EÖTVÖS LORÁND TUDOMÁNYEGYETEM TERMÉSZETTUDOMÁNYI KAR
FÖLDRAJZ- ÉS FÖLDTUDOMÁNYI INTÉZET
Földrajztudományi Központ
Környezet- és Tájföldrajzi Tanszék

Népességbecslés szabadon hozzáférhető adatok alapján

Population estimation using open data

MAGYAR MÁRTON MÁRK

Geográfus hallgató

geoinformatika szakirány

Diplomamunka

Témavezető:

DR. KOHÁN BALÁZS egyetemi docens

ELTE Környezet- és Tájföldrajzi Tanszék

BUDAPEST, 2019

Tartalomjegyzék

1. Bevezetés

Diplomamunkám témája egy adott térbeli egységekre, ingyenesen hozzáférhető adatok használatával pontos népesség becslésére alkalmas modell fejlesztése és működtetése. Erre az igény a GeoX Térinformatikai Kft-nál[1] végzett munkám során merült fel üzleti térinformatikai felhasználásra. A döntéselőkészítő- és támogató eljárások fejlesztése során fontos változóként szerepelhet adott ország, város illetve terület pontos népessége, amely megbízható, ingyenes formában legutóbb a 2011-es népszámlálás során lett felmérve és publikálva. Friss település- vagy akár címszintű népességadatok statisztikai hivataloktól vagy akár hasonló profillal rendelkező vállalkozásoktól hozzáférhetők, de ezek értékes információként igen költséges megoldások.
Jelen munkában egy olyan költséghatékony alternatívát terveztem létrehozni, amely egy keretrendszerként, tetszőleges számú, méretű országos adatbázist felhasználva képes tetszőleges kimeneti egységek népességét megbecsülni megbízható pontossággal.

Dolgozatomban a folyamat minden lépését a tervezéstől, az adatbázisépítésen keresztül a statisztikai becslés megjelenítéséig részletesen bemutatom.

2. Folyamat megtervezése

A folyamat előtervezése során a követelményeket, rendelkezésre álló adatokat és a célt elemezve a modell két különböző országra való felépítése tűnt célszerűnek. Magyarország mellett Csehországra többet között az egységes EU által nyújtott ingyenes adatforrások, a két ország közel megegyező számú népessége és hasonló területe, valamint a modell ellenőrzésére rendelkezésre álló címszintű lakossági adatbázis léte miatt esett a választás.

A magas elemszámú országos adatbázisok miatt fontosnak tartottam a teljes munkafolyamat előzetes, pontos megtervezését a későbbi műveletek gördülékeny elvégzését elősegítve.

[1] http://www.geox.hu/

- adatok beszerzése (EUROSTAT, OSM, Corine Land Cover, LAU 2, népességadat);
- adatok geoadatbázisban való egyesítése a megfelelő konverziók elvégzésével;
- új, statisztikailag releváns változók létrehozása az adatbázisokban;
- egységes adatszerkezetű, de vizsgálati egységenként különböző táblák felépítése;
- népességbecslő statisztikai modell fejlesztése a táblaszerkezet alapján;
- népességbecslés futtatása Python nyelven írt programban;
- becsült érték visszavezetése térinformatikai szoftverbe.

3. Adatforrások

Az alábbiakban bemutatom a felhasznált szabad adatbázisok jellegzetességeit, tartalmát és elérhetőségüket. Általánosságban elmondható, hogy minden említésre kerülő adatforrás vetülettel rendelkező térbeli adatformátumban hozzáférhető.

3.1. OpenStreetMap

Az OpenStreetMap (OSM) egy szabadon szerkeszthető és felhasználható térképadatbázis és térképszolgáltatás. A kezdeményezést Steve Coast alapító eredetileg az Egyesült Királyság felmérésére indította 2004-ben, miután az állami felmérések hatalmas adatmennyiségéből csak minimális mennyiség volt szabadon hozzáférhető. A szerzői jogok oltalma alól felmentett adatbázis fokozatosan bővült, ahogy 2006-ban a Yahoo, majd 2010-ben a Bing is engedélyezte műholdképeinek szabad felhasználását térképezési céllal. Az aktív önkéntes szerkesztőbázis létszáma 2018. novemberében lépte át az 5 millió főt.

A közösségi szerkesztés azonban a gyorsaság és megbízható helyismeret mellett magában hordozza a véletlen vagy szándékos hibák, pontatlanságok megjelenését az adatbázisban. Ugyan a legtöbb frissítés átfut ellenőrzésen, de mindig maradnak hibás adatok az adatbázisban.

A térkép alapját képező adatbázis többek között a geofabrik honlapról[2] tölthető le ingyenesen, tetszőleges területre, tetszőleges adatmennyiséggel.

3.2. Corine Land Cover

A CORINE felszínborítottsági adatbázis (CLC)[3] az 1985-ben indított Coordination of Information on the Environment nevű program terméke, melynek célja a felszín állapotának felmérése és a változások nyomon követése, kezdetben kizárólag műholdképek alapján. Az első teljes adatbázis 1990-ben készült el, tartalmazva az akkori EU tagállamok felszínborítását 44 különböző osztályba sorolva. Az első verzió megjelenése óta hat frissítés követte az adatbázist; 2000-től kezdve hatévente, fokozatosan több országot lefedve. A legfrissebb, 2018-as adatbázisban így már 39 európai ország szerepel.

A felmérés során a területi alapegység 25 hektár, a vonalas alapegység pedig 100 méter, ezen egységeknél kisebb felszíni elem nem kerül megkülönböztetésre, kivéve a CLC-Change rétegek esetében, amelyek két frissítés közötti változások kimutatására kerülnek generálásra 5 hektáros területi alapegységekkel (FERANEC 2016).

A felszínborítottság osztályozása nagy felbontású (≤ 10 méter) műholdképekről történik CAPI (Computer Assisted Photo Interpretation) módszerrel. Egyes tagállamokban az interpretáció további lépései is automatizálásra kerültek helyi térinformatikai rendszerek alkalmazásával. A 2018-as adatbázis volt az első, amelyben a Sentinel-2 műholdképek adták a vizsgálat alapját (FERANEC 2016). Ugyan az eljárás geometriai pontossága 100 méter alatti, de a végső termék pixelmérete mégis 100 méteres, ugyanis a hatalmas vizsgált terület interpretációját felgyorsítva jelentősen támaszkodik a CLC a térképi generalizálásra, így az adatbázis hivatalos tematikus pontossága 85%.

Az adatbázis előzetes regisztráció után szabadon letölthetővé válik a Copernicus program honlapjáról ESRI GDB[4], GeoPackage[5] és GeoTIFF[6] formátumban is.

---

[2] https://www.geofabrik.de/
[3] https://land.copernicus.eu/pan-european/corine-land-cover
[4] http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Types_of_geodatabases
[5] https://www.opengeospatial.org/standards/geopackage
[6] https://www.opengeospatial.org/standards/geotiff

3.3. GEOSTAT 2011

A GEOSTAT 2011 népességháló[7] egy 1 négyzetkilométeres egységekből álló, az Európai Unió tagállamok területére egységes adatsémával rendelkező népességszám adatbázis. Az adatbázis létrejöttének alapvető feltétele volt a népesség cím pontosságú geokódolása minden tagállam területén, amelyet az Európai Parlament és a Tanács 763/2008/EK Rendelete[8] (2008. július 9.) a népességszámlálásokra vonatkozó szabályozásban előírta, a későbbi statisztikai felhasználás céljából. Az így keletkezett pontos térbeli lakossági adatok tetszőleges területi egységre aggregálhatóak. Bizonyos területek esetében azonban az aggregálás ún. "bottom-up" eljárása adathiány miatt nem követhető, ilyenkor a legkisebb elérhető területi egység lakossági adatát osztják szét az egységek között. Ahhoz, hogy ez az eljárás is minél pontosabban kövesse a valóságot, távérzékelt felvételekről vizsgálják például a felszínhasználatot, az éjszakai kivilágítottságot, a talajpusztulás mértékét és az épületsűrűséget a népsűrűség meghatározása céljából (CHAINEY et al. 2008).

Az adat minősége a hivatalos álláspont szerint minden esetben alacsonyabb, mint az adott nemzeti statisztikai hivatalok által előállított forrásadat, egyrészt az aggregálás folyamata miatt, másrészt a tagállamonként eltérő pontosságot célzó előírások miatt.

3.4. Közigazgatási egységek

A Geographical Information System of the COmmission (GISCO)[9] az EUROSTAT térbeli adatszolgáltató szerve, amely szabad hozzáférhetőséget biztosít a tagállamok publikus téradataihoz. Jelen diplomamunka tárgyát képező vizsgálathoz a LAU 2 szintű[10] adatok felhasználását láttam célszerűnek, amelyet a tagállamok helyi közigazgatási egységei építik fel, a vizsgált országok esetében a települések[11] (VÉRTESY 2019).

[7] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/population-distribution-demography/geostat
[8] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008R0763
[9] https://ec.europa.eu/eurostat/statistics-explained/index.php/Geographical_information_system_of_the_Commission_(GISCO)
[10] https://ec.europa.eu/eurostat/web/nuts/local-administrative-units
[11] https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/communes

3.5. Országos statisztikai hivatalok

Adott település településhierarchiában betöltött szerepe mindkét ország esetében jelentős tényezőként hat a népességére. A települések rang szerinti csoportosításához mind a magyar[12], mind a cseh adatok[13] esetében az adott nemzet statisztikai hivatalán keresztül lehet szabadon hozzáférni.

3.6. GeoX Kft.

A modell visszaellenőrzésére szolgáló címszintű, 2019. évi lakossági adatokat a GeoX Térinformatikai Kft. bocsátotta rendelkezésemre.

4. Használt programok

4.1. ArcMap 10.2

A munkafolyamathoz használt adatfeldolgozó és megjelenítő alkalmazásként az ESRI[14] ArcGIS[15] 10.2 for Desktop programcsalád ArcMap és ArcCatalog alkalmazásait használtam. A program jelen diplomamunka tárgyát képező vizsgálathoz legfontosabb funkciói a térbeli adatok kezeléséhez és elemzéséhez kötődő feladatok támogatása valamint ezek átlátható, rétegszintű megjelenítése. A program képes számos térbeli adatformátum kezelésére, egységesítésre, adatbázisaik térbeli és adatbeli relációinak azonosítására és megjelenítésére. A beépített térinformatikai eszközökön az alapkomponenseken kívül számos kiegészítő alkalmazást lehet fellelni a programhoz, sok esetben akár az aktív felhasználóbázis által publikálva. Ilyen eszköz például a munkafolyamat során használt *CostumGridTools*[16] (Ian Broad 2014), amely a beépített, nem kielégítő részletességű rácshálókezelést bővíti úgy, hogy felezési, negyedelési módszert nyújt, emellett tetszőleges rácsháló-gyártási lehetőséget is kínál.

Az adatok szervezéséhez az ArcGIS Desktop ArcCatalog nevű programját használtam, az ESRI natív file geoadatbázis formátumába konvertálva a legtöbb esetben shape vagy szöveges fájlként kézhez kapott állományokat. Az országos

[12] http://edkvf.kvvm.hu/tartalom/nyomtatvany/kar/telepuleskod.pdf
[13]https://www.czso.cz/csu/czso/small-lexicon-of-municipalities-of-the-czech-republic-2017
[14] https://www.esri.com/en-us/home
[15] https://www.arcgis.com/index.html
[16]https://www.arcgis.com/home/item.html?id=4e2a8fe3f297405d81747df1d1fdb45d

adatbázisok magas elemszáma miatt minden más formátum használata sokkal jelentősebb hely- és processzálási időigényt jelentett volna.

Az ArcGIS által nyújtott automatizálási lehetőségek biztosítják a beépített funkciók szkriptelését tetszőleges paraméterekkel, grafikus formában, akár teljes munkamenetek automatizálását lehetővé téve. A beépített eszközök az ArcGIS Python (ArcPy) nyelven írt alkalmazások, melyek szabadon konfigurálhatók és futtathatók akár az alkalmazáson kívül is. Ez lehetővé tette a terjedelmes adatbázisokból hasonló adattartalmú, de különböző térbeli felbontású kimeneti fájlok létrehozásakor a paraméterezés egyszeri pontos megírását és többszöri futtatását csupán a be- és kimeneti fájlok változtatásával (TOMS–O'BEIRNE 2017).

4.2. Python

A modellezés statisztikai és becslési folyamataiban a Python programozási nyelv 3.7.4 verziójának 32-bites változatát használtam. A Python általános célú, de magas szintű programozási nyelvként 1991-ben jelent meg, ahonnan kezdve folyamatos fejlesztések és bővítések kísérték. Legfontosabb jellemzője, hogy interpreteres programozási nyelv, amely lehetővé teszi az írt programok azonnali forráskódi futtatását (OLIPHANT 2007). További jellemzője, hogy bármikor bővíthető új könyvtárakkal, függvényekkel és modulokkal, amely a korábbi tanulmányaim során szerzett programozási tudás mellett fontos tényező volt a Python választásakor. A folyamat során használt három legfontosabb statisztikai illetve matematikai könyvtár: az adatmanipulációra és analízisre írt pandas 0.25.1 verziója[17], a magas szintű matematikai függvénytárral rendelkező NumPy 1.17.3 verziója[18] és a statisztikai valamint machine learning-re írt scikit-learn 0.21.1 verziója[19] (AVILA –HAUCK 2017).

5. Adatelőállítás, adatszerkezet

5.1. Adatimportálás, adattárolás

A statisztikai vizsgálat végrehajtása előtt a szerteágazó minőségű és struktúrájú forrásadatokat egy egységes adatszerkezettel rendelkező adatbázisban

[17] https://pandas.pydata.org/getpandas.html
[18] https://numpy.org/index.html
[19] https://scikit-learn.org/stable/about.html#citing-scikit-learn

egyesítettem az ArcMap programban. Az adatbázis előállításának első lépése a bemenő adatok egy File Geodatabase-be szervezése volt, egységesített méter alapú vetületi rendszerben. A több esetben kontinens léptékű adatok miatt az Egységes Országos Vetület helyett a Web Mercator (SRID: 3857[20]) alapú vetületi rendszert választottam a File Geodatabase és a benne lévő Feature Class-ok[21] tárolására, megkönnyítve a későbbi adatmanipulációs eljárásokat (KESSLER–BATTERSBY 2019).

A statisztikai feldolgozhatóság miatt minden bemenő adat értéket egységesen szám alapúvá alakítottam, ezzel a szöveges típus értékekből logikai, bináris értékű oszlopokat hoztam létre, melyeket logikai értékként és gyakorisági mutatóként hasznosítottam.

Az adatbázist létrehozásától kezdve kettéosztva kezeltem, a két mintaterület hasonló nevű fájljainak egyszerűbb megkülönböztethetőség okán.

5.2. GEOSTAT 2011

A több mint kétmillió elemű, az egész kontinenst lefedő adatbázisból országkód alapján szűrtem le a vizsgált ország területét, ezzel elkerülve az esetleges adatvesztést, amely a közigazgatási határokkal való adatszűkítés esetében fordulhatna elő, a kilométeres egységek miatt (KOUNADI et al. 2016). A munka során a lakatlan területek kiszűrését megkönnyítette a tény, hogy csak a lakott négyzetek szerepelnek az adatban.

Az 500 méter oldalhosszúságú négyzetek létrehozásához a *CostumGridTools* (Ian Broad, 2014) eszköztárat használtam. A lakatlan vagy hiányos bemeneti adatokkal rendelkező mezők későbbi lépésként történő kiszűrését áthidaltam a GEOSTAT 2011 négyzetháló bemeneti paraméterként való alkalmazásával, amely csak a legalább egy lakossal rendelkező egységekből épül fel. A *CostumGridTools* eszköz lehetőséget ad új négyzethálók létrehozása mellett már létező négyzetek negyedelésére is. Az eszköz ebben az esetben futtatás során minden bemenő négyzet oldalainak szemközti felezőpontjait egy-egy vonallal összeköti, negyedelve a négyzetet. Az Advanced Editing eszköztár *Split Polygons* eszköz használatával a létrehozott vonalas réteggel elmetszve a 1 kilométer

[20] https://epsg.io/3857
[21] https://desktop.arcgis.com/en/arcmap/10.3/manage-data/geodatabases/feature-class-basics.htm

oldalhosszúságú négyzethálót, megkaptam a vizsgálathoz szükséges oldalhosszúságú réteget.

A kilométeres alapegységek népességadatát mint a modellben szükséges bemenő adatot, a *Field Calculator* eszközzel negyedeltem a négyzetet felépítő négy egység közt. A nem egész számú lakosságadatok elkerülése végett az új népességet tároló mezőt csak egész számok befogadására szabályoztam, így minden érték a legközelebbi egészre kerekítve jelenik meg.

5.3. OpenStreetMap

Az OpenStreetMap térképadatbázisok szabadon szerkeszthető jellege miatt nem volt lehetséges egy egységes szűrő alkalmazása a mintaországokra, mivel akár egy útszakasz besorolása[22] is nagyban függ a feltöltő besorolásától, de a legtöbb objektumra léteznek országonként eltérő OSM kategorizálási előírások is (LONDÖGÅRD–LINDBLAD 2018). A folyamat során felhasznált rétegek elemszámai jelentős eltérést mutattak a két országban, ahogy az *1. táblázatban* látszik. Ennek okai között az aktívabb szerkesztő közösség mellett a sűrűbb település- és úthálózatot is említeni lehet (BASH ET AL. 2015).

1. táblázat: felhasznált OSM rétegek elemszámai a magyar és a cseh adatbázisokban (2019.10.08. állapot)

Felhasznált réteg	Darabszám	
	HU	CZ
roads	731 872	1 195 252
railways	16 907	57 281
building	1 206 106	4 651 161
pois_point	140 411	196 195
pois_poly	44 568	60 837
traffic_point	52 210	85 336
traffic_poly	14 693	32 118
transport_point	31 802	28 970
transport_poly	201	232
Összesen	*2 238 770*	*6 307 382*

5.3.1. Út réteg

A letöltött út rétegeket először kategóriák szerint rendeztem. Kiszűrve a minimális elemszámú illetve általános kategóriákba sorolt, amelyek ugyan

[22] https://wiki.openstreetmap.org/wiki/Tagging

elemszám szempontból szignifikáns változók, de leíró jellegükkel minimális statisztikai relevanciájuk miatt nem kerültek a modellbe (*2. táblázat*).

Az ArcMap minden rekord geometriai adatait automatikusan hozzárendeli az attribútumhoz betöltés után, az útszakaszok típusának adott négyzetre eső gyakoriságát, mint változót, mindkét ország esetében ugyanazon ArcPy parancssorral hoztam létre. Első lépésként létrehoztam 19 új szám alapú mezőt, a különböző kategóriák neveivel.

Ezután a parancssorban kijelöltem egy adott kategória összes elemét, majd a kitöltöttem frissen létrehozott kategória oszlopot az kijelölt elemek egyedi azonosító értékeivel, üresen hagyva az összes többi oszlopot:

```
arcpy.SelectLayerByAttribute_management("cz_osm_road_m","NEW_SELECTION","fclass
= 'cycleway'")
arcpy.CalculateField_management("cz_osm_road_m","cycleway",'!code!',"PYTHON")
```

ahol a *cz_osm_road_m* a táblázat neve, az *fclass* a kategória oszlop, *cycleway* az adott érték.

2. táblázat: az OSM út adatbázis kategóriáinak darabszáma (dőlt betűvel a kiszűrt kategóriák)

Kategória	Darabszám	
	HU	CZ
bridleway	*83*	*1 020*
cycleway	**8 042**	**10 942**
footway	**63 431**	**182 997**
living_street	**3 364**	**6 627**
motorway	**3 341**	**4 435**
motorway_link	**2 149**	**2 626**
path	**42 648**	**97 822**
pedestrian	**2 778**	**3 540**
primary	**20 892**	**17 279**
primary_link	**1 157**	**1 720**
residential	**164 996**	**232 466**
secondary	**26 938**	**36 840**
secondary_link	**688**	**815**
service	**106 474**	**186 755**
steps	**8 673**	**16 128**
tertiary	**17 709**	**66 442**
tertiary_link	**512**	**466**
track	*185 912*	*67 063*
track_grade1	*3 523*	*25 828*
track_grade2	*9 461*	*51 033*
track_grade3	*12 556*	*79 659*
track_grade4	*16 445*	*42 179*
track_grade5	*8 795*	*30 945*
trunk	**1 254**	**2 914**
trunk_link	**855**	**2 091**

unclassified	**18 823**	**24 536**
unknown	*373*	*84*

Az országos úthálózati réteget ezek után a *Spatial Join* eszközzel összegeztem és fűztem hozzá a négyzetháló egyes elemeihez. A hozzárendelés során az útszakaszok adott négyzetre eső egyesített hosszát és az különböző útkategóriák egyesített gyakoriságát összesítettem minden elemre.

5.3.2. Vasút

A kötöttpályás rétegek esetében a vasútvonalakra szűrtem az adathalmazt (a villamosokat, metrókat nem figyelembe véve), azonban Magyarország vonatkozásában az elővárosi HÉV vonalak hosszuk és jelentőségük miatt az adathalmaz részét képezik (*3. táblázat*).

3. táblázat: az OSM vasút adatbázis kategóriáinak darabszáma (dőlt betűvel a kiszűrt kategóriák)

Kategória	Darabszám	
	HU	HU
funicular	*2*	*16*
light_rail	**608**	**9**
miniature_railway	*63*	*79*
monorail	*0*	*4*
narrow_gauge	*620*	*621*
rack	*0*	*15*
rail	**13520**	**52333**
subway	*391*	*620*
tram	*1703*	*3584*

5.3.3. Point of interest (POI)

Points of interest (POI) avagy látnivaló objektumokat három különböző tematika alapján tárolja az OSM adatbázis, az objektum méretétől függően, pont vagy poligon formátumban (BAKILLAH et al. 2014). A három tematikát (forgalmi, közlekedési, összesített) mind geometriai, mind adattárolás szempontból egyesítettem országonként egy fájlban. A poligon rétegeket a *Feature to Point eszköz* használatával alakítottam pont alapú réteggé, amely a geometriák súlypontját vette az új pont koordinátáinak. Ezek után ismét a *Spatial Join eszköz* használatával egyesítettem és rendeltem hozzá minden grid alapegységhez a benne található látnivalók számát.

5.3.4. Városközpont index

Az egyesített POI adatbázisból szükségesnek láttam egy precízebben kialakított mutató létrehozását is a városközpontok megbízhatóbb kijelölésére[23], az egyszerű térbeli sűrűségen túl. Ehhez a látnivalók listáját centrum területekre jellemző turisztikai, pénzügyi, kereskedelmi, adminisztrációs, tömegközlekedési és egészségügyi elemekre szűrtem, majd egy 1-től 3-ig terjedő skálán frekventáltsági változót rendeltem minden kategóriához, általános érdeklődés és forgalom alapján.

- 3-as érték*: fast_food , hospital , mall , supermarket*
- 2-es érték: *bank , department_store , school , university*
- 1-es érték: *sports_centre, attraction, hotel, doctors, community_centre, college, town_hall, restaurant, pharmacy, pub, zoo, convenience, doityourself, outdoor_shop, kindergarten, cafe, post_office, clothes, shoe_shop, chemist, furniture_shop, biergarten, toilet, sports_shop, kiosk, bakery, newsagent, greengrocer, bar, hairdresser, beverages, atm, parking_multistorey, fuel, bus_station, railway_station*

Majd az így létrejött értékeket ismét a vizsgálati alap négyzeteire aggregáltam.

5.3.5. Házak

Az építmény réteg esetén a népesség szempontjából legfontosabb lakóházakra szűrtem az adatbázist az ArcMap programban, a következő szűrési feltétellel:

```
"type" in (' ' , 'apartments' , 'detached', 'house' , 'mansion' , 'residential'
, 'residential42' , 'semi' , 'semidetached_house')
```

Ezt követően az építmények kerületét és területét aggregáltam az egy illetve fél négyzetkilométeres vizsgálati egységekre.

5.4. CLC

A felszínborítottság osztályozása a CLC esetében 44 osztályban történik, melyek háromszintű hierarchiába rendezhetők. Ezen osztályokat célszerűnek láttam generalizálni, ezért tematikusan 5 általános felszínborítottsági osztályban egyesítettem a 44 hivatalos osztályt (*4. táblázat*).

[23] http://www.geoindex.hu/adatbazisok/kozponti-helyek-budapesten-adatkorok-elemzeshez/

4. táblázat: a Corine felszínborítottsági adatbázis hivatalos osztályai illetve a vizsgálathoz használt osztályok

CLC kód	CLC kategória név	Saját kategória
111	Continuous_urban_fabric	Urbánus
112	Discontinuous_urban_fabric	Urbánus
121	Industrial_or_commercial_units	Ipari
122	Road_and_rail_networks_and_associated_land	Ipari
123	Port_areas	Ipari
124	Airports	Ipari
131	Mineral_extraction_sites	Ipari
132	Dump_sites	Ipari
133	Construction_sites	Ipari
141	Green_urban_areas	Urbánus
142	Sport_and_leisure_facilities	Urbánus
211	Non-irrigated_arable_land	Mezőgazd
212	Permanently_irrigated_land	Mezőgazd
213	Rice_fields	Mezőgazd
221	Vineyards	Mezőgazd
222	Fruit_trees_and_berry_plantations	Mezőgazd
231	Pastures	Természet
241	Annual_crops_associated_with_permanent_crops	Mezőgazd
242	Complex_cultivation_patterns	Mezőgazd
243	Land_principally_occupied_by_agriculture_with _significant_areas_of_natural_vegetation	Mezőgazd
311	Broad-leaved_forest	Természet
312	Coniferous_forest	Természet
313	Mixed_forest	Természet
321	Natural_grasslands	Természet
322	Moors_and_heathland	Természet
324	Transitional_woodland-shrub	Természet

331	Beaches_dunes_sands	Természet
332	Bare_rocks	Természet
333	Sparsely_vegetated_areas	Természet
334	Burnt_areas	Természet
411	Inland_marshes	Víz
412	Peat_bogs	Víz
511	Water_courses	Víz
512	Water_bodies	Víz

A letöltés után az GeoTIFF formátumú raszteres állományt az ArcMap *Raster To Polygon* eszközével vektorizáltam. Az osztályok egyszerűsítése után létrehoztam öt logikai mezőt az adatbázisban, amelyeket 0 illetve 1 értékkel töltöttem fel, attól függően, hogy a négyzet súlypontjában milyen osztályban tartozó poligon helyezkedik el. Ehhez a *Select By Location* eszközt használtam. Így a négyzetháló minden eleméhez rendeltem egy határozott felszínborítottsági osztályt:

```
arcpy.SelectLayerByAttribute_management("clc_hu_cz_m","NEW_SELECTION","Code_12
in ('111', '112', '141', '142')")
arcpy.CalculateField_management("clc_hu_cz_m","clc_urban",'1',"PYTHON")
```

A bináris mezőn kívül a felszínborítottsági poligonok egy négyzetre eső számát is hozzárendeltem a négyzethálóhoz, mint változó.

5.5. Városrang

A statisztikai hivatalok településrang adatait hozzárendeltem a LAU2 egységek poligonos adataihoz, kategorizálva minden települést. A városrangok jelentős eltérést mutattak a két ország között, ezért négy általános típusba soroltam minden települést (*5. táblázat*). Majd ezen értékeket felhasználva hoztam létre a négy logikai városrang mezőt az adatbázisban, ismét a négyzetek súlypontját használva a besorolás alapjául. Az adatbázisba táplálást a *Spatial Join* eszközzel végeztem.

5. táblázat: Magyarország és Csehország hivatalos településrangjai illetve a vizsgálathoz használt kategóriák

HU		Saját kategória	CZ	
Kategória	Darabszám		Kategória	Darabszám
főváros	1	capital	capital city	1
megyeszékhely, megyei jogú város	18	stat_city	statutory city	25
megyei jogú város	5			
város	304	town	market town	223
			town	578
nagyközség	119	municipality	municipality	5427
község	2 707			

5.6. Népességadat

Az országos címszintű népesség adatbázisokat a *Spatial Join* eszközzel aggregáltam az egységnégyzetekre. A GEOSTAT és a címszintű adatbázis létrejötte közötti 8 év különbség okán a négyzetháló által lefedett területeken kívül is találhatók lakossági adatok (GALWAY et al. 2012). A munkafolyamat során ezek nem kerültek aggregálásra.

5.7. Adatszerkezet

Az adatelőkészítés után a két ország adatbázisait (*6. táblázat*) pontosvesszővel tagolt szöveges fájl formátumban[24] (CSV) exportáltam a szöveges típusban egységesítve az adatok tárolási módját.

6. táblázat: a modellbe betáplált adatbázis adatszerkezete

Mező neve	Tartalom
GRD_ID	azonosítómező
GEOSTAT_nep	2011-es népesség
build_count	Épület darabszám
build_area	Épület terület
build_peri	Épület kerület
road_count	Út darabszám
road_lenght	Út hossz
cycleway	Bicikliút darabszám
footway	Járda darabszám

Mező neve	Tartalom
tertiary_link	Mellékút felvezető darabszám
trunk	Gyorsforgalmi út darabszám
unclassified	Kategorizálatlan út darabszám
trunk_link	Gyorsforgalmi felvezető darabszám
rail_count	Vasútvonal darabszám
rail_lenght	Vasútvonal hossz
poi_count	Látnivaló darabszám
poi_score	Városközpont index
geox_nep	2019-es népesség

[24] https://frictionlessdata.io/docs/csv/

living_street	Sétálóutca darabszám
motorway	Autópálya darabszám
motorway_link	Autópálya felvezető darabszám
path	Földút darabszám
pedestrian	Gyalogosok által használt út darabszám
primary	Főút darabszám
primary_link	Főút felvezető darabszám
residential	Lakóövezeti út darabszám
secondary	Másodrendű út darabszám
secondary_link	Másodrendű út darabszám
service	Szervízút darabszám
steps	Lépcső darabszám
tertiary	Mellékút darabszám

clc_urban	Települési felszín
clc_industry	Ipari felszín
clc_agricult	Mezőgazdasági felszín
clc_nature	Természeti felszín
clc_water	Vizi felszín
clc_count	Felszínkategória darabszám
capital	Főváros
municipality	Falu
stat_city	Adminisztrációs jelentőséggel bíró város
town	Város

6. Módszertan

A munkafolyamat következő része az előkészített forrásadatok modellbe illesztése, és statisztikai módszerrel való becsült érték kinyerése. A folyamat célja egy adott országra általánosan érvényes képlet kiszámítása volt. Ezt a statisztikai modell a bemenő adatokban felfedezhető, a lakosság térbeli eloszlását leíró tendenciáinak változókra bontásával végzi. Fontos kritérium volt a modell és az eljárás közérthetősége, és a becsült érték későbbi egyszerű visszavezetése és megjelenítése térinformatikai szoftverben.

6.1. Többváltozós lineáris regresszió

A népességbecslést a bemenő adatok, mint változók közötti regresszió számítással végeztem el. A regresszió számítás lehetővé teszi lineáris kapcsolat felállítását a függő és több független változó között (PÖDÖR 2016). Az elemzéshez a többváltozós regresszió-analízis statisztikai módszerét alkalmaztam, amely következő képlettel írható fel:

$$y = b + a^1x^1 + a^2x^2 + a^3x^3 + \ldots + a^nx^n,$$

ahol az y a függő (becsült népesség értéke), az x^1, x^2, x^3,..., x^n a független változók illetve a b, a^1, a^2, a^3,..., a^n pedig a regressziós együtthatók. A regressziós együtthatók az adott változók hatását mutatják a függő változóra. Például, ha az egyenletben egyedül a x^i értéken változtatunk egy egységnyit, akkor a függő érték a^i értékkel fog megváltozni. Az adatelőkészítés során létrehozott logikai mezőket a statisztikában minőségi (dummy) változóként kezeli a modell, a 0 értéket a minőségi jellemző hiányának betudva, míg az előfordulását az 1 érték jelzi (DOMÁN 2005).

A modell felépítése során arra törekedtem, hogy minél kevesebb 0-tól jelentősen eltérő változó kerüljön beépítésre, ami jelen modellnél nagyobb terület egységek minőségi változóinak bevonásának eredményeként valósult meg.

A regresszió alapú becslés pontosságát a folyamat során a statisztikai hiba értékekkel ellenőriztem. A két használt érték egyike a Mean Absolute Error (MAE) amely a becsült és a tényleges adat közötti átlag eltérést mutatja:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=n}^{n} |x^i - y^i|$$

ahol az y^i a becsült érték, x^i a pontos érték és n a rekordok száma. A másik mutató pedig a Root-Mean-Square Deviation (RMSD), amellyel a túlbecsült érték és a pontos érték közötti eltérést, azaz reziduum négyzetének átlagából vontam gyököt, megkapva a jellemző eltérést a két érték között:

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\widehat{y_t} - y_t)^2}{T}}$$

ahol az $\widehat{y_t}$ a becsült érték, y_t a független változó értéke és a *T* a futtások száma.

6.2. Python modell

A modell felépítését az előzetesen telepítésre került statisztikai és adatkezelési modulok (pandas, numpy, sklearn) behívásával kezdtem.

```
import pandas as pd
import numpy as np
import sklearn
from sklearn.model_selection import train_test_split
```

```
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from pd import dataframe
```

Ezután a korábban .CSV formátumba exportált adatbázisokat olvasásra behívó pandas funkció következik:

```
dataset = pd.read_csv('c:/valami/adat.csv')
```

A függő, illetve független változókat oszlopnév alapján jelöltem ki. A pandas modul lehetővé teszi a beolvasott fájlok oszloponkénti olvasását, ami jelentősen megkönnyítette a változók hozzáadását illetve kivételét, akár két futtatás között is:

```
X = dataset[['GEOSTAT_nep', 'build_count', 'build_area', 'build_peri',
'road_count', 'road_lenght', 'cycleway', 'footway', 'living_street', 'motorway',
'motorway_link', 'path', 'pedestrian', 'primary', 'primary_link', 'residential',
'secondary', 'secondary_link', 'service', 'steps', 'tertiary', 'tertiary_link',
'trunk', 'unclassified', 'trunk_link', 'rail_count', 'rail_lenght', 'poi_count',
'poi_score', 'clc_count', 'clc_urban', 'clc_industry', 'clc_agricult',
'clc_nature', 'clc_water','capital','municipality','stat_city','town']].values
```

Hasonló egyszerűséggel cserélhettem a becsült érték oszlopát is, ugyan a munkafolyamat során egyedül a házszám szintű 2019-es népességszámot használtam visszaellenőrző értékként:

```
y = dataset['geox_nep'].values
```

A későbbi értelmezés érdekében, sorszámok helyett a független változóknak tetszőleges nevet is megadhattam:

```
z = ['GEOSTAT_nep', 'build_count', 'build_area', 'build_peri', 'road_count',
'road_lenght', 'cycleway', 'footway', 'living_street',
'motorway','motorway_link', 'path', 'pedestrian', 'primary', 'primary_link',
'residential', 'secondary', 'secondary_link', 'service', 'steps', 'tertiary',
'tertiary_link', 'trunk', 'unclassified', 'trunk_link', 'rail_count',
'rail_lenght', 'poi_count', 'poi_score', 'clc_count', 'clc_urban',
'clc_industry', 'clc_agricult', 'clc_nature',
'clc_water','capital','municipality','stat_city','town']
```

Ezt követően a tanuló, illetve teszt területek adott adatbázisban egymáshoz viszonyított arányát és mintavétel véletlenszerűségét határoztam meg:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=1)
```

Az sklearn modul statisztikai módszerei közül a lineáris regressziót választva és kijelölve a tanulóterületeit, elvégeztem a modell betanítását az adat 80%-án:

```
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

A tényleges értékbecslés futtatása előtt a hibás, jelentéktelen vagy éppen népesség függő változóinak kiszűrése érdekében megvizsgáltam a betanításból származó regressziós együtthatók táblázatát:

```
coeff_df = pd.DataFrame(regressor.coef_, z, columns=['Együttható'])
print(coeff_df)
```

Ha a táblázat értékei túlnyomórészt 1 körüliek voltak, kevés kiugró értékkel, akkor lefuttattam a becslést a bemenő adat véletlenszerűen választott 20%-ára:

```
y_pred = regressor.predict(X_test)
```

Ellenőrzés céljából táblázatba illeszthető az első és az utolsó 5 egység tényleges és a becsült értéke, de lehetőség van az összes sor megjelenítésére is, a táblázat maximum dimenzióit növelve:

```
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
print(df)
pd.set_option('display.max_rows', 50)
pd.set_option('display.max_columns', 500)
```

Ugyancsak a becslés pontosságát hivatott ellenőrizni a három alábbi statisztikai hibamutató, a sklearn modul részeként:

```
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test,
y_pred)))
```

Végül a modell futása során a konzolban megjelenő regressziós együtthatókat és becsült értékeket egy tetszőleges helyen létrehozható szöveges fájlként exportálhatóvé tettem:

```
f = open("c:/valami/kimenet.txt","w")
f.write(str(coeff_df) + "\n")
f.write(str(df))
```

Ebben a szöveges fájlban egyesítve jelentek meg az együtthatók értékei, soronként a tényleges és becsült népességértékek illetve a becslés statisztikai hibajelzői (*1. ábra*).

6.3. Népességbecslés a képlettel

A modellből kinyert együtthatókat tartalmazó szöveges fájlt felhasználva az ArcMap keretein belül végeztem el a tényleges becslést. Ehhez létrehoztam egy új mezőt a becsült népességszámnak, amelyet a *Field Calculator* eszközzel minden egységre kiszámítottam az adott modell futásból származó együtthatókkal:

```
pred = ([GEOSTAT_nep] * 1.000522) +
([build_count] * -0.028065) + ...
```

Ezt követően minden adatbázishoz rendeltem egy, a valós és a becsült népességszám különbségét mutató differencia mezőt a *Field Calculator* eszközzel:

```
diff = Abs( [geox_nep] - [pred] )
```

Coefficient

build_count	-0.028065
build_area	-0.000226
build_peri	0.001133
road_count	0.852728
road_lenght	-0.002496
cycleway	1.853003
footway	-0.768441
living_street	-2.454333
motorway	0.100325

	Actual	Predicted
0	1	2.481985
1	3	2.104180
2	5	5.894637
3	14	19.639221
4	38	33.731938
...	...	...
8819	35	36.921524
8820	25	18.956774
8821	242	244.968787
8822	35	33.744497
8823	25	21.016387

Mean Absolute Error: 19.117435772441556

Root Mean Squared Error: 92.29272708443098

1. ábra: a modell kimeneti szöveges fájljának tartalma

7. Eredmények

7.1. Népességbecslés

A modell alapfunkcióját kialakítottam, szabadon variálható, szűrhető és pontosítható számított értékek formájában. A bemeneti adatbázis szintén számos formában betáplálható. Országtól, adatmennyiségtől és alapegységtől függően a modell különböző együtthatókkal és pontossággal hajtja végre a népességbecslést. A folyamatot 4 különböző adatbázison (Magyarország és Csehország 1km*1km és 500m*500m felbontású négyzethálóin) teszteltem 3 különböző paraméterezésű futtatással, a legpontosabb térbeli és statisztikai becslés érdekében.

7.1.1. Minden változó, minden érték

A modellt a bemenő adatbázisok minden sorát és mind a 39 független változót figyelembe véve, tehát változtatás nélkül, alapesetben futtattam. Az országos összesített értékek a magyarországi 1 kilométeres négyzetháló esetében mutatnak pontos, 1% alatti eltérést (*7. táblázat*). Az egységnégyzetekre eső becsült népesség

mindhárom futás esetében nagyjából hasonló pontossági értéket mutat, az MAE értéke 100 alatti illetve a RMSD értéke 200 alatti értéket mutat, minimális növekedést mutatva a kifinomultabb modellfutások esetében (*2-5. ábra*).

7. táblázat: a négy országos adatbázisra becsült népességértékek a modell változtatás nélküli futtatása esetén

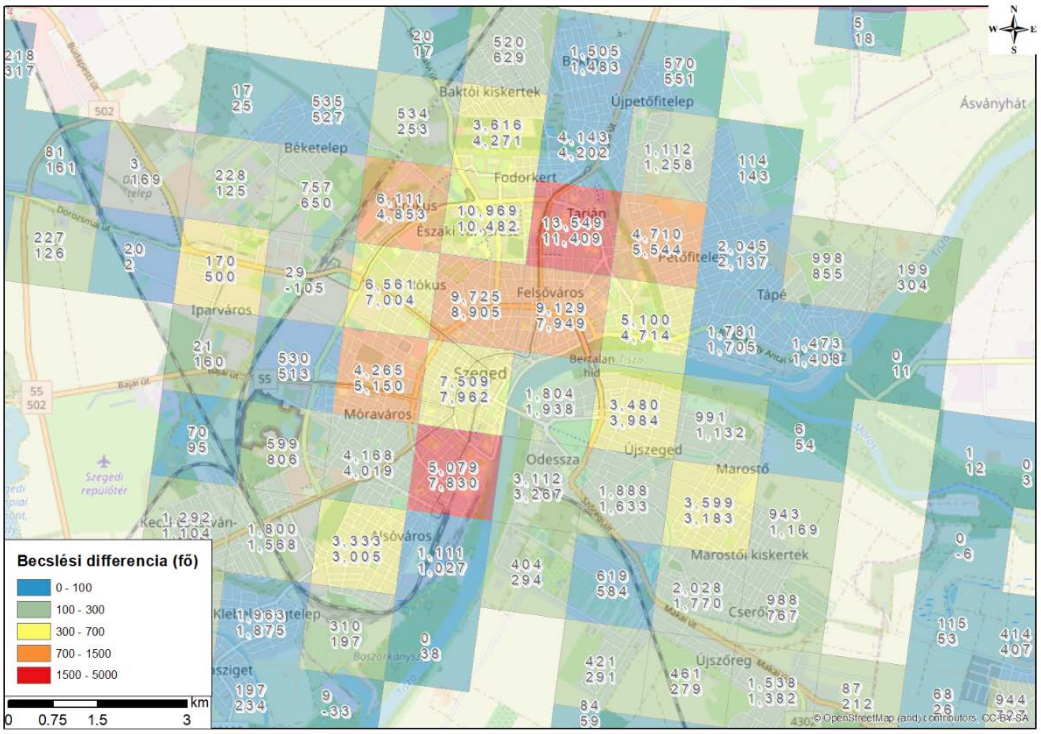
Bemenő adatok		Pontos lakosság	Becsült lakosság	Különbség	MAE	RMSD
HU	1km*1km	9 767 012	9 690 930	76 082.19	59.00381	150.02051
	500m*500m	9 767 032	11 228 266.1	1 461 234	48.85545	133.72092
CZ	1km*1km	10 382 108	10 376 687.1	5 420.896	19.11743	92.29272
	500m*500m	10 382 153	19 127 782.8	8 745 630	39.03633	121.56035

7.1.2. Minden változó, szűrt bemenő adatok

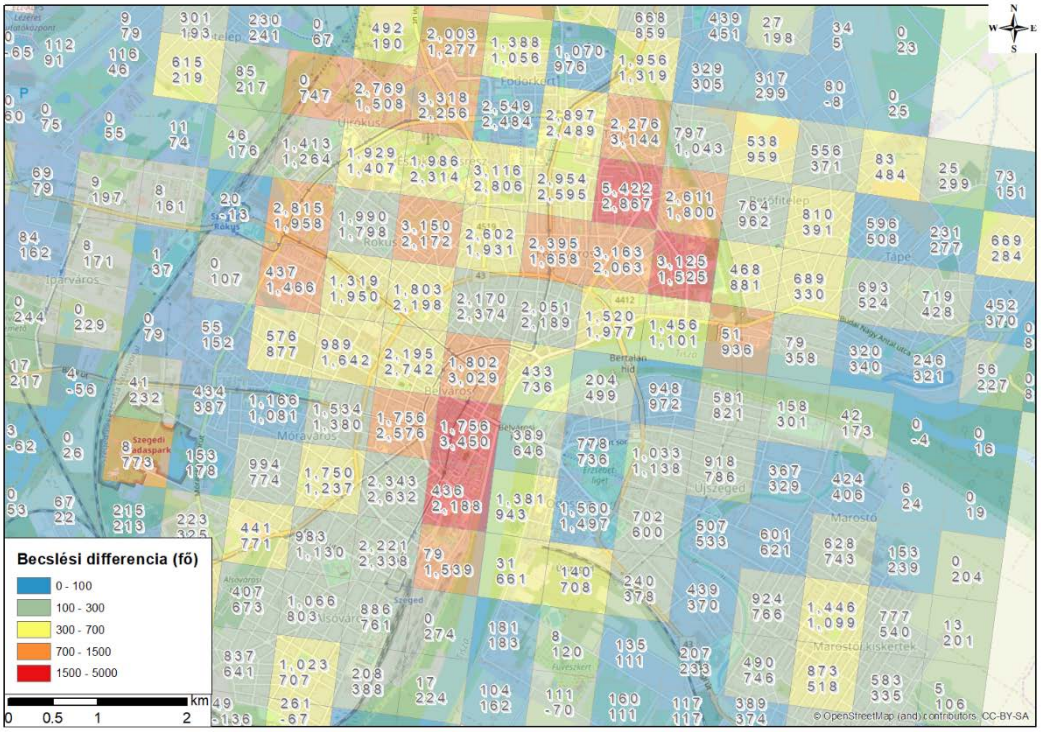
Az első paraméterezési eljárás során a bemenő adatokból kiszűrtem az összes olyan ritkán lakott egységnégyzetet, ahol a 2011-es felmérés szerint 10 főnél kevesebb volt az állandó lakosok száma. Így a modell a betanulást egy szűkebb adathalmazon végezte, kevesebb kiugró értékkel, ami például Magyarországon a tanya területek esetében minimális bemenő adatra (utak, épületek) is tényleges népesség értéket mutatott (*8. táblázat*) (*6-9. ábra*).

8. táblázat: a négy országos adatbázisra becsült népességértékek a szűrt bemeneti adatokkal való futtatás estén

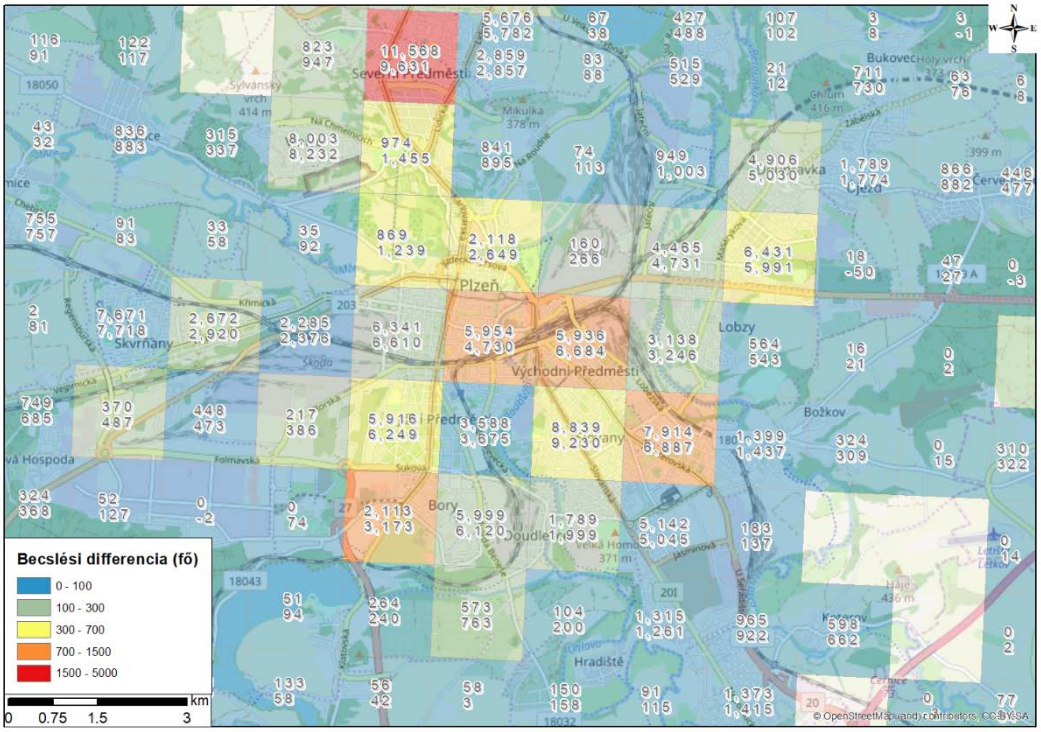
Bemenő adatok		Pontos lakosság	Becsült lakosság	Különbség	MAE	RMSD
HU	1km*1km	9 767 012	9 827 014	60 002.4	69.08855	168.15568
	500m*500m	9 767 032	11 903 454.64	2 136 423	73.03638	175.96323
CZ	1km*1km	10 382 108	9 463 540.8	918 567.2	23.60008	101.05990
	500m*500m	10 382 153	23 503 401.40	13 121 248	49.09059	133.01080



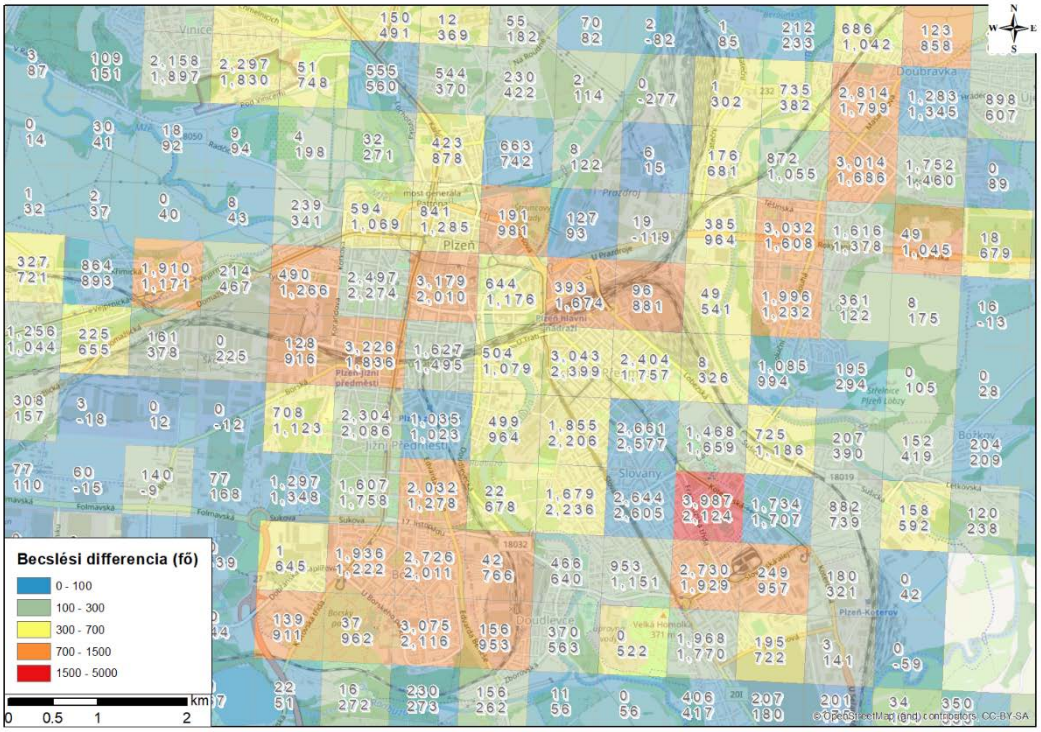
2. ábra: a népességbecslés pontossága Szeged 1:60 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, változtatás nélküli modellfutás esetén.



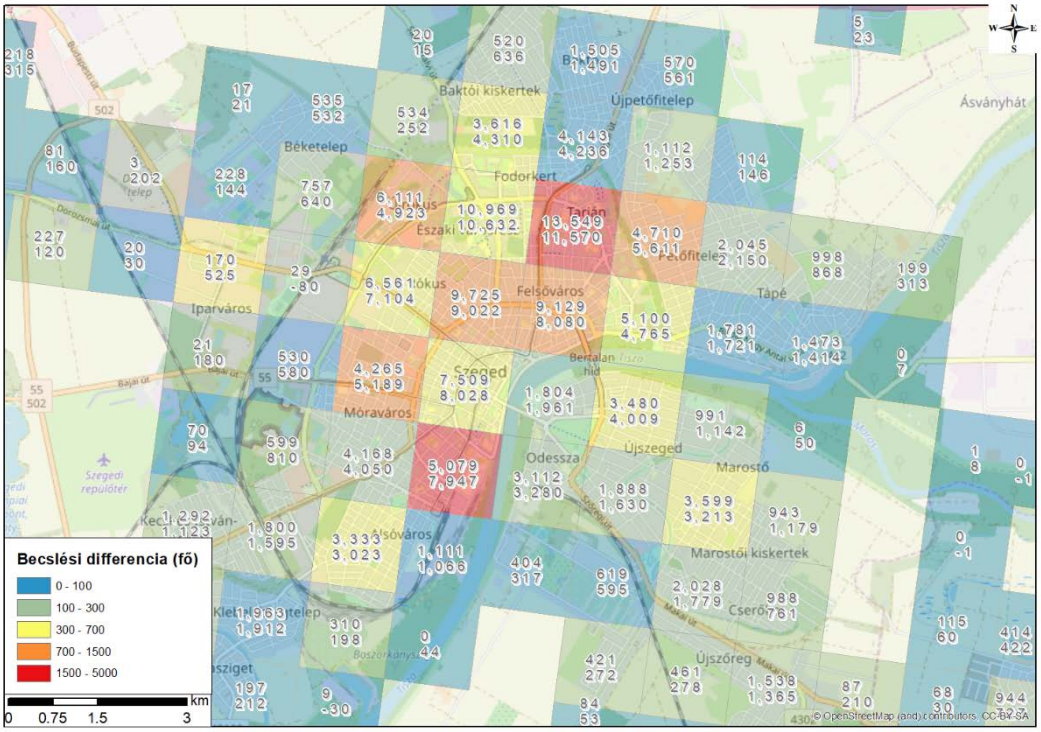
3. ábra: a népességbecslés pontossága Szeged 1:40 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, változtatás nélküli modellfutás esetén.



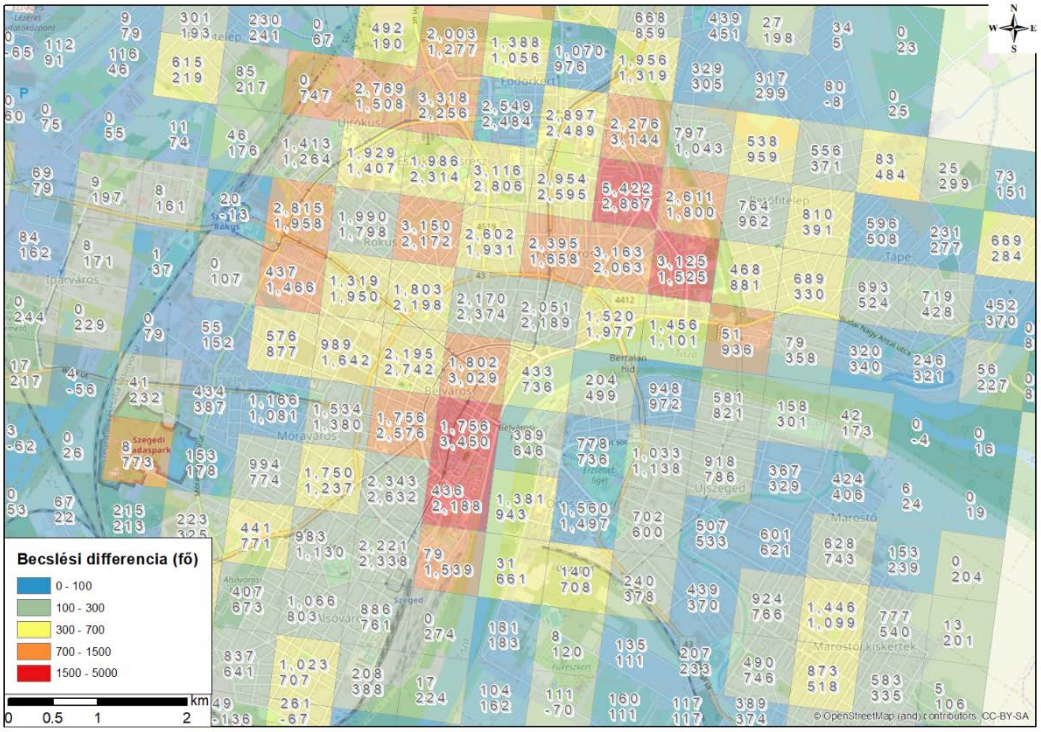
4. ábra: a népességbecslés pontossága Plzeň 1:60 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, változtatás nélküli modellfutás esetén.



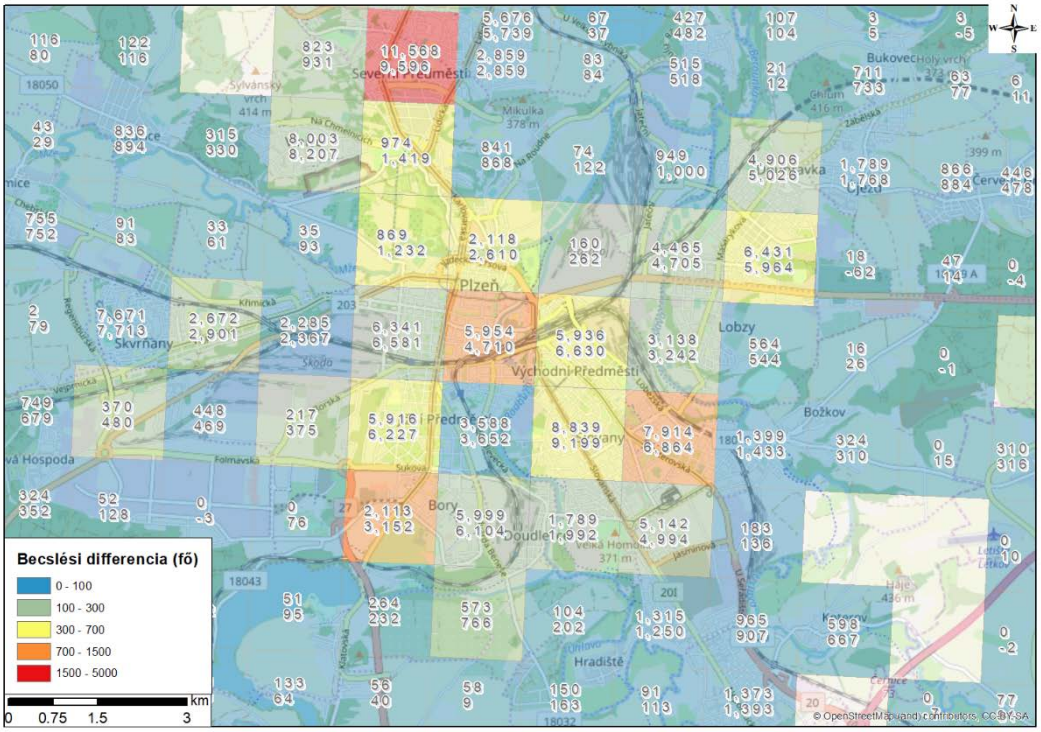
5. ábra: a népességbecslés pontossága Plzeň 1:40 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, változtatás nélküli modellfutás esetén.



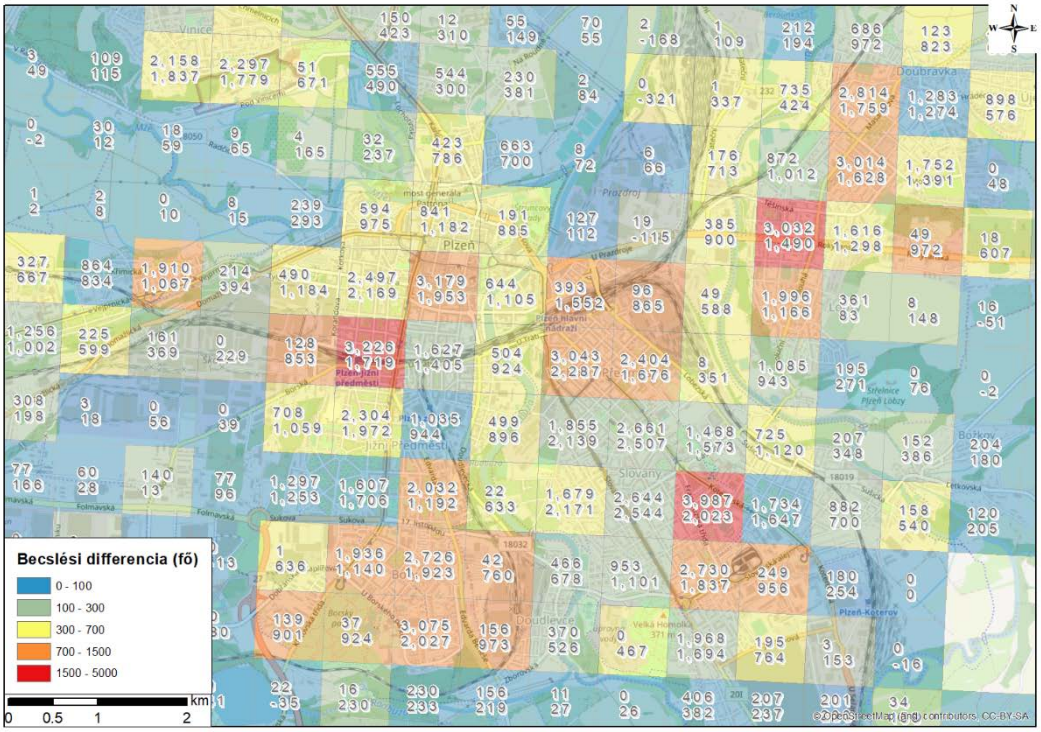
6. ábra: a népességbecslés pontossága Szeged 1:60 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt bemeneti értékekkel futtatott modell esetén.



7. ábra: a népességbecslés pontossága Szeged 1:40 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt bemeneti értékekkel futtatott modell esetén.



8. ábra: a népességbecslés pontossága Plzeň 1:60 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt bemeneti értékekkel futtatott modell esetén.



9. ábra: a népességbecslés pontossága Plzeň 1:40 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt bemeneti értékekkel futtatott modell esetén.

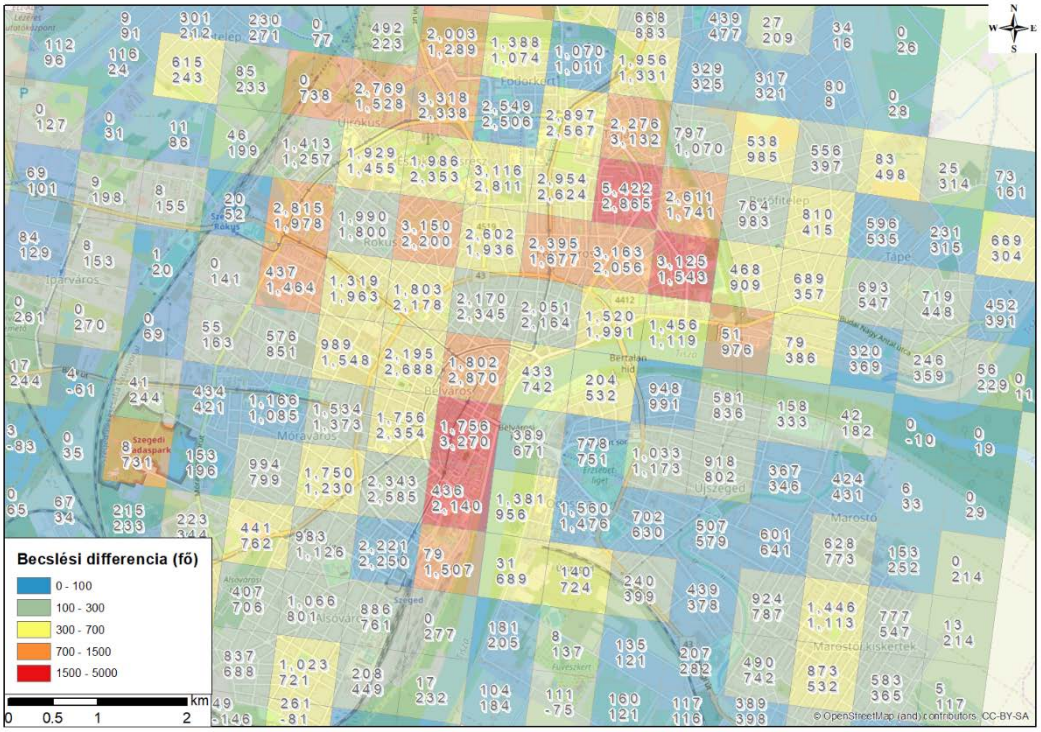
7.1.3. Csak releváns változók, nincs kiugró érték

A harmadik esetben a modell futtatások során meghatározott együtthatók közelítése volt a cél. Alaphelyzetben futtatva, az népességet befolyásoló együtthatók a 10^{-4} és 10^{2} közötti tartományban szóródtak szét, egyes látszólag jelentéktelen tényezők jelenléte századára csökkentve az együttható jelentőségét, míg más, lakossággal közvetlen összefüggést sejtető tényezők (lépcső típusú útszakaszok) elhanyagolható jelentőséggel befolyásolták a becslést. Kivettem a kategorizálatlan útszakaszok tényleges magyarázati érték hiányában. A fővárosi lakosság túlnyomó többsége a többi településsel szemben mind a két tesztelt ország esetében a településrang változók kivételét tette szükségessé az együtthatók normalizálása érdekében. Az együtthatók fokozatos szűrése után 23 független változóra szűkítettem a modellbe bemenő adatként szereplő oszlopokat, jelentősen csökkentve az elemek szórását és a becsült és tényleges értékek közötti differenciát illetve a statisztikai hiba értékeket is redukálva (*9. táblázat*) (*10-14. ábra*).

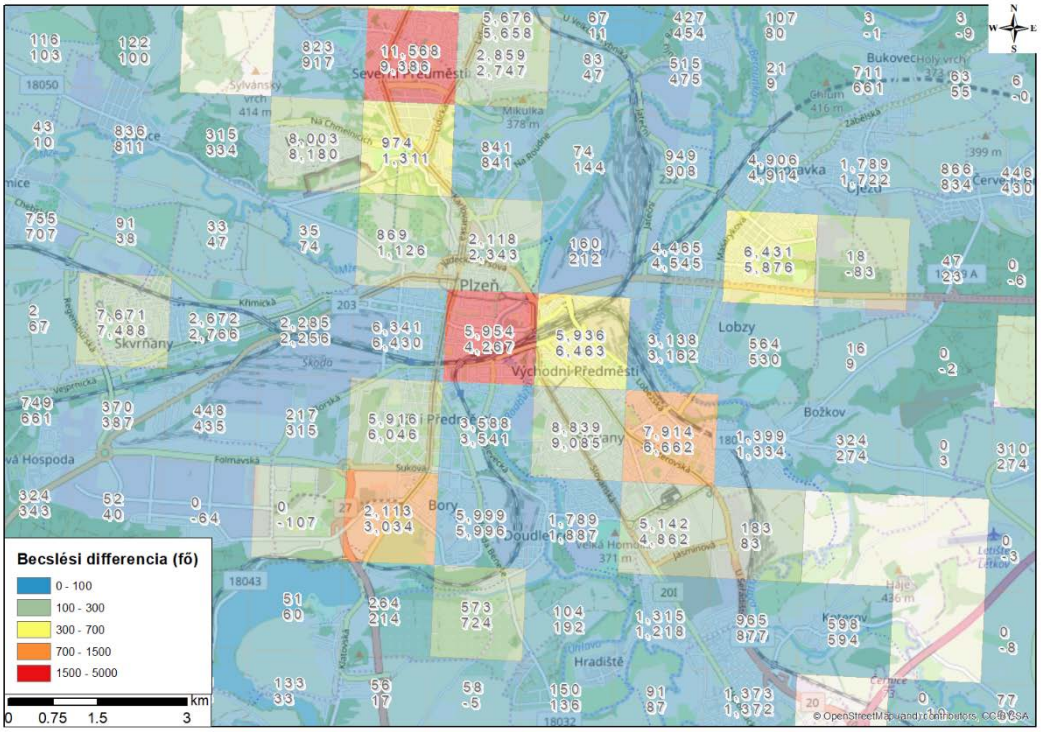
9. táblázat: a négy országos adatbázisra becsült népességértékek a modell változóinak szűrése esetén

Bemenő adatok		Pontos lakosság	Becsült lakosság	Különbség	MAE	RMSD
HU	1km*1km	9 767 012	9 779 643	12 631.08	58.57417	149.28139
	500m*500m	9 767 032	9 966 129.8	199 097.8	50.32133	138.42493
CZ	1km*1km	10 382 108	10 324 892.3	57 215.72	19.18314	92.59203
	500m*500m	10 382 153	12 529 867.7	2 147 715	39.31568	122.54240

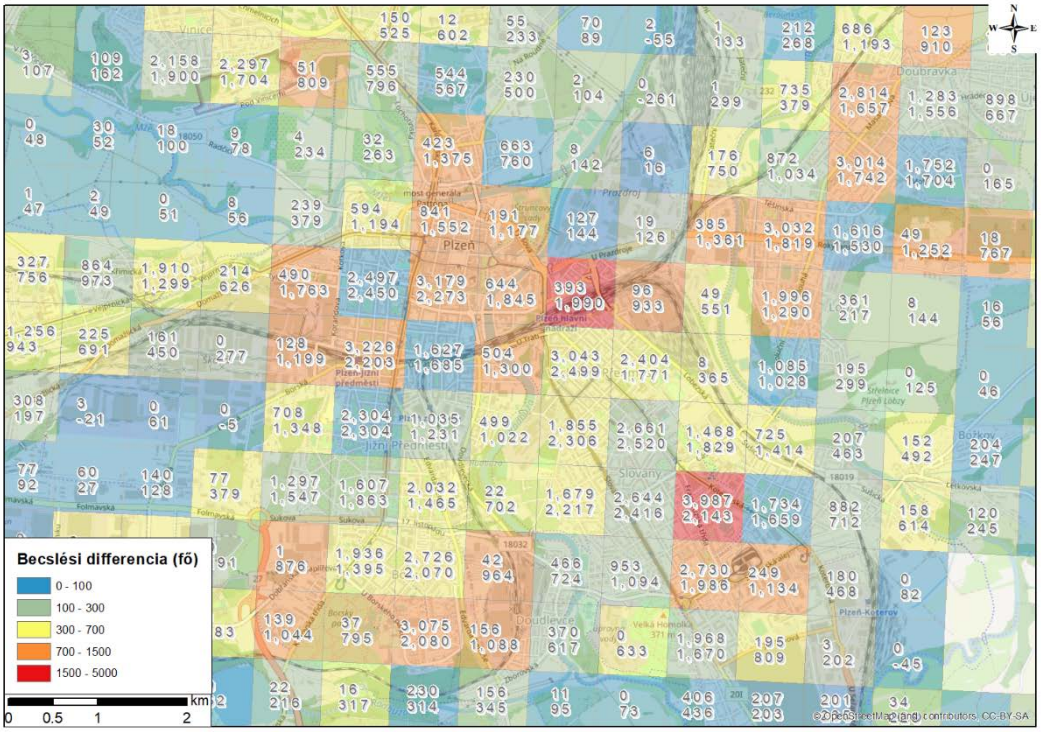
10. ábra: a népességbecslés pontossága Szeged 1:60 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt változókkal futtatott modell esetén.



11. ábra: a népességbecslés pontossága Szeged 1:40 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt változókkal futtatott modell esetén.



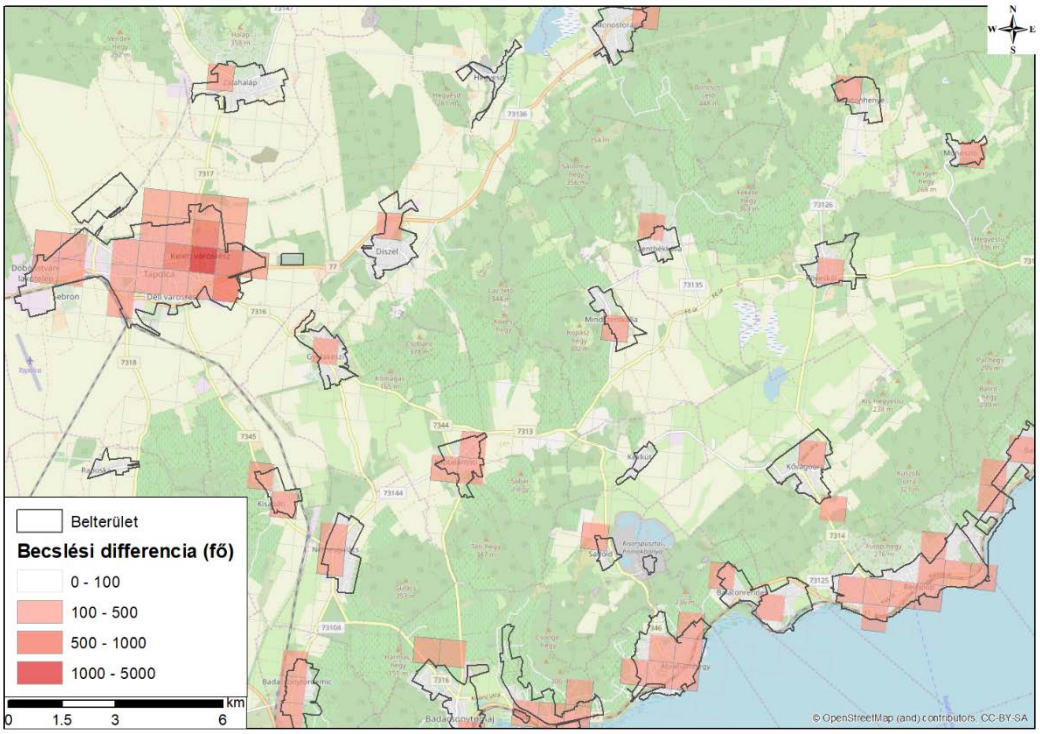
12. ábra: a népességbecslés pontossága Plzeň 1:60 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt változókkal futtatott modell esetén.



13. ábra: a népességbecslés pontossága Plzeň 1:40 000 OSM alapú térképén. A négyzetekben lévő felirat első sora a tényleges, a második a becsült népességszám, szűrt változókkal futtatott modell esetén.

7.2. Városközpont meghatározás

A becsült népességszámok valóditól való eltérése, azaz a differencia mező alkalmas települési centrumok pontos kijelölésére is, ugyanis felfedezhető



14. ábra: a becsült és tényleges népesség közötti különbség növekedése városközpontok esetén, a Balatonfelvidék 1:60 000 OSM alapú térképén.

15. ábra: a becsült és tényleges népesség közötti különbség növekedése városközpontok esetén, Jevišovice és környéke 1:60 000 OSM alapú térképén.

tendencia a becslési pontatlanság növekedésében a belvárosi területeken (14-15. ábra).

7.3. Gyakorlati tapasztalatok

A munkafolyamat kigondolása, a modell felépítése és működésre bírása során tapasztalatot gyűjtöttem a térinformatikai adatok, eszközök és fejlesztések terén. Adatok szempontjából a hivatalos forrásból származó (EUSTAT, KSH) adatbázisok esetében nem jelentett problémát az egységesítés, ellenben a szabadon szerkesztett adatok esetében (OSM) többször kellett saját szabályozók mentén generalizálnom, jelentős adatmennyiséget kiszűrtem pontatlanság, átfedés illetve gyakrabban a kategorizálás hiánya miatt. Kétségtelen, hogy ez az eljárás a szakterületre való megfelelő holisztikus rálátás nélkül jelentősen csökkentheti az eredmények reprezentatív jellegét.

A munkafolyamat tetemes része az adatbázis megteremtésével zajlott, ahol lehetett gyorsító, egyszerűsítő eljárásokhoz folyamodtam (CHAPMAN et al. 2000). Az adatmanipuláció tekintetében, az egyes esetekben százezres nagyságú táblák frissítését illetve módosítását jelentősen gyorsította az adatok indexelése és a programban megnyitott egyéb rétegek számának minimalizálása. Az adatelőkészítés repetitív természeténél fogva a munkán gyorsított az ArcMap eszközeinek egyszeri pontos kalibrálása utáni sorozatos újrafuttatása, egyedül a be- és kimeneti paramétereket cserélve.

7.4. Fejlesztési lehetőségek

A modell leghasznosabb fejlesztési pontjának a bemeneti adatbázisok körének bővítését tartom. Egy következő, erre a munkára épülő fejlesztés kezdeti lépéseként, az EU tagállamokra egységesen keletkező adatokat szükséges priorizálni az EUROSTAT grid alapegység egyszerű adaptálhatósága miatt. Ilyen például az EEA által 2017-ben szabadon elérhetővé tett EUDEM felszínmodell[25], amelyet integrálva e modellbe a tengerszint feletti magasság is független változóként szerepelhet, rámutatva az kertvárosi területek benépesülési gyorsaságának a tengerszint feletti magassággal való összefüggéseire. Ugyanakkor érdekes párhuzamot vonhatna egy tetszőleges mérettel és geometriával rendelkező

[25] https://www.eea.europa.eu/data-and-maps/data/copernicus-land-monitoring-service-eu-dem

alapegységgel való futtatás eredménye az előbbiekben bemutatott egyezményes méretű és formájú egységekkel futtatott példával.
A távlati céljaim között szerepel az ArcMap ArcPy alapú adatmanipulációs és a Python regressziós modell folyamatok egy kezelőfelülettel rendelkező programban való egyesítése, amely az előre meghatározott bemeneti fájlok betöltése után képes lenne végrehajtani a becslést egy felhasználói szintű user számára is.

8. Konklúzió

Diplomamunkám célja egy népességbecslő program létrehozása volt, amely költséghatékony módon szabadon hozzáférhető adatokat használva hajtja végre a becslést. A dolgozatban részleteztem a modellel szemben támasztott elvárásokat, a felhasznált adatokat, az adatbázis előkészítésének lépéseit, a Python modell létrehozását illetve a becsült népesség pontosságának értékelését mind adat, mind térbeli vonatkozásban.

Az adatbázist az EU minden tagországára egységes formátumban elérhető GEOSTAT2011 népességadatbázisból, a Corine Land Cover 2019 felszínborítottsági adatbázisból és a LAU 2 statisztikai egységek adatbázisából építettem fel. Továbbá integrálásra került a közösségi szerkesztés-alapú OpenStreetMap térképadatbázis számos rétege. A becslést Magyarország és Csehország adatain futtattam, felosztva az országok területeit 1 kilométeres és 500 méteres négyzetekre.

A munkafolyamat végén sikerült egy pontos becslésre alkalmas modell fejlesztése, amely megfelelő mennyiségű és minőségű bemenő adat esetében akár 0,05% eltéréssel képes megbecsülni a 2019-es hivatalos népességadatot. Ez a pontosság függ a vizsgált országtól, vizsgálati alapegységek méretétől és a modell pontos paraméterezésétől is. Az elsődleges funkcióján kívül a modell hasznosítható városközpont kijelölésére is. Ez a település centrumok körüli becslési bizonytalanság növekedésében jelentkezik, amelyek tematikus térképeken megjelenítve egyértelmű átfedést mutatnak a belvárosi zónákkal.

A jövőben tervezem folytatni a modell bővítését bemeneti adatok és kifinomultabb statisztikai eljárások terén, illetve a gépi tanulás (machine learning)

eljárások további lehetőségeinek feltérképezését és felhasználását az üzleti térinformatikában.

9. Summary

In this paper I present a statistical model that estimates population by using only openly availabe data instead of purchasing the most recent address level population data. The database used for the estimation consisted of two different groups of data. One being the European Union-wide, regulated and unified datasets of population (GEOSTAT 2011), land cover (Corine Land Cover), and administrative units of the settlements (LAU 2). The other datagroup used was from the globally accessible OpenStreetMap which has a vast list of infrastructural data, but is a less reliable source due to its crowdsourced nature. Given the free access to data on every country in the EU, Hungary and Czechia were chosen as testing areas due to their populations being similar sized.

Both countries had two different sized grids, a one kilometer and a 500-meter one, that were used as the base when executing the estimation. The statistical method of linear regression was used for finding the correlation between population and the various spatial factors such as road density, number of points of interest, and type of land use. With precise parameterizing, the estimation showed a minimal, 0,05% difference from the current sum of the population. Also, the difference between the estimated and the current population sum served as a precise variable for locating grids that had city center areas within them. With this model a cost effective tool has been created for different business GIS purposes.

10. Irodalomjegyzék

AVILA J.,–HAUCK T. 2017: scikit-learn Cookbook - Second Edition: Over 80 recipes for machine learning in Python with scikit-learn. – *Packt Publishing*, Birmingham. pp. 294-304.

BAKILLAH M.,–LIANG S.,–MOBASHERI A.,–JOKAR ARSANJANI J.,–ZIPF A. 2014: Fine-resolution population mapping using OpenStreetMap points-of-interest. – *International Journal of Geographical Information Science*. 28(9). pp. 1940–1963.

BAST H.,–STORANDT S.,–WEIDNER S. 2015: Fine-grained population estimation. – *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. article No.17

CHAINEY S.,–TOMPSON L.,–UHLIG S. 2008: The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. – *Security Journal*. 21(1-2). pp. 4–28.

CHAPMAN P.,–CLINTON J.,–KERBER R.,–KHABAZA T.,–REINARTZ T.,–SHEARER C.,–WIRTH R. 2000: Crisp-dm 1.0 step-by-step data mining guide. Technical report. – *The CRISP-DM consortium*. pp. 20-22.

DOMÁN CS. 2005: Többváltozós korreláció- és regressziószámítás Oktatási segédlet. – *Miskolci Egyetem Gazdaságtudományi Kar Üzleti Információgazdálkodási és Módszertani Intézet Üzleti Statisztika és Előrejelzési Tanszék*. pp 4-6.

FERANEC J. 2016: Project CORINE Land Cover. – In: FERANEC J.–SOUKUP T.–HAZEU G.–JAFFRAIN G. (szerk.): European Landscape Dynamics: CORINE Land Cover Data, *CRC Press*, Boca Raton. pp. 9-17.

GALWAY L.P.–BELL N.–SAE A.S.–HAGOPIAN A.–BURNHAM G.–FLAXMAN A.D.–WEISS W.M.–RAJARATNAM J.K.–TAKARO T.K. 2012: A two-stage cluster sampling method using gridded population data, a GIS, and Google EarthTM imagery in a population-based mortality survey in Iraq. – *International Journal of Health Geographics* 11. Article number: 12. pp. 2-4.

KESSLER F.,–BATTERSBY S. 2019: Working with Map Projections: A Guide to their Selection. – *CRC Press*. Boca Raton. pp. 66-67.

KOUNADI O.,–RISTEA A.,–LEITNER M.,–LANGFORD C. 2018: Population at risk: using areal interpolation and Twitter messages to create population models for burglaries and robberies. – *Cartography and Geographic Information Science*. 45:3. pp. 205-220.

LONDÖGÅRD H.–LINDBLAD H. 2018: Improving the OpenStreetMap Data Set using Deep Learning. – *https://lup.lub.lu.se/student-papers/search/publication/8951995* pp. 26.

OLIPHANT T. E. 2007: Python for Scientific Computing. – *Computing in Science & Engineering*. vol. 9. no. 3. pp. 10-20.

PÖDÖR Z. 2016: Többváltozós lineáris regresszió a gyakorlatban. – *Dimenziók: matematikai közlemények.* 4. pp. 50-51.

TOMS S.,–O'BEIRNE D. 2017 : ArcPy and ArcGIS - Second Edition. – *Packt Publishing*, Birmingham. pp. 60-124.

VÉRTESY L. 2019: Local Debt Burden at LAU2 level in the EU countries. – *European Financial Systems 2019. Proceedings of the 16th International Scientific Conference*. pp. 616-617.

NYILATKOZAT

Név: Magyar Márton Márk

ELTE Természettudományi Kar, szak: Geográfus

NEPTUN azonosító: NXO127

Diplomamunka címe:

Népességbecslés szabadon hozzáférhető adatok alapján

A **diplomamunka** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 20 _____

Magyar Márton Márk