Statisztikai Szemle

Közzététel: 2022. november 21.

A tanulmány címe:
Szövegbányászati változók előrejelzési lehetőségei gazdasági idősorokon és paneladatokon

Szerző:
FELLNER ÁKOS
PhD-hallgató. Pécsi Tudományegyetem, Regionális Gazdaság és Politika Doktori Iskola
E-mail: fellner.akos@ktk.pte.hu
DOI: https://doi.org/10.20311/stat2022.11.hu0999

Fellner Ákos

Szövegbányászati változók előrejelzési lehetőségei gazdasági idősorokon és paneladatokon

The Impact of Online Text Mining Variables on Economic Time Series and Panel Data Forecast

Fellner Ákos, a Pécsi Tudományegyetem Regionális Gazdaság és Politika Doktori Iskolájának PhD-hallgatója
E-mail: fellner.akos@ktk.pte.hu

Tanulmányomban kísérletet teszek az online gazdasági szövegbányászat legfontosabb kérdéseinek összefoglaló bemutatására, különös tekintettel a gazdasági előrejelzésekre. Ismertetem a szófelhők és a szövegprofilok előállításának módjait, illetve az elmúlt évtized legfontosabb kutatási irányait. A tanulmány során egy egyszerű szófelhős vizsgálat (Google Trend) és egy endogén gazdasági mutató (GDP) kapcsolatával illusztrálva bemutatom az idősoros és a panelregressziós *forecast* lehetőségeit.

Kulcsszavak: online szövegbányászat, *forecast*, idősoros vs panelregresszió

The paper deals with the usage of online text mining in economic forecast modelling and in research of innovation milieu and tacit investment attitudes. Firstly the basics of online text mining algorithms and its statistical relevancies are presented, especially the uprising problem of statistical categorization of word clouds and/or textual profiles. The last part of the paper contains short research about GDP and text mining forecast models in four USA states between 2010 and 2015, focusing on the forecast accuracy of time series models and panel regressions.

Keywords: online text mining, forecast, time series vs panel models

A szövegbányászat legnagyobb előnye a gazdasági kutatások terén az, hogy az eljárás online verziójában nagyon rövid idő alatt sok, elsősorban puha vagy rejtett *(tacit)* információt tudunk gépi tanulásos vagy egyéb úton kinyerni. Hátránya az erős nyelvfüggőség és az ökonometriai modellekben való elhelyezhetőség nehézségei (elsősorban a megfelelő aggregációs szint megválasztása idősoros vagy panelmodellekben). Puha vagy rejtett információk alatt olyan információkat értünk, amelyek leginkább a gazdasági szereplők viselkedésével, várakozásaival kapcsolatosak. A szövegbányászat ezeket kulcsszavakkal, szófelhőkkel vagy szövegprofilokkal ragadja meg, oly módon,

hogy statisztikailag klaszterezi azt, hogy bizonyos gazdasági szereplők milyen szófelhőkkel dolgoznak, illetve milyen jellegzetes beazonosítható szövegprofillal rendelkeznek.

Módszertani értelemben az online gazdasági szövegbányászatnak általában kétféle iránya van. Az egyik egy monolitikus építkezés, amennyiben a kutatásnak már létezik egy előzetes prekoncepciója, a szövegbányászat pedig ennek igazolására, avagy elvetésére, esetleg kidolgozására használatos. A másik esetben az online szövegbányászatnak a gazdaságban heurisztikus funkciója van, azaz a kutatás nem rendelkezik előzetes prekoncepcióval. A gazdasági előrejelzésekre alkalmazott online szövegbányászat rendszerint ez utóbbinál használatos, tekintve, hogy *forecasting* esetében nem az a feladat, hogy igazoljunk egy hipotézist, hanem az, hogy feltárjuk azokat a rejtett hatásokat (a modellben általában exogén változókként szerepelnek), amelyek befolyásolhatják a *forecast* modell pontosságát.

Az online szövegbányászat története során először puszta információkinyerés volt az elsődleges cél. Az evolúciós folyamat első kérdésköre arra fókuszált, hogy miként lehet elsősorban statisztikailag csoportosítani a releváns szavakat, szövegelemeket, annak definiálása mellett, hogy mitől számít egyáltalán relevánsnak egy nyelvi elem a keresett kérdéskör szempontjából *(Jiang, 2012).* Fontos további lépés volt a szövegek összegzésének kérdésköre. A számtalan különböző, de egymásra vonatkozó szöveg esetén szükség volt olyan, gépi tanulásos alapú összefoglalására, amely az adott témában való megértést segítette *(Nenkova–McKeown, 2012).* Ezután következett a különböző nagyobb szöveg-aggregátumok képzése, klaszterezése. Klasszikusan a bibliometriai felmérések, illetve a metaanalízisek *(Aggarwal–Zhai, 2012)* ilyenek. Jelentős problémának bizonyult a látens szemantikus tartalmak szövegbányászati kinyerése *(Crain et al., 2012)*. Fontos előrelépés az öntanuló szövegbányászati algoritmusok kifejlesztése, ami humán ágens közreműködése nélkül végzi a nyelvi elemek modularizációit *(Aggarwal–Zhai, 2012).* A legújabb kutatások a multimédiás szövegbányászat irányába tartanak, mikor is már nem csupán az internetről letöltött szövegek elemzése a fő cél, hanem azoknak az audiovizuális adatoknak az értelmezése, amelyek szövegeket tartalmaznak *(Zha et al., 2012).*

Tanulmányomban először bemutatom az online szövegbányászat alapvető módszereit, a szófelhők és szövegprofilok előállításának legfontosabb lépéseit. Ezután az elmúlt évtized legfontosabb szövegbányászati módszereit igénybe vevő gazdasági kutatásokat ismertetem. A tanulmány utolsó részében a gazdasági szövegbányászatot felhasználó modellezés legfontosabb ökonometriai problémáját, a különböző aggregáltsági szintekből fakadó dilemmát vázolom fel egy konkrét példán keresztül.

1. A kulcsszavaktól a szövegprofilig

Az online szövegbányászat alapvetően két formában jelenik meg a gazdasági modellezésben: egyrészről kulcsszavak, szófelhők formájában, másrészről szövegprofilok idősoros vagy panelregressziós előrejelző modelljeiben mint exogén magyarázóváltozó. Az endogén vagy függő változók ezekben a modellekben rendszerint a termelési vagy egyéb monetáris függvényekből ismert mutatók.

Kulcsszavak vagy kulcsszófelhők esetében a releváns kulcsszavak kiszűrése a fő feladat, amit egyrészről idősoros szegmentációval, másrészről a sajátvektoros modularitáson alapuló klaszterezéssel érünk el, egyszerűbb esetben indexálással. A szövegprofilok esetében ezzel ellentétben rendszerint grammatikai elemek mentén vagy szentiment(vélemény-)analízissel létrehozott szövegekre vagy szavakra súlyozott indexértékkel állnak elő a változók. Ezeket az értékeket ma már rendszerint szövegelemző szoftverek állítják elő, így nem szükséges az egyes nyelvi elemek manuális számolása. Például a Google Trend szókeresője az egyszerű, kulcsszavas keresésekre igen elterjedt alkalmazás. Amennyiben nem értelmezhető idősor a nyelvi változókra, úgy a fix vagy a véletlen hatáson alapuló panelregresszió a lehetséges modellezési keret. A szövegbányászati változó a panelmodellekben is – az idősoros modellekhez hasonlóan – rendszerint exogén magyarázóváltozó.

Mind a szófelhők, mind a szövegprofilok esetében statisztikai értelemben klaszterezésről beszélünk. Mindkettő voltaképpen egy, az időben többé-kevésbé állandó klaszter, rendszerint valamilyen hányadosértékként jelentkezik a regressziókban mint exogén magyarázóváltozó. A szövegprofil abban különbözik a szófelhőtől, hogy olyan grammatikai összefüggéseket is vizsgál, amit a szófelhős felmérés nem. A szövegprofil tehát jóval összetettebb szöveg-bányászati elem, mint csupán a szavak előfordulására, vagy azok statisztikai távolságára alkalmazott szófelhő.

A szövegprofilok úgy állnak elő, hogy egyes mintákon megnézik a különböző grammatikai és lexikai, esetlegesen stilisztikai elemek előfordulásának arányait, majd súlyozzák a szöveg teljes egészére nézve. Ezután empirikusan tesztelik az eredményeket egy másik hasonló mintán. Ha megfelelő korrelációkat találnak, illetve a minták közötti statisztikai tesztek is megfelelőek (leginkább a mintafüggetlenséget és a hibatagok kovarianciáját mérő tesztek a legfontosabbak), akkor az adott nyelvi elemek egy ún. tanítószótárba kerülnek, majd egységes szövegprofillá alakulnak. Az esetek többségében ezt a folyamatot folyamatosan iterálják, vagy gépi tanulásos módon, vagy humán ágensek segítségével.

A gépi tanulásos eljárás azt jelenti, hogy az adatbázist modulációelméleti klaszterezéssel felosztják teszt-, tanító és kiértékelő részre, és egy algoritmus a teszt és a keresztvalidációs hibatagok eloszlási fokokra vonatkoztatott eloszlását folyamatosan újra klaszterezi. Amennyiben a teszt és a keresztvalidációs regresszió hibatagja magas, úgy alacsony találati illeszkedésről beszélünk, azaz a modellünk túl kevés eseményt magyaráz. Amennyiben a tesztregresszió hibatagja alacsony és a keresztvalidációs hibatag magas, abban az esetben túldeterminált találati illeszkedésről beszélünk, azaz a modellünk túl sok eseményt magyaráz. A gépi tanulásos algoritmus feladata ebben az esetben, hogy az optimális hibaeloszlási sávban iterálja a teszt és a validációs adatbázis elemeit, és folyamatosan újraoptimalizálja a klasztereket.

A gépi tanulásos módszer egyik hatalmas hozzáadott értéke, hogy komplex iterációs folyamaton keresztül megtalálja az optimálisan illeszkedő adatbázis-klasztereket. Ezzel elkerülhető például az alacsony találati illeszkedés (amikor a teszt- és a keresztvalidációs regresszió hibatagja egyaránt magas), ami arra utal, hogy a modell túl kevés eseményt magyaráz. Ugyanakkor a másik véglet, a túldeterminált illeszkedés is megelőzhető (amikor a tesztregressziós hibatag alacsony, de a keresztvalidációs regressziós hibatag magas).

2. A szövegbányászat alkalmazása a gazdasági életben

A gazdasági előrejelzések témája igen sokrétű lehet, de az online szövegbányászat valamennyi esetében használható, amennyiben megfelelő mennyiségű és minőségű adat nyerhető ki az elemzéshez. Bár az aggregáció kérdésköre esetében makroökonómiai példát fogunk megvizsgálni, ettől függetlenül az online szövegbányászat *forecast* célokra nem csupán makroökonómiai modellek esetén alkalmazható. Az alábbiakban összegyűjtöttünk pár példát a teljesség igénye nélkül, amelyek nagyon jól mutatják azokat a területeket, ahol a gazdasági előrejelzés online szövegbányászatot használ.

Robert J. Shiller (2017) az 1930-as és a 2008-as gazdasági világválság narratíváit vizsgálva idősorosan kimutatta, hogy egyes nyelvi elemek és gazdasági válsághullámok a vírusos járványokra jellemző mintázattal terjednek a gazdaságban, így arra a következtetésre jutott, hogy ezeknek a nyelvi elemeknek (jellemzően gazdasági életre vonatkozó metaforák) trendszerű megjelenése a járványok terjedési mintázata szerint előre jelezhet gazdasági válságokat. Shiller

a Kermack–McKendrick-féle matematikai fertőzésmodell mentén vizsgálódik, ami alapvetően a fertőzött populációra mint adatbázisra épül, azonban számos egyéb extern hatást nem vesz figyelembe. Mindenesetre a narratív gazdaságtan területén forradalmi meglátás volt, hogy a gazdasági híreket a fertőzési mintázatok terjedésének analógiájára gondolta el.

Azqueta-Gavaldon (2020) három fő makrogazdasági mutató mentén (politikai bizonytalanság, kereskedelmi bizonytalanság, illetve belföldi szabályozás) vizsgálta az Európai Uniós befektetési anomáliákat 2000 és 2019 között, negyedéves idősoros bontásban. A szófelhőt gazdasági és politikai folyóiratok cikkeiből nyerték ki, SVAR- *(Structural Vector Autoregression)* analízis segítségével. Ezeket vetették össze VAR-regresszió segítségével, idősoros makrogazdasági mutatókkal Spanyolország, Németország, Franciaország és Olaszország esetében. Az eredmények azt mutatták, hogy Franciaország, Olaszország és Spanyolország vonatkozásában a befektetések bizonytalanságai leginkább a politikai bizonytalanságokhoz voltak köthetőek, a német gazdaság pedig leginkább a kereskedelmi bizonytalanságokra reagált rosszul. A belföldi szabályozások bizonytalanágai pedig leginkább az olasz és a spanyol befektetésekre voltak negatív hatással.

A pénzügyi kockázati előrejelzés és a szövegbányászat szempontjából *Li, Cai és Hu (2021)* kutatásai ígéretesek. Pénzügyi kockázati kompozit indexek idősoros ADL- *(Autoregressive Distributed Lag)* vizsgálatával kísérelték meg a már meglevő kompozit indexeket további extern hatások változóival úgy bővíteni, hogy az alkalmazott GARCH- *(Generalised Autoregressive Conditional Heterocedasticity) forecast* modell információs kritériuma és becslése pontosabb legyen. A kompozit index endogén változói között szerepelt a pénzügyi rendszer törékenységi mutatója, a pénzügyi innovációk száma, a hitelezési hajlandóság és a morális pénzügyi kockázatok. Az idősorokban megkíséreltek rezsimeket elkülöníteni, majd szövegbányászati módszerrel ezen rezsimek időtartamában végeztek kulcsszavas kereséseket pénzügyi cikkek szövegeiben. A kulcsszavas keresés olyan elemekkel tudta a pénzügyi kockázatindexet bővíteni, amelyek a GARCH-előrejelzéseknek sokkal nagyobb pontosságot és megbízhatóságot adtak.

Cicea és Marinescu (2021) publikációmetrikai analízis segítségével vizsgálta a gazdasági kibocsájtás és a külföldi tőkebefektetések kapcsolatát. Speciális klaszteranalízisre épülő eredmények szerint az elmúlt évtizedben a gazdasági növekedéssel kapcsolatos jellegzetes szófelhők a kemény gazdasági mutatók felől fokozatosan áttevődtek a puha gazdasági mutatók irányába.

Szintén publikációmetrikai metaanalízist végzett *Zhao, Seibert és Lumpkin (2010),* akik a Hunter–Schmitt-modell mentén arra keresték a választ, hogy a szakirodalmak 2007-ben mit tartottak a legfontosabb különbségnek az innová-

torok és a menedzserek gondolkodásában, valamint hogy ezek a különbségek milyen összefüggésben voltak a növekedési és a profitabilitási mutatókkal. A Hunter–Schmitt-modell egy olyan, modularitáselven működő osztályozási módszertan, amely elsősorban nagy elemszámú mintán működik megfelelően. Az eredmények azt mutatták, hogy a profitabilitási és a növekedési mutatók az innovátorok szemszögéből ott voltak a legmagasabbak, ahol alapvetően a céges gondolkodást a nyitottság jellemezte. Ellenben azoknál a cégeknél, ahol nem az innovativitás volt előtérben, hanem a piacszerzés, sokkal inkább a kockázattűrő mutatók voltak jellemzőbbek.

Li, Shang és Wang tanulmánya *(2013)* arra keresi a választ, hogy hogyan lehet a fogyasztóiár-index volatilitásának változásait előre jelezni a közösségi média releváns kulcsszavainak segítségével, szegmentált idősorokon. Modelljük makrogazdasági mutatók volatilitásrezsimjeinek autoregresszív elemzését végzi, majd összehasonlítja azon releváns kulcsszavak ugyanazon autoregressziós vizsgálatával. Eredményül azt kapták, hogy a szövegbányászati változóval specifikált MIDAS *(Mixed Data Sampling) forecast* modell sokkal nagyobb találati pontossággal jelzi a fogyasztóiárindex-volatilitást, mint az ugyanazon idősorokból számított ARMA- *(Autoregressive Moving Average),* illetve ADL- *(Autoregressive Distributed Lag)* modellek.

A regionális gazdasági kutatások terén kiemelkedő *Obschonka és munkatársainak* kutatása *(2020),* amelyben amerikai startupcégek Twitter-bejegyzéseinek Big5[1]-elemzésével mérték azok regionális gazdasági kibocsátásának változását 2009 és 2015 között. A fő kérdés, amire a választ keresték, az volt, hogy melyik startupmentalitás járul hozzá leginkább a regionális gazdasági kibocsátás növekedéséhez. A szövegkorpuszt frissen indult startupcégek Twitter-bejegyzéseiből állították össze, elemzését pedig a *World Well Being Project* és a Big5 angol nyelvű tanítószótárára alapozták. A regionális gazdasági kibocsátás endogén változóinak a következő mutatókat választották: népsűrűség, munkanélküliségi ráta, átlagjövedelem, egyetemi végzettség, ipari koncentrációs index. Exogén változó pedig a térbeli súlymátrix és a cég Big5-szövegprofilindexének értéke (szoftver által számolva). Panelregressziós eljárással, Pooled OLS-becsléssel arra az eredményre jutottak, hogy azok a startupcégek adták a legnagyobb hozzáadott értéket a lokális kibocsátás terén, ahol leginkább a nyitottság és a tudatosság voltak a vezető értékek. További kutatások is igazolják, hogy erős korreláció mutatkozik a társadalmi és kulturális nyitottság, valamint a vállalkozói piaci hatékonyság között *(McClelland, 1961; Stuetzer et al., 2016).*

[1] Nyitottság, lelkiismeretesség, extraverzió, szorongás, neuroticizmus (*openness, conscientiousness, extroversion, anxiety, neuroticism*).

Block és munkatársai (2018) a kutatásuk során üzleti befektetők Big5-profilját próbálták megalkotni logisztikus regressziós modellben. Meglehetősen sokváltozós modelleken vizsgálták, hogy a befektetők Big5-profilja milyen összefüggésben áll a befektetői kör nagyságával, összetételével, az üzleti folyamatban elfoglalt helyükkel, a befektetés sikerével, vagy sikertelenségével. Bár látható, hogy utóbbi két kutatás nem kifejezetten *forecast* problémára épül, ennek ellenére szépen mutatja, hogyan használható az online szövegbányászat komplex gazdasági kérdésekben.

Szót kell ejtenünk a gazdasági szövegbányászat korlátairól is. Az egyik fontos probléma, hogy nehezen képezhetőek folytonos idősorok a szövegbányászati változókkal. Ha előrejelzésről van szó, az idősoros modell mindenképpen pontosabb előrejelzést mutat, mint a panelregresszió. Idősorok azért képezhetőek nehezen ezzel a változótípussal, mert az adatok forrása számos esetben hiányos, az emberek vagy beszélnek egy témáról, vagy nem. Természetesen a statisztikai metaanalízis lehet egy bizonyos szinten megoldás, de idősoros becslések pontosságát semmiképpen sem tudja produkálni. Számos esetben ilyenkor vagy panelregressziót vagyunk kénytelenek használni, vagy idősorra transzformálni a hiányos adatokat, azonban ezeknek a transzformációknak nagy a torzításuk. A következő gond, hogy az aggregációs szint nagymértékben tudja a szövegbányászati változók szórását befolyásolni. Szintén fontos probléma az univerzalizálhatóság kérdése: a nyelv egyén- és kultúrafüggő, míg a statisztika nem. Ezeket a modellspecifikáció esetében mind figyelembe kell venni (általában bináris változókként), de ez egyben sajnos szűkíti is a szövegbányászat felhasználási esélyeit.

3. Empíria: idősoros vagy panelregresszió?

Szövegbányászati változókat tartalmazó modellek esetében az aggregációs szint megválasztásának kérdése mellett a legfontosabb a hiányzó adatsorok problémája. Ugyanis nem bizonyos, hogy minden T-, illetve T–p-megfigyelési időpontra rendelkezésre áll adat. A probléma nem ugyanaz, mint a MIDAS-modellezés esetében, hiszen nem arról van szó, hogy az egyik magyarázóváltozó frekvenciája alacsony, a másiké pedig magas, hanem az, hogy adott esetben hiányosságok mutatkozhatnak. Ezek rendszerint úgy fordulnak elő, hogy a modell endogén változói kontinuus makrogazdasági mutatók, exogén változói pedig diszkontinuus frekvenciák. Ezekben az esetekben kérdéses, hogy az

előjelzési modell esetében idősoros, avagy panelregressziós modellt érdemes-e választani.

Tekintsünk át egy viszonylag egyszerű példát! Megvizsgáltam a Google Trend segítségével, hogy 2010 és 2015 között az USA négy keleti tagállamában (Connecticut, Georgia, Delaware, Florida) a felhasználók milyen gyakorisággal kerestek az *inflation* kifejezésre. Az adatokat éves és negyedéves, országos, illetve tagállami szintű bontásban aggregáltam, egyszerű átlagolással becsülve értékeket, majd a GDP vonatkozásában szintén az annak megfelelő országos, tagállami, éves és negyedéves adatokat alapul véve idősoros ARMA-modellre alapozott *forecast*ot (ARMAX), illetve fix és véletlenhatás-panelregressziót végeztem el, azt követően pedig mindegyik előrejelzésnek összehasonlítottam a Theil-féle U_2- *(unbiased)* index értékét. A Theil-féle U-index általánosan használt eljárás egy magyarázómodell előrejelző értékének megállapításához.

1. táblázat

A Theil-féle U_2- *(unbiased)* index értékei a *GDP* és az *inflation* szavakra való rákeresésre, éves és negyedéves, országos, illetve tagállami szinteken 2010 és 2015 között, ARMAX (1, 0, 1)-modell-előrejelzés alapján

The values of the Theil U_2- (unbiased) index for the GDP and Inflation word searches, on yearly and quarterly, national, as well as Member States level between 2010 and 2015, based on ARMAX (1, 0, 1) model prognosis

ARMAX (1, 0, 1) éves	Theil's U_2 (statikus)	Theil's U_2 (dinamikus)
USA	0,7343	3,1559
Connecticut	0,4983	0,5636
Delaware	0,6646	0,8771
Georgia	0,2324	0,4768
Florida	0,5757	1,4084
ARMAX (1,0, 1) negyedéves	Theil's U_2 (statikus)	Theil's U_2 (dinamikus)
USA	0,9026	9,3138
Connecticut	0,8022	1,3500
Delaware	0,9165	2,1083
Georgia	0,6174	6,2654
Florida	0,9780	4,9908

Forrás: Google Trend, Bureau of Economic Analysis of United States.

A Theil-féle U-értékek általánosan használtak a *forecast* modellek pontosságának, illetve megbízhatóságának a mérésére. A Theil-féle érték megadására két általánosan elfogadott módszer lehetséges, mindegyik a *forecast* hibadekompozícióján alapul. Az első módszer szerint a Theil-féle U_1-érték 0 és 1 közé eső érték, és minél inkább tendál 0 felé, a becslés annál pontosabb. A másik

számítási módszer szerint az U_2-érték a *forecast* modell pontosságát a naiv becsléshez képest határozza meg: ha az U_2-érték 1 közelében mozog, akkor az egyszerű naiv becslési eljárás megfelelőbb, mint a használt *forecast* modell, amennyiben 1 fölötti értéket vesz fel, gyengébb, amennyiben 1 alattit, erősebb becslést mutat, mint a naiv becslési eljárás. (Naiv becslésnek nevezzük, amikor t+1 értékét úgy becsüljük, hogy az előző t-differenciát egyszerűen hozzáadjuk.) A pontosság tekintetében az 1 alatti, minél kisebb U_2-érték a kívánatos (1. táblázat).

Láthatjuk, hogy idősoros *forecast* modelljeink előrejelzési pontossága minden aggregációs, illetve időbeli szegmentációs szinten meglehetősen rossz (magas U_2-értékek). Ez azt jelenti, hogy a szövegbányászati változók nem működnek megfelelően idősoros modellek esetében. Megvizsgáltam ugyanezt az adatbázist fix, random panelregressziós modellel, majd ugyanezen változókon a modellek *forecast* erősségét néztem meg, szintén Theil-féle U_2- indexek segítségével. Ezt a 2. táblázatban foglaltam össze.

2. táblázat

A Theil-féle U_2- *(unbiased)* index értékei a GDP és az *inflation* szavakra való rákeresésekre, éves és negyedéves, országos, illetve tagállami szinteken 2010 és 2015 között, fix és random hatások panelregressziója mentén

The values of the Theil U_2- (unbiased) index for the GDP and Inflation word searches, on yearly and quarterly, national, as well as Member States level between 2010 and 2015, along panel regression of fixed and random effects (In case of panel regressions the first digit stands for the cross-section division, the second one for the time series segmentation)

Panel, éves	Theil's U_2 (fix)	Theil's U_2 (random)
USA	0,73429	3,1559
Connecticut	0,00431	0,006
Delaware	0,00141	0,0304
Georgia	0,00076	0,0166
Florida	0,00212	0,0137
Panel, negyedéves	Theil's U_2 (fix)	Theil's U_2 (random)
USA	0,00143	0,0146
Connecticut	0,00219	0,0086
Delaware	0,00862	0,0117
Georgia	0,00280	0,0167
Florida	0,00395	0,0144

Megjegyzés: a panelregressziók esetében az első számjegy a keresztmetszeti felosztást, a második az idősoros szegmentációt tartalmazza.
Forrás: Google Trend, Bureau of Economic Analysis of United States.

Látható, hogy a panelregressziós előrejelzések sokkal megbízhatóbbak, mint idősoros társaik (nagyon alacsony U_2-értékek). Ez amiatt van, hogy a panelregressziók transzformációs eljárásai (azaz amikor egyszerre vizsgálódunk keresztmetszeti és idősoros módon) sokkal jobban kiegyenlítik a változók szórásából származó torzító hatásokat, mint idősoros modellek esetében. Természetesen ezen a kismintás vizsgálaton nem vonhatunk le általános következtetéseket valamennyi ilyen típusú regressziós vizsgálatra, ugyanakkor már egy ilyen egyszerű minta és modellezés is arra enged következtetni, hogy az aggregáltsági szintre a szövegbányászati magyarázóváltozók sokkal érzékenyebbek, mint egyéb gazdasági mutatók változói.

4. Összegzés

Az elmúlt évtizedben a gazdasági online szövegbányászat szerepe felértékelődött. Online szövegbányászati módszerrel kinyert magyarázóváltozót rendszerint akkor építenek be gazdasági (jelen esetben leginkább előrejelző) modellbe, amennyiben valamilyen olyan összefüggésnek keresik a gazdasági hatását, amelyre a statisztikai adatbázisok klasszikusan nem szolgáltatnak adatot. Bemutatott példáinkon láttuk, hogy ezek olyan összefüggések, amelyek többnyire valamilyen gazdasági szereplővel kapcsolatos, a statisztikai hivatalok adatbázisa számára általában rejtett változókhoz köthetőek (kockázati befektetők viselkedése, startupcégek belső értékrendje, a menedzser- és innovátorattitűdök különbsége, a közösségi média szereplői, a pénzpiaci környezet változásai).

Egy általunk összeállított egyszerű adatsoron (GDP és Google Trend) és négy alapvető előrejelző modell (statikus/dinamikus ARMAX, fix/random hatás panelregressziója) segítségével megvizsgáltam a kulcsszavakkal kapcsolatos aggregációs problémát. Azt tapasztaltam, hogy az idősoros, illetve a panelregressziós aggregáltsági szintek szignifikánsan befolyásolják az ugyanazon magyarázóváltozókat tartalmazó előrejelző modell hatékonyságát. Minél kisebb az aggregáltsági szint, különösen panelregresszió esetében, annál jobb a *forcasting* modell Theil-féle U_2-értéke. Idősoros aggregáció esetén azt feltételeztük, hogy ennek oka az aggregációs transzformáció során fellépő hibatag. A panelregressziós transzformációk esetében egyszerre alkalmazunk keresztmetszeti és idősoros módszert, ami az előrejelzés megbízhatóságát növelte, ám ennek ellenére az aggregációs szintek közötti különbség megmaradt. Nyilvánvaló hátránya a panelregressziós modelleknek, hogy egy adott időre viszont nem végezhetünk konkrét előrejelzést.

Irodalom

Aggarwal, C. C. – Zhai, C. (2012): A survey of text classification algorithms. In: *Mining text data.* pp. 163–222; 77–128. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_6

Azqueta-Gavaldon, A. (2020): *Text-mining in macroeconomics: the wealth of words.* Doctoral dissertation, University of Glasgow.

Block, J. H. – Fisch, C. O. – Obschonka, M. – Sandner, P. G. (2018): A personality perspective on business angel syndication. *Journal of Banking & Finance.* Vol. 100. pp. 306–327. https://doi.org/10.1016/j.jbankfin.2018.10.006

Cicea, C. – Marinescu, C. (2021): Bibliometric analysis of foreign direct investment and economic growth relationship. A research agenda. *Journal of Business Economics and Management*. Vol. 22. No. 2. pp. 445–466. https://doi.org/10.3846/jbem.2020.14018

Crain, S. P. – Zhou, K. – Yang, S. H. – Zha, H. (2012): Dimensionality reduction and topic modeling: From latent semantic indexing to latent dirichlet allocation and beyond. In: *Mining text data.* pp. 129–161. Springer, Boston, MA.

Jiang, J. (2012): Information extraction from text. In: *Mining text data.* pp. 11–41. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_2

Li, X. – Shang, W. – Wang, S. (2013): Incorporation of Social Media Data into Macroeconomic Forecast Systems: A Mixed Frequency Modelling Approach. In: *PACIS.* p. 57.

Li, Z. – Cai, Y. – Hu, S. (2021): Research on systemic financial risk measurement based on HMM and text mining: a case of China financial market. *IEEE Access.* Vol. 9. pp. 22171–22185. https://doi.org/10.1109/ACCESS.2021.3055967

McClelland, D. C. (1961): *Achieving society.* Vol. 92051. Simon and Schuster. https://doi.org/10.1037/14359-000

Nenkova, A. – McKeown, K. (2012): A survey of text summarization techniques. In: *Mining text data.* pp. 43–76. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_3

Obschonka, M. – Lee, N. – Rodríguez-Pose, A. – Eichstaedt, J. C. – Ebert, T. (2020): Big data methods, social media, and the psychology of entrepreneurial regions: capturing cross-county personality traits and their impact on entrepreneurship in the USA. *Small Business Economics.* Vol. 55. No. 3. pp. 567–588. https://doi.org/10.1007/s11187-019-00204-2

Shiller, R. J. (2017): Narrative economics. *American Economic Review*. Vol. 107. No. 4. pp. 967–1004. https://doi.org/10.3386/w23075

Stuetzer, M. – Obschonka, M. – Audretsch, D. B. – Wyrwich, M. – Rentfrow, P. J. – Coombes, M. – Satchell, M. (2016): Industry structure, entrepreneurship, and culture: an empirical analysis using historical coalfields. *European Economic Review*. Vol. 86. pp. 52–72. https://doi.org/10.1016/j.euroecorev.2015.08.012

Zha, Z. J. – Wang, M. – Shen, J. – Chua, T. S. (2012): Text mining in multimedia. In: *Mining Text Data.* pp. 361–384. Springer, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_11

Zhao, H. – Seibert, S. E. – Lumpkin, G. T. (2010): The relationship of personality to entrepreneurial intentions and performance: A meta-analytic review. *Journal of management.* Vol. 36. No. 2. pp. 381–404.