kell alkalmazni. Formálisan a következőképpen írható fel a teszt:

(17)

Amennyiben mindkét tesztnél elutasítva lesz a nullhipotézis, azaz az egyik szerint a SAR a másik szerint a SEM alkalmazandó, akkor szükséges a próbák lokálisan robosztus változatának elvégzése. Az 𝑅𝐿𝑀𝜆 az autoregresszív struktúrát bevonva teszteli a reziduális autokorreláció fennállását, az 𝑅𝐿𝑀𝜌 pedig a térbeli hiba modellben (SEM) nézi meg, hogy az eredményváltozó területileg késleltetett értékeinek megfigyelhető-e szignifikáns hatása – lásd: Elhorst (2010).

….

## III. 5. Egyéb machine learning modellek

A megszokott statisztikai módszerek mellett alkalmaztam két gépi tanulási módszertant is, az egyik az oktatási összetétel alapján történő nem felügyelt, k-means klaszterezés, a másik pedig egy random forest volt, ami az oktatási változók prediktív minőségét vizsgálta.

### III. 5. 1. Oktatási profilok nem felügyelt klaszterezése

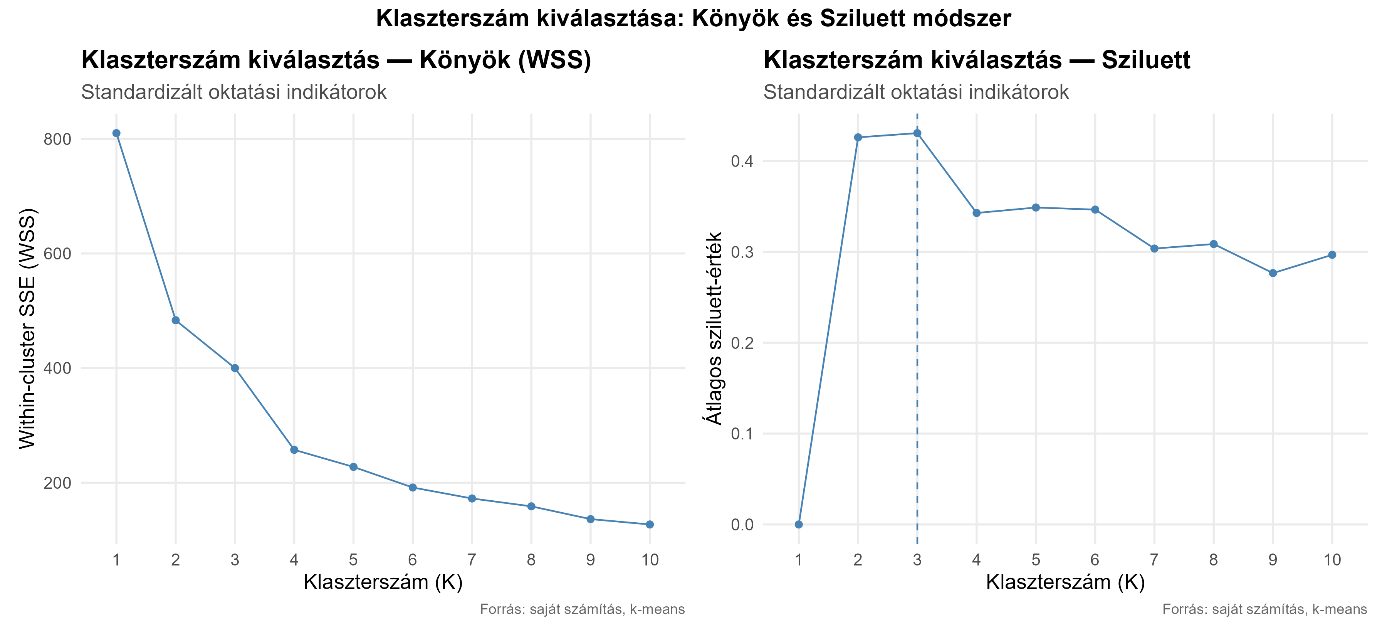
A területi autokorreláció alapján történő besorolás mellett alkalmaztam egy nem felügyelt k-means klaszterezési eljárást is. A model futtatása előtt meg kell tisztítani az adatbázist, minden nem oktatással kapcsolatos indikátor, azonosítót, időbélyeget, továbbá standardizálni kell az adatokat, hogy azok azonos skálán összehasonlíthatóak legyenek:

(18)

* : a i-edik régió j változójának eredeti értéke
* : az átlaga és mintabeli szórása a j változónak

Hogy eldöntsem a megfelelő klaszterek számát, az elbow-módszert (klaszteren belüli szórások, WSS, 1. ábra, bal oldal), illetve az átlagos sziluett együtthatót vizsgáltam (1. ábra, jobb oldal).

1. ábra Klaszterszám kiválasztása



A könyök módszernél azt a klaszterszámot keresem, ahonnan a WSS csak kis mértékben csökken, ez k = 4-nál van. A sziluette módszernál pedig a legnagyobb értéket felvevő klasztert választjuk, ami a k = 3.

A klaszterezés stabilitásának megállapításához replikálom a k-means klaszterezési eljárást 50 alkalommal véletlenszerűen, majd kiszámolom az Adjusted Rand Indexet (továbbiakban ARI) a következő módon:

(19)

RI a Rand Index, ami két partíció közötti egyezést méri. Egy átlagos ARI 1 körül nagyon stabil klaszterezést jelent.

### III. 5. 2. Véletlen erdő

A másik alkalmazott gépi tanulási módszer a Random Forest, vagy Véleltlen erdő (továbbiakban RF) regresszió, amely egy döntési fákon alapuló módszertan. Az RF lényege, hogy sok száz, néha akár ezer fát épít fel, különböző bootstrap mintákon majd ezen fák választott csoportjának becsült értékeit átlagolja. Ezen úgynevezett bagging technika csökkenti a varianciát és ellenállóvá teszi a modellt az túlillesztéssel szemben. Ha B darab döntési fát építünk, akkor x megfigyelés becsült értéke a következő:

(20)

Tb(x) a b-edik fa előrejelzése.

A modell a következő 3 fő hiperparaméterrel optimalizáltam:

* épített fák száma
* minden vágásnál a változók száma
* minimális levélméret

Az egyes fák építésekor a magyarázóváltozók csak egy véletlenszerű részhalmaza lesz figyelembe véve, így csökkentve a fák hasonlóságát és egyes változók esetleges túlsúlyát. A RF modellek ereje, hogy nincs linearitási feltevésük, képesek komplexebb, nem lineáris struktúrákat, illetve a változók közötti kölcsönhatásokat is kezelni. Továbbá a modell out-of-bag (OOB) megfigyelésekkel keresztvalidál, azaz minden fa építésénél a bootstrap-fa minta egy része kimarad, ezek szolgálnak az OOB hibaarány becslésre. Ez egy torzítatlan mérőszám a modell általánosíthatósgára.

A RF modelleknél mérhető a változók fontossága is. A permutációs fontosság definíciója alapján rangsorolom a változókat, azaz adott X változót az OOB-ben permutáljuk és mérjük a modell hibájának növekedését, formálisan:

(21)

MSEOOB az OOB hiba, míg a permutáció utáni hiba. Minél nagyobb a különbség, annál fontosabbnak tekinthető a változó.

A modell teljesítményét nem csak OOB hibákkal vizsgáltam, hanem egy időalapú tanuló-teszt adatszeparálást alkalmaztam. A tanulóhalmaz 2013-2019 közötti értékeket tartalmazza, míg a teszthalmaz 2020-2023 közötti adatokat. Ezen felosztás lehetővé teszi, hogy a modell általánosító képességét, valódi előrejelzési környeztben vizsgáljam és a COVID a covid utáni strutúrális töréseket is figyelembe vegyem. A teszthalmazon a teljesítményt Root Mean Squared Error (RMSE):

(22)

és determinációs együttható alapján értékelem:

(23)

A modellt a dolgozatban két függő változóra fogom alkalmazni: a régiók GDP-jére és a foglalkoztatási rátára.

### III. 5. 3. Propensity Score Matching

A Propensity Score Matching (továbbiakban PSM) módszer célja, hogy kezelje a megfigyelt adatokban előforduló szelekciós torzítást, amikor a kezelési változó nem véletlenszerűen, hanem valamilyen háttérban levő hatás miatt alakul valamilyen módon az eloszlása. A módszer Rosenbaum és Rubin (1983) alapján, arra feltevésre alapszik, hogyha megfelelő kovariánsokra kondíciónálunk, akkor a kezelési hatás torzítatlanul becsülhető. Ezt a kondícionálást egy úgyneveztt propensity scorerral lehet megvalósítani. Az eljárás a következő 4 lépésből áll:

1. Lineáris regresszió

A kezelt és kontroll csoportok(például nők = 1 és férfiak = 0) átlagos kimeneti különbségét egyszerű lineáris regresszióval becsüljük, így kiszűrhetjük a nem szignifikáns változókat.

1. Propensity score becslése, logisztikus regresszióval

A propensity score definíció szerint a következő:

(24)

azaz az i-edik egység valószínűsége, hogy kezelést kap, a háttérváltozók függvényében. Erre a korábban szűrt változókkal történő logisztikus regresszió alapú becslést alkalmazom:

(25)

Mivel a kezelés bináris változóként van definiálva , ezért a módszer közvetlenül alkalmazható olyan dichotóm változókra is, mint például a férfi/női különbségek vizsgálata.

1. Matching

Miután minden egységhez hozzá lett rendelve egy propesity score, a kezelt egységeket egy hasonló értékű kontrollhoz rendeljük.

A hozzárendelési szabályoknak több módszere is van:

* legközelebbi szomszéd
* caliper: csak adott távolságon belüli illesztés
* kernel vagy rétegzett illesztés

1. Kezelési hatás értékelése

A kezelés hatása a kezelt és a hozzájuk illesztett kontroll változók kimeneti különbségével becsült. Ezen különbségek szignifikanciáját t-próbával vizsgálom, melynek nullhipotézise szerint nincs különbség kezelt és illesztett értékek között.

Eszerint, ha a nullhipotézist elvetése mellett kell dönteni, akkor számít a kezelés hatása, vagyis jelen dolgozat esetében számítani fog a nem változó hatása az oktatási profilok tekintetében.

### III. 5. 4. Vektor Autoregresszió

Sima VAR és SVAR

# IV. Adatok bemutatása

Minden használt adat az Eurostat adatbázisából lett letöltve és a mellékelt R kóddal lettek egy nagy adatbázissá szerkesztve. NUTS 2-es szintű területi besorolású adatokkal dolgoztam minden esetben, ami egy Európai Unió szerte egységes klasszifikálása a régióknak, így könnyítve az elemzések elvégzését.

Oktatással kapcsolatban a korábban említett 3 fő indikátornak, a korai iskolaelhagyóknak, a felsőfokú végzettség arányának, illetve az oktatásban/képzésben való részvételi aránynak az éves adatait használtam 2013-2023 közötti időszakban. A három indikátorból kettőnel az a jobb, ha az érték nagyobb, míg a korai iskolaelhagyásnál az a jó, ha az érték kicsi, ezért ott transzformáltam az adatokat, miután a konvergenciákat kiszámoltam (a reciprokukat vettem), így minden változó esetében a nagyobb érték jelenti a jobb teljesítményt. Ezen változókat, ahol elérhető volt férfi/női bontásban is használtam, arra a célra, hogy a nemek közötti különbségeket feltárhassam.

Azért esett erre az időszakra a választás, mivel az adatok nagyrésze 2013-as kezdődátummal érhető el úgy, hogy viszonylag kevés hiányzó érték legyen az adatbázisban. Továbbá ebben az évtizedben számos jelentős fejlemény történt. Implementálni kezdték az Unió oktatási programját, mint például az Európa 2020 program stratégia céljait, hogy a lemorzsolódási arány 10% alá csökkenjen és a felsőfokú végzettséget szerzők aránya 40% felett legyen. Illetve, ezen időszak már magában foglalja a COVID-19 világjárvány okozta zavaró hatásokat, így egy komplex képet kaphatunk annak hatásáról is.

Regionális lefedettséget tekintve a fő fókusz az Európai Unió tagállamain van, de az adatok rendelkezésre állása okán tagjelölt országok, illetve EFTA országok régió is megjelennek az adatbázisban. Továbbá az Egyesült Királyság nem szerepel a vizsgált területek között, mivel 2019 óta nem jelent adatokat az Eurostat felé. Minden adatot a legfrissebb, 2021-es NUTS osztályozáshoz igazítottam. Ehhez biztosítani kellett, hogy a kódok és a határok egyezzenek: például az Eurostat néha frissítette a régiók kódjait (Görögország esetében a régebbi „GR” kód helyett az „EL” kódot használta stb.). Arról is gondoskodtam, hogy az elemzés térbeli súlymátrixa ezeknek a régióknak feleljen meg.

Sajnos, az Eurostat adatbázisai rendelkeznek hiányzó adatokkal. Ezen hiányok nagy mértékben befolyásolhatják a modelleket, akár meg is akadályozhatják azok építését, értelmezését. Ezért a „Multivariate Imputation by Chained Equations” (továbbiakban MICE), azaz „Többváltozós hiányzó-adat pótlás láncolt regressziós egyenletekkel” algoritmussal pótoltam a hiányzó adatokat. A MICE módszer előnye, hogy a hiányzó értékeket nem egyszerű átlaggal vagy mediánnal helyettesíti, hanem a változók közötti összefüggéseket figyelembe véve pótolja, így reálisabb becsléseket ad. Ez különösen fontos térbeli paneladatoknál, ahol a megfigyelések időbeli és regionális mintázata is lényeges, továbbá a hiányzó adatsorok elhagyása nem lehetséges, mivel a modellek komplex mivolta ezt nem engedi meg. Az algoritmus a következőképpen működik:

1. Előkészítés

Először is kivettem az adatbázisból, az összes térbeli azonosító adatot, az országon kívül, mivel ezen változók túl sokrétűek, amelyek nem stabil becsléshez vezetnek. Majd a quickpred() függvénnyel megalkottam a prediktor mátrixot, ami a következőképpen írható le:

(26)

1. Pótló modell építése

Majd ezen mátrix alapján prediktív átlagillesztés (PMM) módszerével pótoltam a hiányzó adatokat a következőképpen:

(27)

az i-edik megfigyelés regresszióval becsült értéke . Minden hiányzó értékre meghatároztam a lehetséges donorok halmzát Di és ezek közül véletlenszerűen választok egyet, amit beillesztek az adatbázisba, ezzel megőrizve az adatok empirikus eloszlásának karakterisztikáját.

1. Többszörös pótlás

Ezután a fent leírt módszerrel 5 különböző adatbázist készítettem, melyek 10 Gibbs- típusú iteráció után épültek fel:

(28)

1. Kész adatbázis választása

Végezetül a több elkészített adatbázis közül kiválasztottam egyet, amelybe visszaállítottam minden geometriai és egyéb azonosítót. Az elemzésem alapjául ezen adatbázis szolgált.

# V. Eredmények

## V. I. Szomszédsági viszonyok definiálása

A szomszédságok meghatározására többféle módon bonyolítható le, kettőt vettem számításba a meghatározáskor. Az egyik, hogy azokat a területi egységeket tekintjük szomszédosnak, amelyeknek a központjai egy előre megadott távolságon belül helyezkednek el egymáshoz képest. Ez a NUTS 2-es területi egységek szintjén nem egy optimális választás, mivel ezen területek nagyon változó méretűek lehetnek, így előfordul, hogyha túl kicsi távolság lesz megadva, akkor egyes területeknek nem lesz szomszédja, míg, ha túl nagy, akkor, ugyan mindenkinek lesz szomszédja, de lesznek olyan régiók, amelyeknek sok tíz szomszédja lesz (például Luxemburg) és lesznek olyanok, amelynek még így is alig lesz 5-6 szomszédja (például Dél-Spanyolországi egységek).

Így inkább a második módszerrel határoztam meg a szomszédokat, amely egy egyszerű szabály szerint működik, aminek közös oldala, vagy sarka van, az szomszédos is. Ezt másnéven királynő szomszédságnak is neveznek, a királynő sakkbábu lépeseinek nyomán. Így a következő szomszédsági hálózatot kapjuk:

2. ábra Szomszédsági hálózat

A map of europe with lines and dots

AI-generated content may be incorrect.

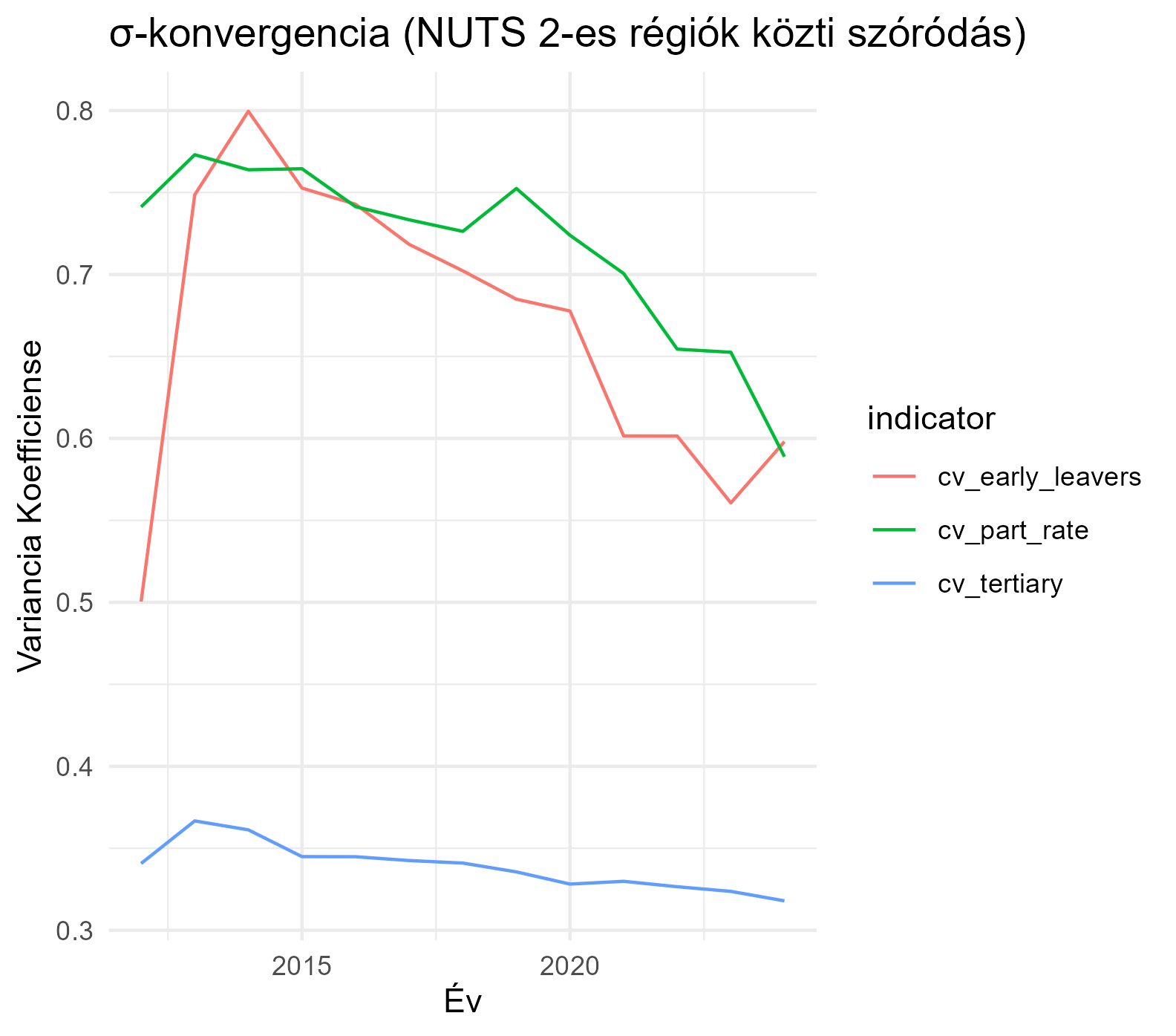
Ezen szomszédság meghatározási mód mellett a szigetek mellett, mint Málta, Ciprus stb. még nem lett szomszédja 1 finn és 1 norvég területnek. Ezeket eltávolítottam az adatbázisból és az összes modell ezen területi egységek nélkül épült. Továbbra is maradtak „elszigetelt” szomszédsági viszonyokkal rendelkező területek (pl. Írország), de ezeknek a hatása fontos lehet az Európai Unió fejlődését tekintve, így nem vettem ki őket az adatbázisból.

## V. II. Modellek eredményei

***Konvergencia jelenléte az oktatás fejlődésében***

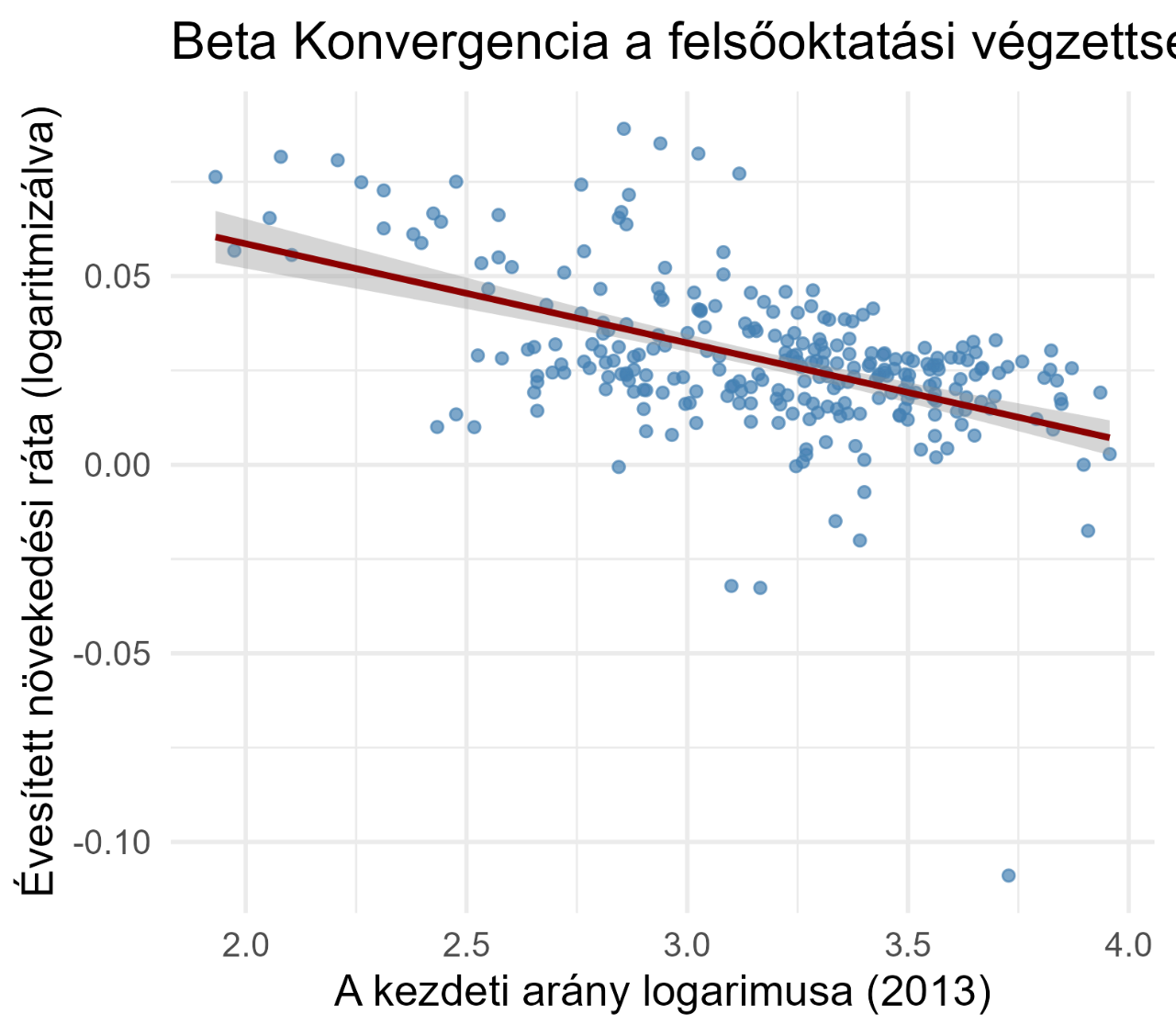
Az általam vizsgált évtizedben jelentős változások voltak megfigyelhetőek az Európai Unióban az oktatási indikátorok terén. Ezen emelkedés betudható az EU 2020-as oktatás fejlesztési célkitűzéseinek, legfőképpen a felsőfokú végzettséggel rendelkezők arányánál figyelhető ez meg. 2013-ban a NUTS 2-es régiók lakóinak mindössze 20%-a rendelkezett felsőfokú végzettséggel, de ez a mérőszám 2023-ra a régiók túlnyomó részénél 35-40%-os sávban van. A korai iskolaelhagyások aránya jelentősen csökkent, mivel a legtöbb ország folyamatos intézkedésekkel próbálta szűkíteni a lemorzsolódók arányát. Az Unió által meghatározott 10%-os küszöbérték környékén, vagy ennél jobb értéken mozog jelenleg a régiók többsége. A felnőttkori oktatásban való részévétel vegyes tendenciákat mutatott, egyes régiók erős, egész életen át tartó tanulási kultúrát alakítottak ki, míg mások stagnáltak, vagy visszaesést tapasztaltak (különösen 2020 körül, a COVID-19 járvány idejében).

3. ábra Szigma konvergencia alakulása



A szigma konvergencia számszerűsítésénél, az 1. ábra egyes mutatók keresztmetszeti szórását ábrázolja, bemutatva a regionális egyenlőtlenségek alakulását. A korai iskolaelhagyás esetében a szóródás meredeken csökkent, illetve ugyanígy alakult a felnőttkori oktatásban való részvétel is. A felsőfokú végzettségek szerzők aránya viszont nem csökkent, csak szinten tartotta magát, sőt egy kis divergencia is megfigyelhető volt. Tehát, míg szinte minden régió javította a felsőfokú végzettséggel rendelkező felnőttek arányát, addig a már eleve vezető pozícióban levő régiók (leginkább fővárosok, innovációs központok) szintén jelentősen javultak, így továbbra is jelentős különbségek maradtak.

4. ábra Béta konvergencia a felsőfokú végzettséget szerzők között



A korábbi szigma konvergencia eredményeket a béta konvergenciák is megerősítik (2.ábra). Például egy keresztmetszeti lineáris regressziójában szignifikáns, negatív β értéket (-0,025) látunk, azaz az alacsonyabb kiindulási arányú régiókban nagyobb arányú növekedés volt megfigyelhető az elmúlt évtizedben.

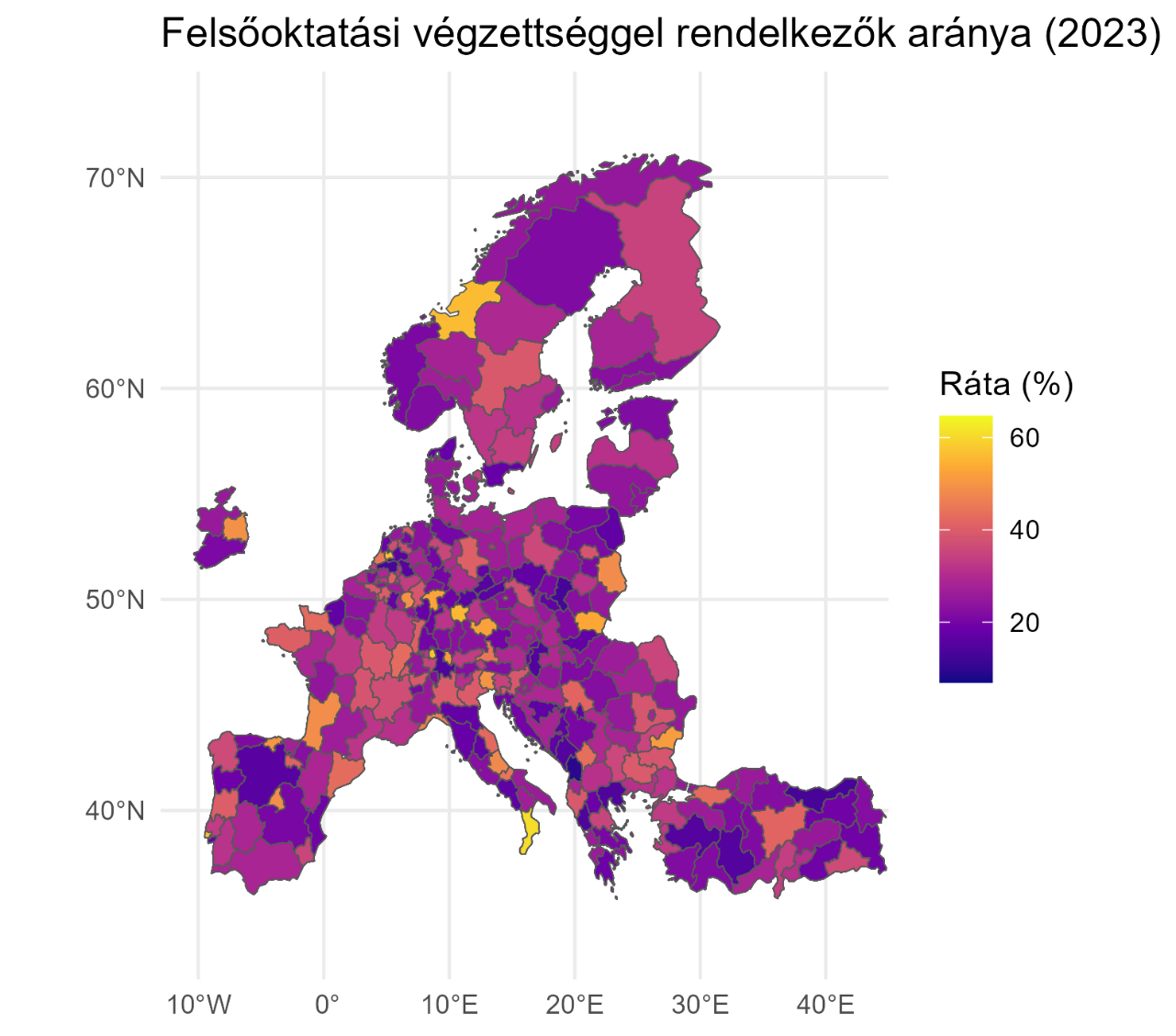
A 2. ábra szemlélteti ezt az összefüggést: a felsőfokú oktatásban részt vevők kezdeti arányának és a százalékos növekedésnek a szórásdiagramja lefelé irányuló tendenciát mutat, megerősítve, hogy a kezdetben lemaradó régiók javultak a leggyorsabban. Hasonlóan jelentős béta-konvergenciát figyelünk meg a többi mutató esetében is (korai iskolaelhagyás és képzésben való részvétel), amit minden esetben a kezdeti értékek negatív együtthatói jeleznek, vagyis a legrosszabb kiindulási helyzetű régiók javultak a leginkább. Fontos azonban ezt összehangolni a szigma eredményekkel: a β-konvergencia jelenléte szükséges, de nem elégséges a σ-konvergenciához. A felsőfokú oktatás esetében, bár a gyengébb teljesítményű régiók relatív értelemben gyorsabban növekedtek, az abszolút különbség továbbra is nagy maradt, mert a jól teljesítő régiók is tovább fejlődtek (pl. a felsőfokú végzettség aránya ~50%-ról ~65%-ra emelkedett). Ez összhangban van azzal a koncepcióval, hogy a béta-konvergencia akkor is bekövetkezhet, ha a szigma-konvergencia megtorpan vagy akár visszafordul, ha a vezető régiók abszolút értelemben továbbra is előre haladnak.

Összefoglalva, néhány oktatási eredményben biztató konvergenciát látunk, de nem mindegyikben. A korai iskolaelhagyás mind szigma-, mind béta-konvergenciát mutat, a lemorzsolódási arányok regionális egyenlőtlenségei jelentősen csökkentek, és a kezdetben legrosszabb helyzetben lévő régiók javultak a leginkább. A felsőoktatás egyértelmű béta-konvergenciát mutat (a növekedési ráták felzárkózása), de csak marginális szigma-konvergenciát, ami azt jelenti, hogy a felsőoktatási végzettség terén az általános egyenlőtlenség továbbra is magas. A felnőttkori tanulásban való részvétel alig mutat konvergenciát, ami arra utal, hogy az egész életen át tartó tanulás terén mélyen gyökerező kulturális vagy rendszerbeli különbségek (például az északi és a dél-európai gyakorlatok között) továbbra is megmaradtak. Ezek a minták arra utalnak, hogy a célzott erőfeszítések sikeresek voltak bizonyos alapfokú oktatási eredmények kiegyenlítésében (például a korai iskolaelhagyás csökkentésében), míg a felsőoktatás bővítése, bár széles körben elterjedt, még mindig sok régiót messze elmarad a vezetőktől.

***Területi mintázatok***

Az oktatási eredmények földrajzi eloszlása Európában egy markáns központ-perem mintázatot mutat.

5. ábra Felsőoktatási végzettséggel rendelkezők aránya (2023)



A 3. ábra a felsőoktatási végzettséget mutatja be régiónként, 2023-ban szemléltető példaként. Azok a régiók, amelyek a legmagasabb felsőoktatási végzettséggel rendelkező felnőtt aránnyal rendelkeznek (gyakran 40% feletti értékekkel) koncentrálódva vannak fővárosokra, illetve tudásalapú gazdasági központokra. Ilyenek például, Stockholm, Madrid, valamint Hollandia és Finnország számos régiója, melyek kiemelkednek ezen mérőszám alapján. Ezzel szemben alacsonyabb arány figyelhető meg Dél-Olaszország több részén, Spanyolország egyes részein, Bulgáriában, Romániában és Magyarországon is. Egyértelmű észak–dél és nyugat–kelet irányú gradiens figyelhető meg: Észak- és Nyugat-Európa ad otthont a legmagasabb képzettségű régiók többségének, míg Dél- és Kelet-Európa számos része lemarad. Vannak kivételek – például Spanyolországban éles belső kontrasztok figyelhetők meg (Madrid és a Baszkföld messze meghaladja a többi spanyol régiót az oktatási szint tekintetében), és más, egyébként magas oktatási színvonalú országok néhány viszonylag vidéki része csak mérsékelt felsőfokú végzettségi arányokkal rendelkezik.

Statisztikai tesztek kimutatnak területi autokorrelációt:

2. táblázat Területi autokorreláció mérőszámok

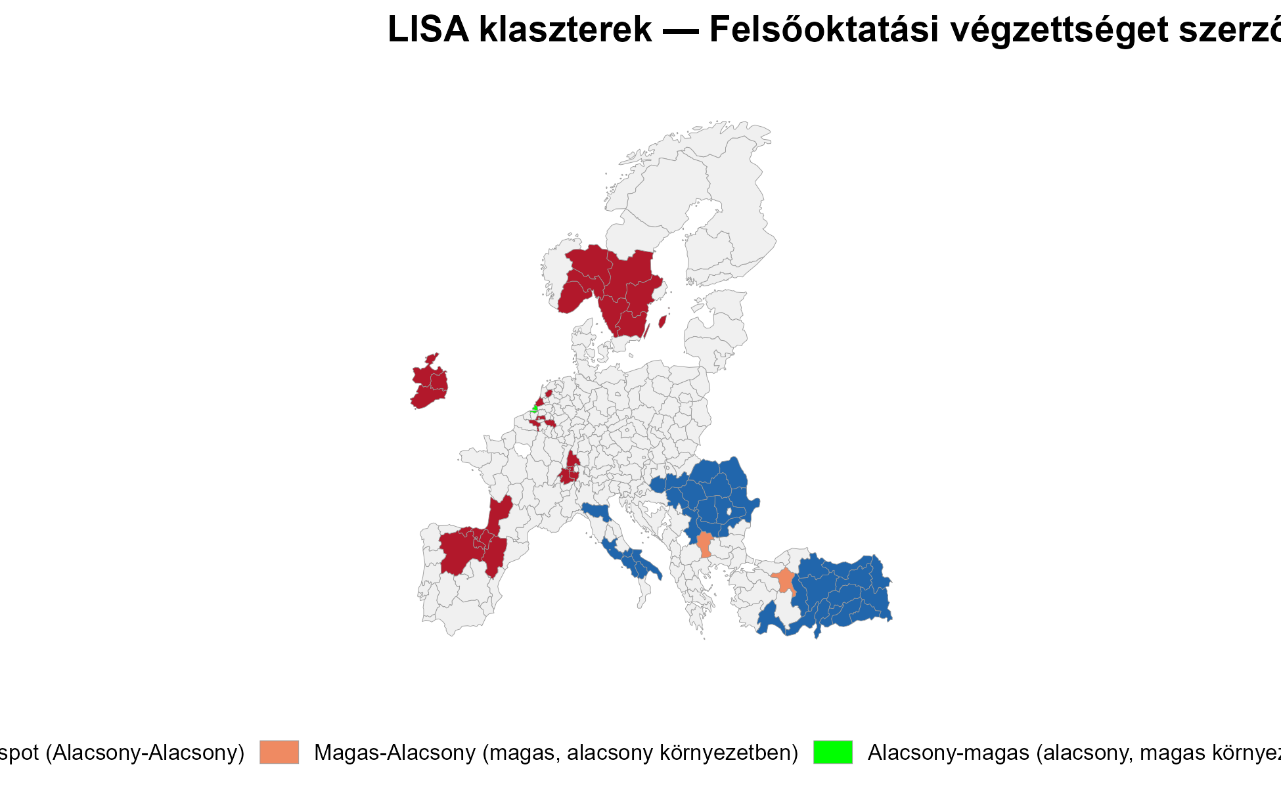
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Év | Korai iskolaelhagyás aránya | | Felnőtt oktatásban résztvevők aránya | | Felsőoktatási végzettséggel rendelkezők aránya | |
|  | Moran’s I | Geary’s C | Moran’s I | Geary’s C | Moran’s I | Geary’s C |
| 2013 | 0,738 | 0,169 | 0,748 | 0,197 | 0,539 | 0,347 |
| 2018 | 0,653 | 0,234 | 0,750 | 0,208 | 0,495 | 0,388 |
| 2023 | 0,485 | 0,438 | 0,744 | 0,205 | 0,488 | 0,395 |
| Átlag | 0,625 | 0,280 | 0,748 | 0,203 | 0,508 | 0,376 |

A felsőoktatási végzettséggel kapcsolatban a Moran’s I mutató a vizsgált időszakban 0,45-0,55 intervallumban mozgott, ami azt jelenti, hogy időszaktól függetlenül, hasonló eredmények mindig csoportosultak. A korai iskolaelhagyás aránya szintén jelentős térbeli autokorrelációt mutat, mind a Moran’s I, szintúgy a Geary’s C mutató tekintetében. Ugyanakkor érdemes megjegyezni, hogy 2013-ról 2023-ra 0,74-ről 0,48-ra csökkent a Moran’s I mutató, ami arra utal, hogy a magas és alacsony lemorzsolódási arányú régiók térbeli csoportosulása kevésbé volt markáns, mivel az általános egyenlőtlenségek csökkentek. Ezzel szemben a felnőttkori képzésben való részvétel térbeli eloszlása rendkívül magas és meglehetősen stabil maradt (Moran-index nagyjából 0,75 mind 2013-ban, mind 2023-ban), ami azt jelzi, hogy a 2013-ban az egész életen át tartó tanulás terén élen járó régiók 2023-ban is nagyrészt élen jártak, a lemaradók pedig továbbra is lemaradók maradtak, tartós térbeli szakadék alakult ki.

Ezek a statisztikák megerősítik, hogy az oktatási eredmények eloszlása nem véletlenszerű térbeli eloszlás, hanem inkább a mögöttes földrajzi, gazdasági és történelmi mintákat tükrözi. Az egyetemek és képzett munkaerőt igénylő iparágak bizonyos területeken való koncentrációja, a magasan képzett munkavállalók fővárosokban történő migrációja, valamint a nemzeti szinten irányított oktatáspolitika hozzájárul ahhoz, hogy szomszédos régiók hasonló oktatási szinten tudjanak mozogni.

Ugyan a Moran’s I és Geary’s C mutatók megmutatták azt, hogy azonos oktatási profillal rendelkező régiók csoportosulnak, nem mutatják meg pontosan ezeket a csoportosulásokat. Erre az úgynevezett LISA (Local Indicators of Spatial Association) mutatók alkalmasak, ezek közül a Local Moran’s I-t használtam, amellyel megtaláltam a hot, illetve cold spotokat.

6. ábra Local Moran's I osztályozás



A 4.ábrán a felsőoktatási végzettséget osztottam fel négy csoportba a szomszédos régiók értéke alapján (10%-os szignifikancia szint mellett):

* Magas- magas (átlag feletti felsőfokú végzettségű régió, amelyet hasonlóan magas értékekkel rendelkező szomszédos régiók vesznek körül): ilyen például Írország, Madrid környéke, illetve Svédország és Norvégia déli része
* Alacsony- alacsony (alacsony felsőfokú végzettséggel rendelkező régió, amelyet alacsony végzettséggel rendelkező szomszédok vesznek körül): ilyen régió csoportosulás például Románia, Bulgária egy része és Magyarország déli része
* Magas- alacsony (magas végzettséggel rendelkező régió, amelynek szomszédai lemaradnak): Ez az elemzésem alapján csak Törökország egy középső régiójára igaz
* Alacsony- magas (alacsony végzettséggel rendelkező régió, amelyet magasabb végzettséggel rendelkező szomszédok vesznek körül): Ez Rotterdam és Antwerpen környékére igaz.

A LISA besorolási térképből is egyértelműen kirajzolódik az, hogy az Északi és Nyugati régiók fejlettebbek a Déli, illetve Keleti régióknál, mivel az előbbi területeken csak átlagnál jobb csoportosulások voltak megtalálhatóak, míg a későbbieknél csupa alacsony értékekkel rendelkező régiók tömötülnek.

Ezek a lokális mintázatok előtérbe helyezik, hogy a regionális oktatási egyenlőtlenségek gyakran olyan földrajzi hatással bírnak, amelyek átnyúlnak az adminisztratív határokon. Az egymással határos, alacsony oktatási színvonalú régiók csoportja arra utal, hogy egyetlen régióban végrehajtott elszigetelt beavatkozások nem feltétlenül elegendőek, egy szélesebb körű, összehangolt stratégia (esetleg régiókon átívelő vagy nemzeti/EU programok által vezetett) lehet szükséges az egész csoport felemeléséhez. Hasonlóképpen, a magas oktatási színvonalú csoportok elhelyezkedésének megértése információval szolgálhat arról, hogy a vezető régiók hogyan oszthatják meg egymással a bevált gyakorlatokat, vagy hogyan hozhatnak létre hálózatokat szomszédjaik támogatására.

A területi autokorreláció által megjelenő mintázatok alapján készült csoportosításon kívül nem felügyelt klaszterezési eljárással is megnéztem, milyen csoportokba lennének egyes régiók sorolva, ha csak az oktatási indikátorok lennének figyelembe véve, a területi és szomszédsági viszonyok nem.

7. ábra K-means klaszterezés az oktatási profil alapján (2023)

A map of europe with red and blue dots

AI-generated content may be incorrect.

A 7. ábrán az első klaszter az átlagnál jobb oktatási profillal rendelkezők, azaz magas a felsőfokú végzettséggel rendelkezők aránya, sokan vesznek részt felnőttkori oktatásban és elenyésző a korai iskolaelhagyók aránya. A második klaszter ennek az ellentétje, míg a harmadik klaszter azon

***Nemi különbségek a felsőoktatásban***

***PSM***

***Az oktatás és a szociális helyzet, illetve gazdasági növekedés kapcsolata***

***RF***

***További megfigyelések***

***…***

# VI. Kitekintés, Magyarország helyzete

3: <https://www.ksh.hu/stadat_files/okt/hu/okt0001.html>

VAR idősoros adatokra, kevés terület adat híján

# VII. Korlátok

Adatok elérhetősége és teljessége

Szomszédság definiálása

EU kohéziós költések nuts2 dokumentálása

Modell assumptions

# VIII. Konklúzió

…

# Irodalomjegyzék

**There are no sources in the current document.**

# Mellékletek