Külgazdaság, LXVI. évf., 2022. november–december (28–44. o.)

A makroökonómiai válság-előrejelzés lehetőségei szövegbányászatilag specifikált random panelregressziós modellekkel

FELLNER ÁKOS*

A gazdaságiválság-előrejelzést két klasszikus nézet jellemzi. Az egyik szerint válság-előrejelzésre az idősoros modellek a legalkalmasabbak, a másik szerint a legerősebb válságszignál a fogyasztói árindex, illetve a különböző befektetői bizalmi indexek változása. Ez a tanulmány magyar példán keresztül mutatja be a szövegbányászati módszerekkel specifikált random hatású panelmodell (REPR) működését amellett érvelve, hogy a modell hibatagjainak értékei jóval pontosabb összefüggések feltárását teszik lehetővé panelregresszió használatával, mint idősoros modellek alkalmazásával, illetve válságszignálok esetén sokkal fontosabb a külkereskedelmi árindexek monitorozása, mint a fogyasztói árindexé vagy a befektetői bizalmi indexé.

Journal of Economic Literature (JEL) kódok: C10, C80, E17.
Kulcsszavak: válság-előrejelző modellek, gazdasági szövegbányászat, panelregresszió.

Abstract

The potential of macroeconomic crisis forecasting with text-mining specified random panel regression models

ÁKOS FELLNER

There are two classical views on economic crisis forecasting: one is that time series models are the most appropriate for projections of crises, and the other is that the strongest crisis signal is the change in the consumer price index and the various investor confidence indices. This paper

* *Fellner Ákos* PhD-hallgató, Pécsi Tudományegyetem Regionális Gazdaság és Politika Doktori Iskola. E-mail: fellner.akos@ktk.pte.hu

A kézirat 2022. december 12-én érkezett szerkesztőségünkbe.
https://doi.org/10.47630/KULG.2022.66.11-12.28

presents a Hungarian example of a random effects panel model (REPR) specified by text mining methods, arguing that the values of the error terms of the model allow for a much more accurate correlation using panel regression than using time series models, and that monitoring the foreign trade price index is more important than considering the consumer price index or the investor confidence index for crisis signals.

Journal of Economic Literature (EL) codes: C10, C80, E17.
Keywords: economic crisis forecast, economic text mining, panel regressions.

Bevezetés

A gazdasági szövegbányászat az elmúlt közel másfél évtizedben került a gazdasági modellezések specifikálásának előterébe. Ennek oka, hogy a szövegbányászat módszereivel olyan rejtett gazdasági hatások szűrhetők ki, amelyek lényegileg befolyásolhatják a leíró vagy előre jelző modelleket. A válság-előrejelzés két fő gondolatkörköz kapcsolódik: egyrészt a modellek szükségszerűen idősorosak, másrészt a fogyasztói árindex és különböző befektetői indexek a válságok megfelelő indikátorai. Ez a nézet azonban a nemzetközi szakirodalom alapján is vitatható. Az idősoros modellek legnagyobb problémája, hogy csak erősen torzítva tudják kezelni azokat a rejtett változókat, amelyek a klasszikus makrogazdasági mutatók mellett specifikálni képesek az előrejelzéseket. A fogyasztói, illetve a befektetői bizalmi indexekkel szembeni legfőbb ellenvetés, hogy ökonometriai értelemben nem bizonyítható egyértelműen, hogy ezeknek a mutatóknak valóban szükségszerű oksági kapcsolatuk van-e a válsághullámok megjelenésével.

Ez a tanulmány ezért olyan alternatív modellspecifikációt vizsgál, amely mindkét nézettől eltávolodik. Az előre jelző modellek nemzetközi szakirodalmában szintén régóta jelen van a panelregresszió módszertana, ami több szempontból szerencsésebb az idősoros modelleknél. Legfőbb jellemzője, hogy olyan diszkontinuus magyarázó változókat tartalmaz, amelyek az idősoros modellekben nem szerepelnek. A szövegbányászati változók diszkontinuus változók, ezért idősoros modellek specifikációiban történő szerepeltetésük megkérdőjelezhető. A panelregresszió sokkal rugalmasabb keretet nyújt a diszkontinuus szövegbányászati változók alkalmazásához. A később részletezett eredményeim egyben alátámasztják a külgazdasági multiplikátorhatás működését, azaz igazolják, hogy az árindexek közül az importárindex a fogyasztói árindexnél sokkal fontosabb válságindikátor.

A tanulmány először a gazdasági szövegbányászat néhány gazdasági kontextusát tekinti át, majd az árindexek és befektetői bizalmi index ökonometriai alkalmazhatóságáról tesz kritikai megállapításokat. Ezután a modellezés alapkérdéseit vizsgálja, majd hazai adatokon mutatja be a random panelregresszió működését választott kulcsszavak alapján. Arra keresi a választ, hogy az idősoros vagy a szövegbányászati szempontból specifikált panelmodell rendelkezik-e jobb hibatesztekkel, illetve igazolódik-e, hogy a külgazdasági multiplikátorhatás terén az importárindex változása fontosabb válságszignál, mint a fogyasztói árindexé.

A szövegbányászat néhány gazdasági kontextusa

A makroökonómiai válság-előrejelzési modellek terén az elmúlt évtizedben a nemzetközi szakirodalomban különleges szerepet kaptak a szövegbányászati módszereket használók. A *szövegbányászati változó* egy ökonometriai modellben egy bizonyos *kulcsszónak, kulcsszófelhőnek* (kulcsszavak csoportjának) vagy egy szöveges tulajdonság értékének az időbeli frekvenciája. A szövegbányászati változó időben diszkontinuus. *Szövegbányászati módszernek* nevezik azt a szegmentációs eljárást, amely valamilyen statisztikai algoritmus szerint csoportosítja ezeket a kulcsszavakat, kulcsszófelhőket vagy szöveges tulajdonságokat. Ezek azért rendelkeznek magas hozzáadott értékkel a modellspecifikációban, mert *tacit összetevőket* tudnak beemelni a modellalkotásba. A gazdaságtanban, illetve az ökonometriában a tacit, vagy más néven rejtett tényező nem jelenti szükségszerűen ugyanazt. Az előbbiben rendszerint rejtett tényezők azok az összetevők, amelyek valamilyen formában az emberi tevékenységhez vagy annak kulturális kontextusához köthetők. Az ökonometriában azt a változót nevezik rejtettnek, amelynek hibatagja autokorrelál a modell hibatagjával, vagy erős multikollinearitást mutat a többi változóval. A két meghatározás nem esik szükségszerűen egybe. Ez a tanulmány a rejtett hatásokat az első értelmezés szerint használja. A válság-előrejelző modellek tacit összetevői azok a gazdasági vagy társadalmi folyamatok, amelyek nem a klasszikus makroökonómiai idősorokban aggregálódnak. Ilyen klasszikus tacit indikátorok a különböző gazdasági bizalmi indexek, amelyek a befektetők befektetési, illetve a fogyasztók fogyasztási és/vagy megtakarítási hajlandóságát hivatottak mérni, azonban a köztük lévő valós oksági összefüggéseket nem képesek igazolni. Ezek az indexek alkalmasak ugyan arra, hogy általános képet közvetítsenek a piaci várakozásokról és idősorosan

is megadhatók, nem alkalmasak azonban okságilag megbízható következtetések levonására.

Az online szövegbányászati módszerekkel kinyert magyarázó változók sokkal árnyaltabb képet adnak a makrogazdasági folyamatok változásáról, mint az általános bizalmi indexek, azaz ökonometriai értelemben sokkal kisebb a kovarianciájuk a modell hibatagjával (Li et al., 2021). A szövegbányászati modellspecifikáció lényege, hogy a válságciklusokat, illetve azokat a rezsimváltásokat, amelyek a gazdasági válság-előrejelző mutatók idősorait jellemzik, valamilyen szófelhő megjelenésével előre jelzik. Ez a szófelhő azoknak a szavaknak az előfordulása, amelyek tipikusan bizonyos válságok előtt jelentkeznek a gazdasági diskurzusokban és az általános közbeszédben.

Klasszikus válság-előrejelzési mutatók a különböző *árindexek* és a befektetői jövőképet, illetve hajlandóságot mérő indexek. Ezek arról informálnak, hogy az adott gazdasági tevékenységnek mennyi a költsége, azaz mennyire kifizetődő a gazdasági szereplőknek az adott tevékenység folytatása, illetve a befektetők egy adott országban mennyire bizakodóak befektetésük megtérülését illetően. Az árindexeket havi idősoros egységekben mérik, ami potenciális válságok begyűrűzésének megfelelő előrejelzése lehet. Ha egy adott árindex megnő, az azt jelenti, hogy a piaci szereplőknek többet kell fizetniük ugyanazért a termékért, szolgáltatásért vagy tevékenységért. Az árindexek azért mutatnak viszonylag hamar válságszignálokat, mert közvetlenül piaci hatásokat tükröznek, eltérően az olyan többszörösen aggregált makrogazdasági mutatóktól, mint a bruttó belföldi termék (GDP), a bruttó nemzeti termék (GNP), a bruttó hozzáadott érték (GVA) vagy a munkanélküliségi ráta. Klasszikus válságszignál a fogyasztói árindex, bár sokan mérik a válság begyűrűzését a munkanélküliségi ráta változásával. Ezek az adatok viszont sokszor nem pontosan tükrözik a valós válságfolyamatokat (például a szürkegazdaság hatása miatt).

A *külkereskedelmi multiplikátorhatás* miatt azonban vitatható, hogy valóban a fogyasztói árindex változása a legjobb árindexszignál válság-előrejelzésre. Harrod (1939) és Samuelson (1992:1232–1238) szerint ugyanis, ami igazából befolyásolja a keresleti és kínálati változásokat nyitott gazdaság esetén, az nem a belföldön előállított áruk vagy szolgáltatások belföldi piaci kereslete, hanem az exporttevékenység és az importtermékek fogyasztása. Ez a külkereskedelmi multiplikátorhatás lényege nyitott gazdaság esetében. Vagyis, amikor az a kérdés, hogy melyik árindexmutatót indokolt és célszerű vizsgálni egy válság begyűrűzésének gyorsabb feltérképezéséhez, akkor feltételezésem szerint érdemesebb az *import- és exportárindexek* változásait nyomon követni a belföldi fogyasztói árindex vagy a fogyasztói kosár alakulása

helyett. A gazdasági szövegbányászat szempontjából ebben az esetben az a döntő kérdés, hogy pontosabb modell alkotható-e akkor, ha szövegbányászati változók szerepelnek a modellben exogén változókként, illetve a külgazdasági folyamatokat mutató árindexek sokkal pontosabban jelzik-e előre a válságot, mint a klasszikusan használt fogyasztói árindex.

A nemzetközi szakirodalomban máig eldöntetlen kérdés, hogy ha online szövegbányászattal kinyert magyarázó változókat szerepeltetnek egy makrogazdasági előre jelző modellben, akkor arra idősoros, panelregressziós, esetleg metaanalitikus vagy modularitási elven működő algoritmusokban kerüljön-e sor.

A jelen tanulmány nem tárgyal metaanalitikus és modularitási elven működő modelleket, mert bemutatásuk szétfeszítené az írás tartalmi és terjedelmi kereteit. Ráadásul ezek az előre jelző *(forecast)* erő vonatkozásában nagyon bonyolult és erősen vitatható kimenetellel rendelkeznek. Idősoros modellek alkalmazása pedig azért nem indokolt, mert megbízhatóságuk erősen vitatható. Az idősoros elemzéshez ugyanis folytonos adatsorokra van szükség, miközben a szövegbányászati változókra az adatsorok vonatkozásában gyakran nem lehet biztosítani folytonosságot. Ezt legfeljebb különböző transzformációs eljárásokkal lehet elérni, amelyek viszont számottevő mértékben torzítják a minta eloszlását.

Az árindex és a bizalmi index válság-előrejelzésben

A gazdasági válságszignálok a *neoklasszikus megközelítések* szerint három hullámban jelentkeznek. Az első hullám a befektetések volumenét, a második a nemzetközi monetáris környezet megváltozását, a harmadik pedig az árképzések módosulását érinti (Zarnowitz & Moore, 1982). Vagyis a neoklasszikus értelmezés szerint az árindexek változása mutatja legutoljára a válságokat, mert először a termelői és az azt kiszolgáló pénzügyi láncolatok szignáljai az elsődlegesek, az árképzés a piacnak már csak az erre adott reakciója.

A *neokeynesiánus irányzat* szerint a sorrend fordított, felfogásában ugyanis az aggregált keresleti és aggregált kínálati oldal komplex változása okozza az infláció változását. Mindezt ezután követik a termelői és monetáris láncolatokban végbemenő változások (Nekarda & Ramey, 2020).

A különböző árindexek és bizalmi indexek esetében kérdéses az, hogy melyik tekinthető lényegesebbnek a válságfolyamatok vizsgálata, a válság-előrejelzés szempontjából. Nyitott gazdaság tételezése esetén nem indokolt kiemelt jelentőséget tu-

lajdonítani a fogyasztói árindexnek, mert a már hivatkozott külkereskedelmi multiplikátorhatás alapján feltételezhető, hogy a kibocsátást a klasszikus Solow-modellen túl az exporthatás és az importtermékek keresletének változása befolyásolja. Ezek a hatások ugyanolyanok, mint a beruházásokéi, illetve a kormányzati vásárlásokéi. Ha a válság külföldről gyűrűzik be (a krízisek az utóbbi száz évben világgazdasági jellegűek voltak) a nagy fokú világgazdasági nyitottságú és kis gazdasági dimenziójú országokba, akkor a külkereskedelmi multiplikátor-hatás miatt első körben a válságszignálokat sem a belföldi fogyasztás vagy a belföldi termelési láncok körében kell keresni. A külkereskedelmi multiplikátor részletes tárgyalása nem tárgya ennek az írásnak. A témáról kiváló és alapos történeti összefoglalót ad McCombie & Thirlwall (1999).

A befektetői és a fogyasztói várakozásokat, bizalmat mérő indexekkel *(Business Confidence Index, Consumer Confidence Index)* kapcsolatban a legnagyobb gond, hogy ökonometriai szempontból ezeknek az indexeknek az előre jelző ereje problematikus. Khan & Upadhayaya (2020) kimutatta, hogy az OOS *(Out-Of-Sample)* előrejelzések esetében az idősoros hibatagok kovarianciája nem igazolja egyértelműen a kapcsolatot ezen indexek változása és a gazdasági trendek között, illetve a változók idősoros elemzése nem igazán alkalmas oksági kapcsolatok magyarázatára. Mivel ezek az indexek OOS-módszeren alapulnak, a fenti indexek elemzése ugyan általános képet adhat egy válságciklus kialakulásáról, de nem biztos, hogy valós oksági összefüggés áll fenn közöttük.

Modelltípusok: idősoros, moduláris, panel

Az ökonometriában a *rezsimváltás* az idősoroknak az a tulajdonsága, amikor a változóknak valamilyen trendszerű jellemzője megváltozik. A válság ebben a megközelítésben felfogható egy rezsimnek, a válság előrejelzése pedig a rezsimváltás időpontjának (Hamilton, 2016). Kérdés, hogy miként lehetséges a rezsimváltásokat pontosabbá tenni szövegbányászati változók segítségével. Erre általában két megoldás kínálkozik: a monolitikus és a gépi tanulásos módszer. Az első metódus szerint a kutatók feltételeznek bizonyos összefüggéseket, és azokra futtatnak le kulcsszavas kereséseket. Megvizsgálják, hogy a kiválasztott kulcsszavak vonatkozásában hogyan alakulnak a rezsimváltások. A másik megoldás, hogy először a gazdasági mutatókon idősoros szegmentációt végeznek, és megállapítják a rezsimváltások időbeli helyét (ahol a maradványérték varianciája szignifikánsan változik). Majd a szegmentumokon gépi algoritmusok segítségével szófelhőket (azaz kulcsszavak klasztereit)

alkotnak. Ez utóbbi megoldás rendkívül idő- és költségigényes, ugyanis igen nagy mennyiségű szöveget kell analizálni, de ez tekinthető pontosabbnak és objektívebb megközelítésnek.

Az online idősoros szegmentáció kérdésében kiváló összefoglalót nyújt Keogh et al. (2001). Az idősoros szegmentáció alapjai vonatkozásában rendkívül hasznos Sclove (1983), illetve az idősoros rezsimek statisztikai összehasonlíthatóságáról Wang & Wang (2000).

A szövegbányászati változókkal dolgozó előre jelző modellek része az idősoros szegmentáción alapuló tartalomelemzés. Ezt azt jelenti, hogy (1) szegmentációs módszerekkel megvizsgálják a rezsimváltások időpontjait, majd (2) az adott időpontokban gépi tanulásos módszerrel szófelhőket képeznek a rendelkezésre álló online szöveganyagból. Ha nincs előzetes prekoncepció, akkor először el kell végezni az idősoros szegmentációt (ennek részletes módszertani bemutatásától jelenleg eltekintek), majd a rezsimváltások időpontjában szövegmodulációt kell végezni. Ez azt jelenti, hogy kulcsszavak vagy egyéb más szövegbányászati változók alapján modulokat képeznek. Erre több lehetséges eljárás létezik: klaszterezés, vektoranalízis, maradványérték-autoregresszió, neurális hálózatelemzés. Ezek részletes bemutatására ebben a tanulmányban nincs hely, de a kérdésről kiváló szakirodalmi összefoglalót tartalmaz Blondel et al. (2008) és Newman (2006).

Célszerű röviden áttekinteni néhány idősoros modellt használó nemzetközi empirikus kutatást. Azqueta-Gavaldón (2020) az euróövezet monetáris instabilitásait vizsgálta szövegbányászati változók felhasználásával. Módszere az idősoros szegmentáció, és az online szöveganyag VAR- *(Vector Autoregression)* analízisre épülő kulcsszóbányászatának ötvözése volt gépi tanulásos módszer felhasználásával. Miután megállapították a rezsimek időtartamát, megnézték a releváns kulcsszavakat az adott időszakban. Kimutatható volt olyan szófelhő, amely Németország, Franciaország és Olaszország esetében specifikusan volt köthető a gazdasági idősorok rezsimváltásaihoz.

Baker et al. (2016) szintén VAR-analízisre támaszkodva dolgozott ki nemzetközi volatilitást előre jelző rendszert szövegbányászati változók alkalmazásával. A módszer igen hasonló Azqueta-Gavaldón kutatásához, azzal a különbséggel, hogy elsődlegesen nem gépi tanulásos, hanem humán ágensek által végzett előzetes szelekciót alkalmaztak. A kettő közötti különbség az, hogy humán ágensek esetében a megkérdezett egyének szelektálják előre az egyes témákat, illetve klaszterezik a kulcsszavakat, míg gépi tanulásos esetben ugyanezt a feladatot vektoranalitikus eljárással különböző algoritmusok végzik el.

Li et al. (2021) szintén szövegbányászati módszerekkel bővített monetáris válság-előrejelző indexeket a kínai gazdaságra kalibrálva. Az idősoros szegmentáció módszereivel különítettek el rezsimeket, majd a rezsimek időpontjában szöveggyakoriság-vizsgálatot végeztek. A kulcsszavas vizsgálatok alapján új topikokat azonosítottak, majd a már meglévő monetáris válságjelző indexet specifikálták az eredményekkel. A kapott eredményeket GARCH *(Generalized Autoregressive Conditional Heteroskedasticity)* forecast-modellel újra lefuttatták. Igazolták, hogy a szövegbányászattal specifikált index sokkal pontosabban jelezte előre a makrogazdasági válságmutatók romlását, mint a szövegbányászati vizsgálat nélküli GARCH forecast-modell.

Ami a magyar vonatkozásokat illeti, Ágoston (2022) nemzetközi cégek csődelőrejelző módszereit vizsgálta szövegbányászati modellek felhasználásával. Hornyák (2014) a területi innovációk kutatásában rejlő szövegbányászati lehetőségeket tekintette át. Kruzslicz et al. (2016) a klaszterelemzésen alapuló szövegbányászat lehetőségeit foglalja össze. Kovács (2017) a részvénytőzsdék előrejelzéseit vizsgálta szövegbányászati módszerek felhasználásával.

Az idősoros modellek alkalmazásával kapcsolatban azonban számos probléma merül fel, ha szövegbányászati módszerekkel kinyert változóval szeretnék specifikálni a modellt. Miként arról már korábban szó volt, a legfontosabb probléma az, hogy a szövegbányászati változó nem kontinuus adatsor, azaz nem áll rendelkezésre minden azonos időszakaszra információ. Ez még egyszerű kulcsszavas frekvencia esetén sem adott mindig, ugyanis a keresőmotorok sem tudnak mindig adatokat szolgáltatni ugyanazon időszakaszokra. Az egyik legszélesebb körben használt Google Trend keresőmotorja sem tud napi bontásban szókeresési információkat adni. Ezt a problémát sajnos nem oldják meg a MIDAS *(Mixed Data Sampling)* vagy HAR *(Heterogeneous Autoregressive)* típusú idősoros modellek sem, ugyanis idősoros modellezés esetében még eltérő frekvenciájú változónál is követelmény a kontinuitás, ami csak jelentős transzformációs eljárásokkal becsülhető. Ez azonban számottevően torzítja a szöveges változó eloszlását a modellben.

Ennek a hiányosságnak az alternatívájaként jelentek meg a modularitásfunkció elvén működő különböző metaanalitikus eljárások, amelyek elsősorban a hálózatelemzés és a gépi tanulásos módszerek előretörésének voltak következményei. Ezek rendszerint többlépcsős algoritmusok, amelyek arra épülnek, hogy háromféle adatbázist szegmentálnak modulációs elvekkel. Ezek a tanító, a teszt-, illetve az OOS-adatbázisok. A modularitásos szegmentáció hosszú vagy rövid távú, mechanikus vagy smart módon működik. A módszer lényege, hogy gráfelméleti alapokon mű-

ködő modulképzésekkel folyamatosan szegmentálják az adatbázist, egymásra vonatkoztatják a tanító, a teszt- és a forecast OOS-adatbázisokat. Ezt vagy úgy teszik, hogy mindig a kiinduló adatbázishoz viszonyítanak, vagy pedig mindig a legutolsó szegmentációs állapotot veszik figyelembe. A módszer gyengéje, hogy az előrejelzést mindig az OOS-adatbázis szegmentációja határozza meg, így a prognózis ereje és pontossága mindig attól függ, hogy az OOS-modul miként kerül meghatározásra (Kock, 2009).

Ilyen típusú vizsgálatot közöl Cicea & Marinescuo (2021). Publikációmetrikai analízis segítségével vizsgálták a gazdasági kibocsátás és a külfölditőke-befektetések kapcsolatát. Modulációs 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, és pontosabb előrejelzési lehetőségekkel rendelkeznek, mint a kemény gazdasági modellekkel bemutatott idősoros előrejelzések. Tanulmányuk további kielégítő bevezetést nyújt a modulációs elven működő gazdasági topikelemzés szakirodalmába, így ennek részletes felsorolásától itt eltekintek.

Végül érdemes áttekinteni a panelregresszión alapuló modelleket, ugyanis véleményem szerint szövegbányászati modellspecifikációra ez a forma a leginkább alkalmas. Baltagi (2007) módszeresen összegyűjtötte a panelmodellek piaci elemzésben, pénzügyi modellezésben és makrogazdasági összehasonlító elemzésben való általános használhatóságát. Bemutatásában leginkább olyan esetekkel lehet találkozni, ahol a vizsgált terület túlságosan komplex az idősoros modellek alkalmazásához. Emellett számos olyan változót tartalmaz, amely idősorosan nem megfelelően vizsgálható.

A panelregresszió használható regionális gazdasági kibocsátások vizsgálata, illetve nagyobb ipari vagy gazdasági területek összehasonlító elemzése esetén. Baltagi és munkatársai (Baltagi et al., 2008) francia régiók összehasonlításával térképezte fel a benzinárak és a regionális keresetek kapcsolatát. Driver et al. (2004) a gazdasági bizonytalanság hatását mérte fel nagy-britanniai iparágazatok összehasonlításával. Rapach & Wohar (2002) az USA monetáris politikájának hatásait kutatta a befektetői hajlandóság és a cégprofilok vonatkozásában, szintén panelregressziós módszerekkel. Ezeknek az elemzéseknek közös eleme, hogy olyan kevert (keresztmetszeti és idősoros adatsorokat egyaránt tartalmazó) adatokkal dolgoztak, amelyekre jellemző a diszkontinuitás, azaz bizonyos régióra vagy iparágra egy bizonyos időpontban állnak rendelkezésre adatok, de egy másikban nem. A különböző típusú modellek fontosabb jellemzőit az *1. táblázat* foglalja össze.

1. táblázat

A szövegbányászati specifikációk összefoglalása

Modelltípus	Cél	Hatásfok	Alkalmazási környezet
Idősoros	modellspecifikáció	globális érvényesség, magas torzítás	makrogazdaság, pénzügy
Panel	modellspecifikáció	lokális érvényesség, kisebb torzítás	tacit makrogazdasági/piaci környezet
Modularitásos szegmentáció	rejtett összefüggések feltárása	globális/lokális érvényesség, bizonytalanabb előre jelző erő	tacit makrogazdasági/piaci környezet
Metaanalitikus	rejtett összefüggések feltárása	globális/lokális érvényesség, bizonytalanabb előre jelző erő	tacit makrogazdasági/piaci környezet

Forrás: Saját szerkesztés az áttekintett modelleket bemutató szakirodalmi források alapján.

Kulcsszavas panelregressziós vizsgálat magyarországi adatokon

A tanulmány főbb kérdésfeltevéseinek áttekintése után egy egyszerű modellezés következik 2018 és 2021 közötti hazai adatok alapján. Bár az előzőekben Khan & Upadhayaya (2020) alapján utaltam arra, hogy az ár- és különböző befektetői kondíciókat mérő indexek idősoros előre jelző ereje vitatható, ettől függetlenül az egyszerűbb áttekinthetőség kedvéért makroökonómiai válságszignálokként négy árindexet (fogyasztói, ipari, import és export) és a magflációt választottam. Szöveges magyarázó változó a „krízis", a „gazdasági válság", a „recesszió" és a „gazdasági recesszió". Mindegyik magyarázó változót 3 lag késletetett hosszúsággal vizsgáltam, ami jelen frekvenciában 3 hónapos időablakot jelent. Arra a kérdésre kerestem a választ, hogy az első lag késleltetettek melyik modellben mutatnak jobb p-value értékeket, valamint hogy melyik modell hibatesztjei bizonyulnak jobbnak. Benchmarkmodellként egy egyszerű ARMA *(Autoregressive Moving Average)* idősoros modellt használtam, amelyben nem szerepeltettem szövegbányászati magyarázó változókat, a másik választott modell random hatású panelregresszió, amelyben szövegbányászati változókat alkalmaztam. Feltételezésem szerint az utóbbi modell sokkal pontosabban fogja a válságindikátorokat előre jelezni, mint a puszta makroökonómiai mutatókat tartalmazó idősoros modell.

A makroökonómiai mutatók kapcsán feltételeztem, hogy az őket ért pénzpiaci sokkok kiegyenlítettek, valamint azt, hogy a választott mutatók homogenizáltak annyira, hogy ne kelljen a különböző külső sokkhatások befolyásoló hatásával külön foglalkozni. Természetesen lehet úgy is dönteni, hogy a különböző sokkhatásokat figyelembe vesszük, de jelen tanulmányban a szövegbányászati specifikáció működésének a bemutatása a fő cél, nem pedig egy széles körű sokkhatáselemzés a válságszignálokról. Ez érinti mind a maginfláció, mind pedig az árindexek kérdését.

A szöveges magyarázó változók kiválasztása az egyszerűség kedvéért a hétköznapi gyakorlat figyelembevételével történt. Természetesen érzékenyebb módszereket is lehet alkalmazni a kulcsszavak kiválasztására, de feltételezésem szerint a jelen vizsgálatban erre nincs szükség. Végül feltételeztem, hogy a kulcsszavak frekvenciájának vizsgálatára széles körben alkalmazott Google Trend szókereső valós információkat szolgáltat.

2. táblázat

A benchmark ARMA-modell szövegbányászati specifikáció nélkül

ARMA				
Függő változó: d_core_inf_rate				
Magyarázó változó	Koefficiens	SD	p-value	
d_fogy_arindex_1	−0,642298	0,1802	0,0004	***
d_import_arindex_1	0,36429	0,0654	2,58E-08	***
d_export_arindex_1	−0,162669	0,0695	0,0193	**
R^2	0,7			
Mean Error −0,012225				
Root Mean Squared Error 0,14314				
Mean Absolute Error 0,10491				

Forrás: KSH-adatbázis.

Első körben a benchmark ARMA-modellel készült analízis bemutatására kerül sor. Ebben nem szerepeltettem a szövegbányászati változókat. A függő változó a maginfláció, a magyarázó változók pedig az árindexek. A *2. táblázat* az ARMA-modell output adatait tartalmazza. Láthatjuk, hogy az importárindex már az első lag késleltetettben pontosabban jelzi a válságot, mint a fogyasztói árindex (minél kisebb

p-value érték). A magyarázó változók első lag késleltetettjeivel p-value **<0,05, ***<0,001. A fogyasztói árindex és az exportárindex csökken, az importárindex növekszik. Ez megfelel a külgazdasági multiplikátorhatásnak. Az importárindex változásának p-value értéke a legjobb a magyarázó változók között. Ez is igazolja a külgazdasági multiplikátor jelenséget. .A modellilleszkedés jó, R^2 0,7. A modell hibatagjainak információira *(Mean Error, Root Mean Squared Error, Mean Absolute Error)* a modellszelekciónál még visszatérek.

A modell minden tekintetben jól teljesít (R^2 0,7, ami mindenképpen jónak mondható). A p-value esetében az import- és az exportárindex változása a legerősebb az első lag késleltetettben. Sokkal erősebb, mint a fogyasztói árindex változásának esetében. Nem a fogyasztói árindex az, amely a legjobban jelzi a recessziót, hanem az importfolyamatok változása. Az ARMA-modell igazolta a külkereskedelmi multiplikátor-hatás működését válság-előrejelzés esetében is.

Panelregressziót akkor célszerű választani szövegbányászati változók esetében, amikor a szöveges változó diszkontinuus változó, azaz nem áll rendelkezésre adat minden időegységre. Fix hatású panelregressziót nincs értelme ilyen helyzetekben választani, mert nem az a kérdés, hogy egy bizonyos szövegbányászati változó időben hogyan befolyásolja a többit, vagy fordítva, hanem az, hogy milyen szöveges változó és hogyan jelentkezik bizonyos makroökonómiai változó mellett (random hatás). Általános formában:

$$y_{it} = \beta_i X_{it} + (\alpha_i + \varepsilon_{it})$$

ahol y_{it} a függő változó (jelen esetben a maginfláció), $\beta_i X_{it}$ a regresszorok mátrixának becslése (ide tartoznak a makroökonómiai mutatók és a szövegbányászati változók), α_i a random hatás szórásának hatása a mátrixon, ε_{it} pedig a hibatag. A panelregressziós vizsgálatnál Nerlove-transzformációt választottam, a paneleket pedig abban az arányban osztottam fel, hogy idősoros szempontból a legtöbb megfigyelési pont legyen. A becslési eljárás GLS *(Generalised Last Squared)* -becslés volt. Az eredményeket a *3. táblázat* tartalmazza.

3. táblázat

Random panelregresszió a magyarázó változók első, második és harmadik lag késleltetettjeivel, Nerlove-transzformáció alkalmazásával, két keresztmetszeti panelosztással

Random-effekt (GLS)				
Függő változó: dcoreinfrate				
Robust (HAC) standard error				
Magyarázó változó	Koefficiens	SD	p-value	
const	0,165977	0,0209267	2,17E-15	***
krizis_1	0,016862	0,00607162	0,0055	***
krizis_2	−0,00966980	0,0223113	0,6647	
krizis_3	0,00698104	9,06E-05	0	***
gazdasAgivAlsAg_1	−0,0518110	0,0250211	0,0384	**
gazdasAgivAlsAg_2	0,0629292	0,00856074	1,97E-13	***
gazdasAgivAlsAg_3	−0,0402898	0,0569182	0,479	
recesszia_1	0,0147477	0,00783589	0,0598	*
recesszia_2	−0,0312579	0,0365662	0,3926	
recesszia_3	−0,0364149	0,00194494	3,22E-78	***
gazdasAgireces~_1	−0,0169732	0,0128781	0,1875	
gazdasAgireces~_2	0,022143	0,0845206	0,7933	
gazdasAgireces~_3	0,0452321	0,0326037	0,1653	
dfogyarinde_1	−0,240039	0,138205	0,0824	*
dfogyarinde_2	−0,0522059	0,0586649	0,3735	
dfogyarinde_3	−0,383091	0,209191	0,0671	*
dipariarind_1	−0,000581390	0,0032351	0,8574	
dipariarind_2	−0,0571532	0,0452576	0,2066	
dipariarind_3	0,024051	0,0314354	0,4442	
dimportarind_1	0,246154	0,0176407	2,98E-44	***
dimportarind_2	0,266618	0,0428705	5,00E-10	***
dimportarind_3	0,304498	0,212028	0,151	

Random-effekt (GLS)				
Függő változó: dcoreinfrate				
Robust (HAC) standard error				
dexportarind_1	−0,128272	0,0429605	0,0028	***
dexportarind_2	−0,166514	0,0376031	9,50E-06	***
dexportarind_3	−0,358887	0,138875	0,0098	***
Between' variance	0,308251			
Within' variance	0,0109618			
Mean Error	–0,0043551			
Root Mean Squared Error	0,14002			
Mean Absolute Error	0,10674			

Megjegyzések: Google Trend. p-value **<0,05, ***<0,001. A constant (con) p-value értéke jó, ami azt jelenti, a modell jól illeszkedik a regressziós egyenesre.

Forrás: KSH-adatbázis.

A random panelregresszió nemcsak kedvezőbb statisztikai tulajdonságokkal rendelkezik, mint az ARMA-modell, hanem a külkereskedelmi multiplikátor-hatásnak is megfelel. A függő változó szóródása nem a legjobb a középértékhez képest, a p-value értékek azonban mindenhol jobbak, illetve a keresztmetszeti paneleken belüli *(Within Variance)* és a panelek közötti *(Between Variance)* variancia (főleg az előző) igen jók. Az alacsony variancia a keresztmetszeti panelen belül azt jelenti, hogy alacsony hibatagok alacsony korrelációban vannak egymással a keresztmetszeti panelen belül. A három hibaérték összehasonlításánál az tapasztalható, hogy a random panelmodell egyedül a MAE *(Mean Absolute Error)* alapján teljesít rosszabbul, az ME *(Mean Error)* és RMSE *(Root Mean Squared Error)* vonatkozásában jobban. Ezek a hibatesztek arról informálnak, hogy milyen a hibatag hatása a modellre, azaz alapvetően meghatározzák mind a modellválasztást, mind a modellspecifikációt. Vagyis a hibatagok alapján a szövegbányászati változóval specifikált panelmodell megfelelőbb, mint a szövegbányászati változó nélküli ARMA.

Összefoglalás, következtetések

Az elmúlt másfél évtizedben a gazdasági szövegbányászat jelentősége megnőtt az előre jelző modellek körében a modellspecifikáció terén, ugyanis a szövegbányászat számos rejtett gazdasági tényező modellezésére alkalmas. A tanulmány bemutatta a gazdasági szövegbányászat gazdasági előre jelző modellek finomításában betöltött jelentőségét, továbbá a legfontosabb szövegbányászati algoritmusok működését. A külgazdasági multiplikátor-hatás figyelembevételével két regressziós modellt állítottam fel a hazai maginfláció elemzésére: ARMA- és random panelregressziót vizsgáltam, ez utóbbit szövegbányászati változókkal specifikáltam, a benchmark ARMA-modellt nem.

Az elemzéssel arra a kérdésre kerestem a választ, hogy igazolható-e a fogyasztói árindex mint kiemelt válságindikátor létjogosultsága, illetve az idősoros vagy a szövegbányászati szempontból specifikált panelregresszió rendelkezik jobb ökonometriai jellemzőkkel. Eredményül azt kaptam, hogy a háromhibatag értéket tekintve (ME, MAE, RMSE) a szövegbányászatilag specifikált random panelregresszió hatékonyabb az összevetésül választott ARMA-benchmark modellhez képest. A panelmodell azt is igazolta, hogy a külgazdasági mutatók 1 lag késleltetettjei jóval relevánsabbak a válság-előrejelzés szempontjából, mint a fogyasztói árindex 1 lag késleltetettje. Ez mind az előzőekben bemutatott panelregresszióval kapcsolatos összefoglaló szakirodalmat (Baltagi, 2008; Driver, 2004; Rapach & Wohar, 2002), mind a kutatás fő kérdéseit pozitív módon támasztja alá, illetve megerősíti.

Gyakorlati következtetés, hogy szövegbányászati modellspecifikáció esetén sokkal hatékonyabb a panelregressziók használata, mint az idősoros modellek alkalmazása.

A tanulmányban bemutatott kutatási módszertan *tudományosan újszerű következtetése,* hogy igazolja a panelregressziós modellek alkalmazásának létjogosultságát, ami a hazai szakirodalomban eddig háttérbe szorult mind a gazdaságiválság-előrejelzés, mind a gazdasági szövegbányászat területén, a domináns idősoros modellezéssel szemben.

További *kutatási irány* a panelregresszióból nyert eredmények hibatagjainak vizsgálata OOS-mintán. Erre mindenképpen azért van szükség, hogy láthassuk a paneladatokon nyert eredmények mennyire általánosíthatók, ugyanis egyben ez a panelregressziós eljárás egyik legnagyobb *korlátja*. Másik fontos irány a kulcsszavak kiválasztásának statisztikailag megalapozottabb előkészítése. Vizsgálatomban

a leggyakrabban előforduló kulcsszavak frekvenciáit választottam specifikációnak, azonban nincs kizárva, hogy finomabb szűrőkkel bővíthető a szófelhő.

Hivatkozások

Ágoston, N. (2022). Külföldi csődelőrejelző módszerek szisztematikus irodalomelemzése. *Vezetéstudomány / Budapest Management Review, 53*(1), 69–89. https://doi.org/10.14267/veztud.2022.01.06

Azqueta-Gavaldón, A. (2020). *Text-Mining in Macroeconomics: the Wealth of Words* (Doctoral dissertation, University of Glasgow). http://theses.gla.ac.uk/81641/

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics, 131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Baltagi, B. H. (2008). Forecasting with panel data. *Journal of Forecasting,* 27(2), 153–173. https://doi.org/10.1002/for.1047

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008

Cicea, C., & Marinescu, C. (2020). Bibliometric analysis of foreign direct investment and economic growth relationship. A research agenda. *Journal of Business Economics and Management, 22*(2), 445–466. https://doi.org/10.3846/jbem.2020.14018

Driver, C., Imai, K., Temple, P., & Urga, G. (2004). The effect of uncertainty on UK investment authorisation: Homogenous vs. heterogeneous estimators. *Empirical Economics, 29*(1), 115–128. https://doi.org/10.1007/s00181-003-0192-2

Hamilton, J. D. (2016). Macroeconomic Regimes and Regime Shifts. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of Macroeconomics. Vol 2.* (pp. 163–201). Elsevier. https://doi.org/10.1016/bs.hesmac.2016.03.004

Harrod, R. F. (1939). An Essay in Dynamic Theory. *The Economic Journal, 49*(193), 14–33. http://piketty.pse.ens.fr/files/Harrod1939.pdf

Hornyák, M. (2014). Közösségi adatforrások felhasználási lehetőségei a területi kutatás támogatásában. *GRADUS,* 1(2), 230-237. http://real.mtak.hu/110845/1/2014_2_ECO_011_HORNYAK.pdf

Keogh, E., Chu, S., Hart, D., & Pazzani, M. (n.d.). *An online algorithm for segmenting time series.* Proceedings 2001 IEEE International Conference on Data Mining. https://doi.org/10.1109/icdm.2001.989531

Khan, H., & Upadhayaya, S. (2019). Does business confidence matter for investment? *Empirical Economics, 59*(4), 1633–1665. https://doi.org/10.1007/s00181-019-01694-5

Kock, A. (2009). *A guideline to meta-analysis.* Tim Work. Pap. Ser, 2, 1–39. https://webcache.googleusercontent.com/search?q=cache:IxlvepE4PPQJ:https://www.tim.tu-berlin.de/fileadmin/fg101/TIM_Working_Paper_Series/Volume_2/TIM_WPS_Kock_2009.pdf&cd=1&hl=hu&ct=clnk&gl=hu

Kovács, B. (2017). *Tőzsdei hírbányászat a magyar részvénypiacon.* (Doktori értekezés, Pécsi Tudományegyetem, KTK Gazdálkodástani Doktori Iskola) http://pea.lib.pte.hu/handle/pea/23362

Kruzslicz, F., Kovács, B., & Hornyák, M. (2016). Összehasonlító klaszterjellemzés külső, szöveges források bevonásával. *Statisztikai Szemle, 94*(11–12), 1123–1148. https://doi.org/10.20311/stat2016.11-12.hu1123

Li, X., Shang, W., & Wang, Sh. (2013). *Incorporation of Social Media Data into Macroeconomic Forecast Systems: A Mixed Frequency Modelling Approach*. PACIS 2013 Proceedings. 57. https://aisel.aisnet.org/pacis2013/57

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,* 9, 22171–22185. https://doi.org/10.1109/access.2021.3055967

McCombie, J., & Thirlwall, A. (1999). Growth in an international context. *Foundations of International Economics*. https://doi.org/10.4324/9780203017760.ch3

Nekarda, C. J., & Ramey, V. A. (2020). The Cyclical Behavior of the Price-Cost Markup. *Journal of Money, Credit and Banking, 52*(S2), 319–353. Portico. https://doi.org/10.1111/jmcb.12755

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review, E 74*(3). https://doi.org/10.1103/physreve.74.036104

Rapach, D. E., & Wohar, M. E. (2002). Testing the monetary model of exchange rate determination: new evidence from a century of data. *Journal of International Economics, 58*(2), 359–385. https://doi.org/10.1016/s0022-1996(01)00170-2

Samuelson, P. A., & Nordhaus, W. D. (2016). *Közgazdaságtan*. Akadémiai Kiadó. https://doi.org/10.1556/9789630597814

Sclove, S. L. (1983). Time-series segmentation: A model and a method. *Information Sciences, 29*(1), 7–25. https://doi.org/10.1016/0020-0255(83)90007-5

Wang, C., & Wang, X. S. (2000). *Supporting content-based searches on time series via approximation*. Proceedings. 12th International Conference on Scientific and Statistical Database Management. https://doi.org/10.1109/ssdm.2000.869779

Zarnowitz, V., & Moore, G. H. (1982). Sequential Signals of Recession and Recovery. *The Journal of Business, 55*(1), 57. https://doi.org/10.1086/296154