Alkalmazott statisztika? R!

Daróczi Gergely
PhD-jelölt, az Upshift R Kft. ügyvezetője
E-mail: daroczig@rapporter.net

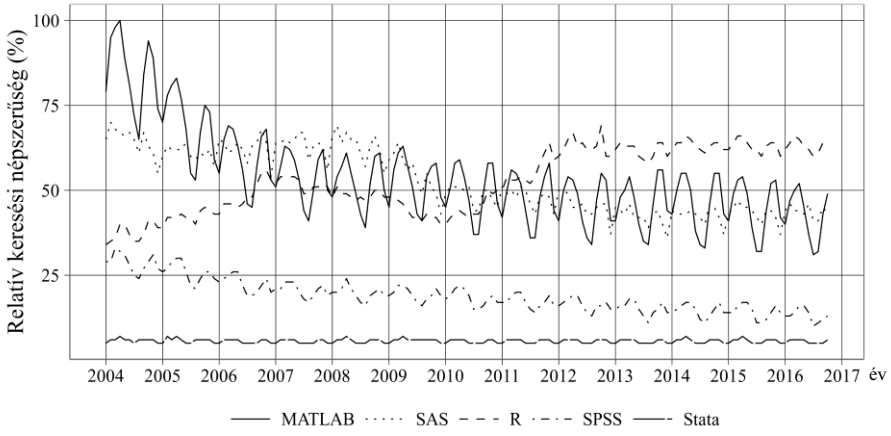
A tanulmány napjaink egyik legnépszerűbb adatelemző eszközének, az R programozási nyelv és statisztikai környezetnek főbb jellemzőit, a sikeréhez vezető út mérföldköveit veszi sorra, továbbá a programmal ismerkedni vágyók számára biztosít rövid felsorolást a legfontosabb R csomagokról, oktatási segédanyagokról, konferenciákról és egyéb rendezvényekről.

TÁRGYSZÓ:
R programozási nyelv.

DOI: 10.20311/stat2016.11-12.hu1108

Az R programozási nyelvvel és adatelemző, statisztikai és adatvizualizációs rendszerrel (*R Core Team* [2016]) kicsivel több mint tíz éve ismerkedtem meg felsőfokú tanulmányaim során, amikor is egy választható gazdaságszociológia kurzus keretén belül a magyarországi burgonyapiac kaotikus viselkedésével (*Vizvári–Bacsi* [1997], *Vizvári* [2002]) volt szerencsém rövidebben foglalkozni. Ezt a személyes emléket azért tartottam fontosnak leírni, mert a káoszelmélettel való ismerkedés in medias res – a kapcsolódó meglehetősen összetettnek tűnő matematikai háttér tárgyalása nélkül –, az alkalmazással indult, és az R-nek köszönhetően a félév végén sikerrel abszolváltam a kurzust. Ezzel párhuzamosan egy új és igen gazdag világ tárult fel előttem az R eszköztárával, amely évekkel később egyik legkedvesebb szabadidős elfoglaltságommá, majd elsődleges munkaeszközömmé vált.

1. ábra. Adatelemző szoftverek online keresettsége, népszerűsége



Noha az R nyelv már több mint húsz éves múltra tekinthet vissza, népszerűségét leginkább az elmúlt öt évben nyerte el, és korábban inkább csak a felsőoktatásban, valamint akadémiai pályán, statisztikusok között volt ismert. Ez jól látszik az 1. ábrán is, ahol a Google Trends adatai alapján öt adatelemzésre alkalmas szoftver online keresési népszerűségét láthatjuk 2004 óta.

Bár ezen adatok nem reprezentálják a szoftverek valódi népszerűségét, és a keresési adatokból nem tudunk a valódi használati adatokra következtetni, annyi mégis látszik, hogy az elmúlt közel tizenöt év alatt komoly változások történtek. A keresési toplista éléről a MATLAB igen gyorsan visszaesett, majd a SAS vette át a vezetést évekig, mígnem 2011 óta egyértelműen az R lett a „legkeresettebb" adatelemző esz-

köz a listában, és az egyéb feltüntetett szoftverekkel ellentétben a kapcsolódó keresések belüli aránya szinte töretlenül növekszik.[1]

Mindenesetre joggal vetődhet fel a kérdés, hogy mire fel ez a nagy népszerűség vagy keresettség?

1. Az R múltja és jelene

Tóth Gergellyel 2013-ban írt cikkünkben (*Daróczi–Tóth* [2013]) részletesebben tárgyaltuk az alkalmazott statisztika fejlődéséhez aktívan hozzájáruló hardver és szoftver eszközök XX. századi történetét, amelyben többek között kitértünk az első digitális számítógépek, a mainframe-ek, a személyi számítógépek és mobil eszközök; továbbá a BMD(P), a SPSS, a SAS, a JMP, a STATA és az S, S+, valamint az R nyelvek fontosabb mérföldköveire is. A továbbiakban ezekből az R története szempontjából legfontosabb momentumokat emelem ki, hogy megérthessük mi vezetett az R mai sikeréhez.

A Scheme nyelv mellett az S nyelv tekinthető az R előfutárának, hiszen az R valójában az S nyelv open-source (nyílt forráskódú) implementációjaként született az 1990-es évek elején. Az S nyelv azonban már jóval korábban, 40 évvel ezelőtt jött létre a C és C++ nyelveket is világra hozó Bell Laboratóriumban. Az S és R nyelvek szoros kapcsolatát jól jellemzi, hogy az S nyelven eredetileg dolgozó *John Chambers* ma az R Development Core Team (az R központi fejlesztőcsapata) tagja, továbbá *Rick Beckert* is örömmel számolt be az S fejlesztésével kapcsolatos tapasztalatairól a 2016-os R konferencián a Stanford Egyetemen (*Smith* [2016]).

A Bell Laboratórium belső hálózatában már az 1970-es évek második felétől használták az elsősorban John Chambers által fejlesztett S programcsomagot. Nagy előnye volt a korábbi, egyedi feladatokra írt Fortran programokkal szemben, hogy egységes parancsok segítették az interaktív adatelemzést, illetve a különböző statisztikai módszerek elvégzéséért felelős függvények (programrészek) könnyen elérhetők voltak a fejlesztők számára. Ezeken kívül a nagyszámítógépekre szánt General Comprehensive Operating Systemről UNIX-ra történő portolás (1980), majd a program (1981), illetve a programkód (1984) megnyitása a külvilág felé garantálhatta leszármazottjai hatalmas sikerét. Az 1980-as évek végére az immáron több mint tíz éves program többszöri átdolgozása után megjelent a „New S" nyelv, amely a korábbi makrók helyett már valódi függvényekre épített. Újabb grafikus eszközök (X11 és

[1] A teljesség kedvéért meg kell jegyezni, hogy a listában nem szerepel például az Excel, amellyel kapcsolatban nagyságrenddel több keresés indult, továbbá a keresési adatok éven belüli szezonalitásából jól látszik, hogy a keresések száma erősen összefügg a tanévek időbeosztásával, tehát a keresési adatok inkább jellemzik a (felső)oktatás résztvevőit, mint például az ipari szereplőket.

PostScript) váltak használhatóvá, kialakul a napjainkban is használt „formula-notation" (szabályjelölés) és az alapértelmezett S3, majd később az S4 metódusok (*Daróczi–Tóth* [2013]).

Bár az S nyelv és programozási környezet, továbbá annak a kereskedelmi implementációja, a TIBCO által szállított S PLUS elérhető, de napjainkban az S nyílt forrású leszármazottja, az R szoftver és variánsai jóval népszerűbbek és általánosabban használtak mind az akadémiai, mind az ipari és egyéb környezetekben. Többek között például a TIOBE-index (a programozási nyelvek népszerűségét számszerűsítő lista) szerint az R a leggyakrabban használt programozási nyelvek sorában is bekerült az első 30 közé (2012), 2015-ben a 12. volt a listán, míg az S kereskedelmi változata (S-PLUS) csak ritkán került be az első 100 közé.

Az R program fejlesztése az S eredményein túl (*Hornik* [2012]) *Gerald Jay Sussman* Scheme nyelvéből kölcsönzött „lexical scoping" (lexikális hatókör) funkciója köré épült, és 1993-ban indult az Auckland-i Egyetemen *Ross Ihaka* és *Robert Gentleman* vezetésével. Azóta az R szoftver központi magjának fejlesztését és karbantartását a jelenleg húsz főt számláló R Development Core Team végzi.

A program nyílt forráskódú: szabadon használható, terjeszthető és módosítható a GPL v2 licenc mellett. A Free Software Foundation által elismert program, a GNU (GNU's Not Unix – a GNU nem Unix) része. Számos platformon (Windows, Macintosh, Linux) ingyenesen elérhető a telepítésre kész változata, sőt, napjainkra sok grafikus felhasználói felület („frontend", „graphical user interface") segíti az R-t használók mindennapi munkáját a hagyományos parancssorok, megoldások és azok integrált környezetben (RStudio, Emacs/ESS, Eclipse/StatET, Architect, TextMate, Notepad++ stb.) való futtathatósága mellett (*Daróczi–Tóth* [2013]). Ezek közül napjainkban az RStudio a leggyakrabban használt és talán legdinamikusabban fejlődő IDE (integrated development environment – integrált fejlesztői környezet) az R felhasználók körében, amelynek létezik asztali (desktop) és szerver változata is. Ez utóbbi a számításokat egy távoli, általában nagy kapacitású számítógépen végzi, a kezelőfelület, a parancssor pedig bármely szabványos internetböngészőből elérhető a kliensoldalról, akár jelszóval védetten is.

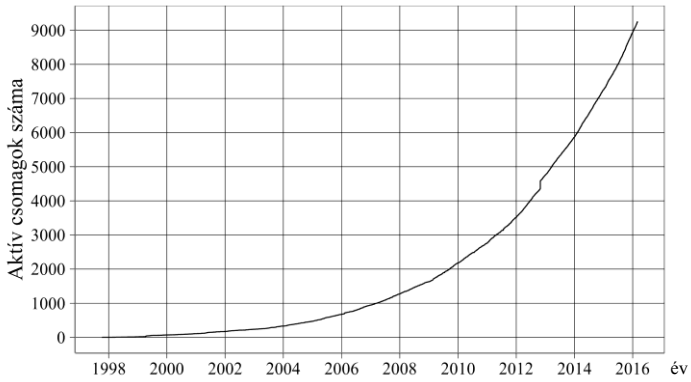
2. R csomagok

Bár az alap R telepítés is tartalmaz számos adatbeolvasásra és átalakításra alkalmas funkciót, valamint az általánosan használt statisztikai próbák és modellek körét, továbbá adatvizualizációra is alkalmas, az R egyre növekvő sikerét az ingyenes és szabadon használható volta mellett (vagy talán inkább az alapján) elsődlegesen a

CRAN (comprehensive R archive network – átfogó R archívum hálózat) csomagtárolónak és a felhasználók által megosztható és az R-hez hasonlóan ingyenesen, általában különösebb korlátozás nélkül megosztott programkódoknak köszönheti.

Napjainkban a CRAN több mint 9 000 R csomagot számlál, amelyek többnyire lefedik a kurrens statisztikai módszerek tárházát, és mára elmondható, hogy alig jelenik meg új statisztikai módszert bemutató tudományos folyóiratcikk a kapcsolódó R csomag publikálása nélkül. A növekedés az elmúlt közel húsz év adatai alapján exponenciálisnak mondható[2] – az első tíz évben összesen ezer, majd évről évre egyre több csomag került publikálásra. 2016-ban mindösszesen hat hónap telt el a 8 000. és a 9 000. aktívan karbantartott új csomag publikálása között.

2. ábra. A CRAN tárolókban megtalálható aktív csomagok száma



Bár a CRAN-re bárki beküldhet R csomagokat, és azokon kizárólag automatikus teszteket futtatnak a hálózat üzemeltetői, a nagy számú felhasználó és az aktív közösség (GitHub, StackOverflow, [R-help] és egyéb levelezőlisták több mint havi 3 000 üzenete stb.) állandó ellenőrzése és visszajelzése egyfajta garanciát jelent a programok karbantartására és további fejlesztésére. E mellett az R Core Development Team kezeskedik az alapcsomagok és néhány további library hibamentes működéséért, illetve mára a valóban standard munkaeszközzé vált R többek között klinikai vizsgálatok esetében is megfelelő felülvizsgálattal és tanúsítványokkal rendelkezik (*The R Foundation for Statistical Computing* [2012]).

A továbbiakban sorra veszem a leggyakrabban, legáltalánosabban használt és általam leginkább nagyra értékelt csomagok tematizált listáját, hogy az R-rel ismerkedni vágyókat segíteni tudjam az első lépésekben. Természetesen ez a lista egyáltalán nem teljes és valamelyest szubjektív válogatás, mégis úgy gondolom, hogy a több mint 9 000 csomag közötti választást nagymértékben tudja segíteni a kezdő felhasználók számára. A csomagok és az R függvények neveit dőlt betűvel jelölöm.

[2] Az adatok forrása: https://gist.github.com/daroczig/3cf06d6db4be2bbe3368

3. Adatmanipuláció

Mint általánosan ismert, az adatelemzéssel foglalkozó projektekre fordított idő 80 százaléka általában az adatok beolvasásával, tisztításával és előkészítésével telik, és mindösszesen a maradék 20 százalék fókuszál a tényleges adatelemzésre, vizualizációra és modellezésre, illetve az eredmények kiértékelésére. Ennek megfelelően R-ben is számos eszköz áll rendelkezésünkre az adatok beolvasására:

– a *read.csv* vagy általánosabb *read.table* függvény segítségével csv, txt vagy egyéb strukturált szöveges állományokat olvashatunk be, nagyobb állományok gyors beolvasásához pedig a külön telepített *readr* csomagot vagy a *data.table* csomag *fread* függvényét javasolt használni;

– a *foreign* csomag lehetőséget biztosít a dbf fájlformátumon túl az általánosan használt egyéb adatelemző programok egyedi fájlformátumait olvasni, mint például az SPSS sav vagy a SAS által használt ssd és xport, továbbá az újabb *haven* csomag is számos kereskedelmi program formátumát támogatja, úgy mint a SAS sas7bdat és sas7bcat, az SPSS sav és por vagy a Stata dta fájljait;

– Excel fájlok legegyszerűbben a *readxl* csomag segítségével olvashatók, amely nagy előnye, hogy sem Java, sem Perl függőségekkel nem rendelkezik, így egyszerűen telepíthető, de a fejlettebb funkciók eléréséhez sokszor megkerülhetetlen valamelyik Java alapú csomag, például a *openxlsx* használata, amely formázott Excel táblák írására is alkalmas,

– a *DBI* csomag egy általános adatbázisréteget biztosít a felhasználók számára, amely segítségével kényelmesen tudunk kapcsolódni például MySQL (*RMySQL*), Postgres és Amazon Redshift (*RPostgreSQL*), Oracle (*ROracle*) MS SQL Server (*RSQLServer*) vagy egyéb adatbázisokhoz ODBC (*RODBC*) vagy JDBC (*RJDBC*) kommunikációs csatornákon keresztül.

Az adatbeolvasáson túl általában szükségünk lesz az adatok ellenőrzésére és tisztítására is. A holland statisztikai hivatal munkatársai által fémjelzett *validate* csomag egyszerűen meghatározható szabályrendszerek alkalmazására nyújt lehetőséget beolvasott adatainkon, amely szabályok automatizált felderítésére is lehetőségünk nyílik a *deductive* csomag segítségével.

A tisztított adatok elemzése előtti adattranszformációk (például adatok rendezése, szűrése, aggregálása, származtatott változók létrehozása, adathiány kezelése, imputálás, táblák összekapcsolása stb.) során már az alap R telepítés is számos hasz-

nos függvényt biztosít, például egy mátrix sorain vagy oszlopain végzett műveletek elvégzésére az *apply* függvénycsalád segítségével vagy a *rowSums*, *colSums* stb. vektorizált megoldásokkal, de a hatékonyabb és felhasználóbarát szintaxis érdekében érdemesebb egyéb csomagokkal is megismerkednünk:

– a *data.table* csomag az R-ben általánosan használt *data.frame* adattömb oszlopműveletekre optimalizált kiterjesztése, amely az alap R telepítéshez képest sokkal gyorsabb és hatékonyabb működést tesz lehetővé nagyobb adatbázisokkal való munka során egy SQL-szerű szintaxis segítségével, és különböző táblák komplex összekapcsolásaira (például overlap, rolling vagy non-eque join) is lehetőséget biztosít;

– napjainkban egyre népszerűbb a *magrittr* csomag által biztosított „pipe" (cső, csővezeték) operátorra épülő, *Hadley Wickham*[3] által karbantartott *dplyr* csomag, amely egyrészt a *data.table*-hez hasonló gyors működést, másrészt olvashatóbb kódot tesz lehetővé az R függvények egymásba ágyazása helyett, a UNIX-ban megszokott pipe operátor segítségével a függvények kimenetét újabb függvényeknek adja át sorról sorra,

– a *reshape2* és *tidyr* csomagok az ún. long (hosszú) és wide (széles) táblák átalakítására nyújtanak lehetőséget a *dcast* és *melt* függvények segítségével, amely többek között a modellezés vagy adatvizualizációhoz szükséges adatformátum előállítása során bizonyul hasznosnak.

Bár R-ben általában adattömbökkel dolgozunk, de olykor elkerülhetetlen az ettől összetettebb adatstruktúrák kezelése. Ilyen esetekben a listák kényelmesen használhatók, azonban az adattömböknél megismert módszereket sajnos nem alkalmazhatjuk. Hasonló metódusokat kínál többek között az *rlist* csomag, amellyel listákon belül szűrhetünk, rendezhetünk vagy alakíthatjuk át, illetve lehetőség szerint transzformálhatjuk adattömb-formátumúra a további elemzésekhez adatainkat.

4. Adatvizualizáció

Az exploratív adatelemzés elengedhetetlen eszköze az adatok vizuális ábrázolása, amely segítséget nyújt az adatban rejlő minták felismerésében. Például konfirmatív

[3] Hadley Wickham számos további, méltán népszerű R csomag (többek között például a *ggplot2*, *readr*, *stringr*, *lubridate*, *tidyr*, *rvest*) eredeti szerzője és karbantartója, e csomagokat összefoglaló névvel csak egyszerűen *hadleyverse* vagy *tidyverse* névvel illetik általában.

faktorelemzés során is igen hasznos a feltárt eredmények képszerű megjelenítése, így az adattal foglalkozó szakemberek elengedhetetlen eszköztárába tartozik az adatvizualizáció is, amelyre már a kezdetek óta igen jól használható módszerekkel és függvényekkel rendelkezik az R.

Már az alaptelepítés is tartalmazza az általánosan használt főbb grafikai megoldásokat a *plot* függvénycsalád személyében, mint például az oszlop- vagy pontdiagram, de ugyanúgy elérhető hisztogram (*hist*) vagy éppen kördiagram (*pie*) is.

Bár már ezekkel a függvényekkel is egyszerűen tudunk egyszerre akár több változót is ábrázolni R-ben, a *lattice* csomag a formula jelölésnek köszönhetően a *grid* rendszerre támaszkodva, kiterjeszthető és testreszabható módon képes sokváltozós vizualizációk megjelenítésére, többek között a *barchart, xyplot, histogram, densityplot, contourplot* vagy például a kördiagram helyett javasolt *dotplot* segítségével. A *lattice* csomaggal készített ábrák általában kedvezőbb esztétikai megítélésben részesülnek mint az alap R függvényekkel generált grafikonok.

A legtöbbször használt grafikai csomag mégis a *ggplot2*, amely *Wilkinson* [1999] Grammar of Graphics (a grafikonok elemei és nyelvtana) R-beli implementációja. Sikerét egyrészt Wilkinson elképzelése megvalósításának, így az egységes interfésznek és az egyszerű, gyors fejlesztést lehetővé tevő használatnak, továbbá a szép megjelenésű grafikonoknak köszönheti. A csomag segítségével a kezdő R felhasználók is publikálásra kész ábrákat tudnak készíteni akár csak néhány könnyen megjegyezhető és meglehetősen intuitív parancs használatával.

Az itt bemutatott, statikus ábrák készítésére alkalmas csomagokon túl napjainkban egyre több olyan alkalmazással is találkozhatunk, amelyek segítségével egyszerűen készíthetünk interaktív grafikonokat – sőt, azok összekapcsolásával akár dashboardokat (irányítópultokat) is. Az általában D3.js (*Bostock–Ogievetsky–Heer* [2011]) alapú vizualizációs technikákat R-en belül a legegyszerűbben a *htmlwidgets* csomag segítségével hívhatjuk meg, amely R csomag automatizálja az adatok Javascript számára való átadását, és általában az interaktív vizualizáció mindösszesen egy-két parancs meghívásával, pusztán R-en belül maradva, tehát bármiféle további Javascript, HTML vagy egyéb szoftverfejlesztési ismeret nélkül is megvalósítható.

Az általános grafikonok készítésére alkalmas csomagokon túl találunk speciális igényeket kielégítő R csomagokat is, például a hálózatelemzés (*igraph*), térbeli statisztika és GIS (*maps, maptools, rgeos, sp, choroplethr*), idősorelemzés (*zoo, xts, forecast*) területén, vagy akár folyamatábrák (*DiagrammeR*) is egyszerűen rajzolhatók R-ben.

Bármely említett adatvizualizációs modul (és lényegében bármely R output) relatíve egyszerűen, párszoros R kóddal interaktív webalkalmazásban, dashboardban is felhasználható a *shiny* csomag és keretrendszer segítségével, amelyet futtathatunk akár saját gépünkön, akár szerver oldali környezetben is.

5. Statisztikai modellek, gépi tanulás

Az R leglényegesebb funkcionalitása azonban talán továbbra is a statisztikai módszerek nagyon gazdag tárháza, és teljes bizonyossággal kijelenthető, hogy nincs olyan programozási nyelv vagy adatelemző környezet, ahol több statisztikai modell vagy gépi tanulás (machine learning) eljárás lenne elérhető, mint az R-ben – legyen az akár egy hagyományos OLS- (ordinary least squares – legkisebb négyzetek módszere) regresszió, kevert modell vagy bayesi megközelítés.

A lineáris modellek már az alap R telepítésben elérhetők az egységes *lm* parancs segítségével, amely lineáris regressziós modelleken túl például varianciaanalízisre is lehetőség nyújt. A program a megadott változók mérési szintje alapján dönt az alkalmazott módszerről, de természetesen a modellek részletesen is paraméterezhetők. A *glm* parancs segítségével általánosított lineáris modellek építhetők, például Poisson-eloszlás alapján vagy többek között logisztikus regresszió segítségével.

A felsőoktatásban hagyományosan bemutatott többdimenziós eljárások közül a hierarchikus klaszterelemzés (*hclust*), a *k*-közép eljárás (*kmeans*), a többdimenziós skálázás (*cmdscale*), a főkomponens-elemzés (*prcomp* és *princomp*) és a faktoranalízis (*factanal*) az alap R telepítés részei, amelyeket egyszerűen kiegészíthetünk akár új rotálási eljárásokkal (*FactoMineR*) vagy például új klaszterező algoritmusokkal (többek között *fpc*, illetve *dbscan*).

Az R egyik fő erénye azonban abban rejlik, hogy nemcsak natív megoldásokat kínál, hanem egyéb programnyelvek – például C, C++ vagy a big data-val terhelt napjainkban egyre inkább a Java – moduljai is jól beágyazhatók a környezetbe, sőt, azokhoz natív interfészen keresztül biztosít hozzáférést. Erre jó példa a *h2o* csomag, amely valójában egy különálló, elosztott rendszereken való futtatásra tervezett, gépi tanulási módszereket igen hatékonyan megvalósító Java programcsomag. A H2O programot R parancsok segítségével is indíthatjuk, az adatok átadása a két környezet között automatikusan megtörténik, és a felhasználó számára észrevétlen módon nyílik lehetőségünk ezt a jól skálázható és meggyőző eredményeket produkáló machine learning eszközt párosoros R szkriptben használni. Hasonló módon érhető el az *xgboost* framework is.

Mindezek alapján ma már véleményem szerint kevéssé érdemes natív R függvényeket és (például a *randomForest*, *gbm*, *C50*, *rpart* vagy *party*) csomagokat használni gépi tanuló algoritmusok futtatására, hiszen a külső modulok segítségével jóval hatékonyabb módon nyílik erre lehetőségünk, de ha mégis erre adnánk fejünket, akkor a *caret* csomag egységes interfészt biztosít a különböző csomagokban elérhető függvények futtatásához.

6. Megismételhetőség

Napjainkban a statisztikai programokról szóló online társalgások központi témáját adják a „megismételhető kutatás” (reproducible research), ún. „annotált” jelentések készítése (literate programming) az R segítségével. Ennek az eljárásnak a lényege, hogy az elemzés folyószövegébe „csempészett” R kódot (ún. „chunk”-ok tartalmát) feldolgozva a kész anyag a szöveg között az eredményeket tartalmazza, ráadásul a szerző által meghatározott formátumban. Így nem szükséges többé táblázatkezelő eszközökben finomítani a statisztikai programok kimenetét, hogy azt majd egy szövegszerkesztőbe átmásolva tudjuk végleges formába önteni, hiszen mindezt megtehetjük egy lépésben is – az adatokra és nem a segédeszközökre koncentrálva.

R-ben minderre már az ezredforduló óta is lehetőségünk nyílt a *Sweave* segítségével PDF kimenetet generálva, azonban napjainkban már jóval népszerűbbek az egyéb fájlformátumokat is támogató, markdown-alapú eszközök. A markdown szöveges fájlformátum egyszerű szimbólumok segítségével jelöli a szöveg formázását (mint például a cím, alcím, táblázatok, beágyazott képek, hiperhivatkozások, referenciák, félkövér és kurzív betűk), amely többek között a *pandoc* program segítségével HTML-, PDF-, DOC- és egyéb dokumentumformátumokba konvertálható. Ez a funkcionalitás R-en belül egyszerűen elérhető a *knitr*, *rmarkdown*, továbbá például a *pander* csomagok használatával.

A sima szöveges (markdown) formátum nagy előnyei közé tartozik az egyszerűség mellett például a Word-dokumentumok kézzel való szerkesztésével szemben, hogy

– a kimeneti formátum bármikor megváltoztatható (legyen szükség akár Power Point prezentációra vagy egy HTML blog-posztra az elemzés végén);

– ha változik a bemeneti adat vagy elképzelésünk a kimenettel kapcsolatban, például a grafikonokat szeretnénk újragenerálni más színpalettát használva, mindezt egy gombnyomással megtehetjük;

– az eredményeket generáló programkód a grafikonok, táblázatok mellett marad, az utólag bármikor ellenőrizhető, felülvizsgálható és újrafuttatható;

– a teljes dokumentum és programkód egyszerűen tárolható verziókövető rendszerekben (például git vagy subversion) és a különböző változatok bármikor elővehetők.

Ennek egy példája a jelen dokumentumot generáló R markdown fájl, amely elérhető a http://bit.ly/statszemle-2016-R címen.

7. Egyéb eszközök

A korábbiakban röviden bemutatott témákon túl számos egyéb, különböző tudomány- vagy üzleti területen is jól használható R csomag született. Ennek részletesebb és még válogatottan is igen hosszú felsorolására terjedelmi okokból nem térek ki, de az érdeklődők számára ajánlom a CRAN tematikus és annotált csomaglistáját a https://cran.r-project.org/web/views címen. Itt, többek között klinikai vizsgálatokkal, farmakokinetikával, pénzüggyel, szövegbányászattal, térstatisztikával, társadalomtudománnyal, idősor-elemzéssel, illetve további kéttucat témával kapcsolatos oldalt tartanak karban a kapcsolódó kutatási kérdéseket, problémákat és az R csomagokat is jól ismerő szakértők.

Ezen listák közül ízelítőképpen a webes technológiákkal kapcsolatos gyűjteményt emelném ki. Itt találhatunk fejlesztőknek szánt csomagokat különböző webes szolgáltatásokkal való kommunikációs interfészek és R kliensek fejlesztésére, online tartalmak feldolgozására (például hogyan tudjuk egy weboldalon közzétett táblázat adatait különösebb manuális beavatkozás nélkül beolvasni), e-mail vagy IM üzenetküldésre alkalmas csomagokat, felhőszolgáltatásokhoz (mint például az Amazon, Google, Microsoft Azure vagy éppen a Facebook) tervezett klienseket, de akár interaktív, Javascript alapú online dashboardok, teljes honlapok és infografikák készítésére alkalmas R csomagokat, sőt R alapú webalkalmazásokat is.

8. R közösség

A rengeteg hasznos csomagon és algoritmuson túl az R népszerűségét és sikerét, véleményem szerint, a mögötte meghúzódó, pontosabban aktívan közreműködő közösségnek köszönheti. A Revolution Analytics 2014-es becslése szerint a világon mintegy 2 millió R felhasználó található, és számuk egyre növekszik. Sajnálatos, hogy a becslés kapcsán nem került közlésre az alkalmazott módszertan, annak megbízhatósága és a várható hiba nagysága, azonban egyéb metrikák alapján is jól látszik az R felhasználók magas létszáma:

– az [R-help] levelezőlistán 1997 és 2015 között közel 400 ezer üzenet született több mint 100 ezer különböző e-mail címről küldve (*Daróczi* [2015] 333. old.);

– 2016 közepéig majdnem 150 ezer kérdés adódott R témában a programozókörökben jól ismert StackOverflow Q&A fórumon, amelyek meghatározó részét igen gyorsan (pár órán belül) megválaszolták;

– a Facebook adatai alapján körülbelül 1,3 millió felhasználójukat érdekli az R nyelv (*Daróczi* [2015] 343. old.);

– az R fejlesztői és felhasználói tábor különösen aktív Twitteren: a Budapesten megrendezett 2016. szeptember 3-i satRday konferenciával kapcsolatban több mint 400 bejegyzés született 48 órán belül, de világszerte naponta több ezer tweetet publikálnak az #rstats jelöléssel rendezvényektől függetlenül is;

– a github.info adatai alapján 2014 végéig közel 35 ezer aktív R kódtárolót hoztak létre a GitHubon, amely alapján az R a 12. legnépszerűbb programozási nyelv a nyílt forráskódú programoknak otthont adó, online verziókezelő-rendszerben.

Az R felhasználók fentebb említett informális kommunikációs csatornáin túl természetesen léteznek hagyományosabb fórumok is az eszmecserére és személyes találkozókra. A legnevesebb R-es esemény az éves useR! konferencia, amelyet évente váltakozva hol Európában, hol az Egyesült Államokban rendeznek meg. Idén a Stanford Egyetemen 800 résztvevővel zajlott az esemény, jövőre Brüsszelben legalább ennyi résztvevőt várnak. Remélhetőleg az idei hibából tanulva sikerül javítani a logisztikán, és nem zárul le a regisztráció két hét alatt, a konferencia kezdete előtt hónapokkal a nagyon magas túljelentkezés hatására.

Ezen univerzális R konferencián túl számos egyéb többnapos nemzetközi rendezvény is működik világszerte. Az „R/Finance" konferenciát Chicago-ban rendezik évente május végén a pénzügyi kérdések iránt érdeklődőknek, körülbelül 300 résztvevővel – testvérrendezvénye a londoni „R in Insurance". A BioC konferenciára évente nyáron kerül sor biostatisztika fókusszal, de létezik az üzleti alkalmazásokra szakosodott előadássorozat is, például a jó pár éve ősszel Londonban és Bostonban is megrendezésre kerülő EARL (effective applications of the R language – hatékony R alkalmazások) konferencia. Kínában immáron 9. éve rendezik meg China-R konferenciát, amelyen a tavalyi látogatottsági adatok alapján körülbelül 4 ezer ember vesz részt évente. Az ERUM (European R Users Meeting – európai R felhasználók találkozója) konferencia idén először került megrendezésre Lengyelországban azzal a céllal, hogy az európai R felhasználók számára nyújtson alternatív találkozási lehetőséget azokban az években, amikor a useR! konferencia az Egyesült Államokban van.

A nagy nemzetközi konferenciákon túl számos egyéb kisebb esemény is színesíti az R felhasználók életét. Az R User Groups, az R felhasználói csoportok célja, hogy a helyi R közösség tagjainak nyújtson lehetőséget a rendszeres találkozásokra és fejlődésre elérhető fizikai távolságban és minimális költségekkel. Hagyományosan a meetup.com oldalon meghirdetett eseményeket 1-2 havonta szervezik, ahol egy hosszabb vagy több rövidebb szakmai előadás után a résztvevők egy pizza, üdítő vagy sör mellett informálisan beszélgetve tudnak ismerkedni egymással. Ezek a be-

szélgetések általában nagyon inspiráló szakmai eszmecserékké és hasznos együttműködésekké, olykor barátságokká mélyülnek.

A magyarországi R meetup 2013 augusztusa óta működik. Mára a tagok száma meghaladja a 800 főt, és havonta váltakozva angol és magyar nyelvű, minden esetben ingyenes előadásokat szervezünk mintegy 60–80 fő részvételével. További információ található a http://www.meetup.com/Budapest-Users-of-R-Network oldalon, ahol szívesen látunk haladó és teljesen kezdő R felhasználókat, továbbá érdeklődő látogatókat is.

A világon először idén került megrendezésre az R Consortium és számos hazai adatelemzéssel foglalkozó cég támogatásával a satRday konferencia Budapesten, amely egynapos, szombatra eső rendezvénysorozattal azt a célt tűztük ki magunk elé, hogy a havi rendszerességű meetupok és a nagy nemzetközi konferenciák között is lehetősége nyíljon az R felhasználóknak megismerkedni a környező országok R fejlesztőivel és felhasználóival. A rendezvény sikerét jelzi, hogy ezen első, kísérleti alkalommal is közel 200 fő regisztrált, és a résztvevők egyharmada külföldről érkezett, összesen 19 országból. A következő satRday konferenca a Dél-afrikai Köztársaságban és Puerto Ricoban lesz jövő év elején, hazánkba várhatóan 2018 tavaszán tér majd vissza.

9. R kiadványok és egyéb média

Természetesen az R nyelv és annak alkalmazásai a konferenciák és találkozók forgatagán túl, akár jóval formálisabb keretek között is elsajátítható. Az R akadémiai háttérének köszönhetően számos szakkönyv született az elmúlt húsz évben: az r-project.org oldal több mint 150 R-rel foglalkozó könyvet sorol fel, elsősorban a Springer és a Chapman & Hall/CRC kiadóktól. Az általános statisztikai tankönyveken és módszertani kiadványokon túl jelentek meg gyakorlatorientált alkalmazásgyűjtemények is különböző tudományterületekre fókuszálva, de találhatunk kifejezetten adatvizualizációval vagy például gépi tanulással foglalkozó köteteket is, illetve kifejezetten SPSS és SAS felhasználók számára is készült könyv (*Muenchen* [2011]).

A hazai R iránt érdeklődők számára magyarul is egyre több segédanyag áll rendelkezésre. Többek között ingyenesen elérhető *Solymosi* [2005], *Abari* [2008] és *Tóth* [2016] online jegyzete, de ezek mellett bátran merem ajánlani *Münnich–Nagy–Abari* [2006] pszichológus hallgatóknak írt, R példákat használó statisztika könyvét vagy *Reiczigel–Harnos–Solymosi* [2014] hasonlóan magas színvonalú és remekül használható biostatisztika és R bevezető tankönyvét.

Ezen szisztematikus gyűjtemények mellett az R Foundation évente 2-3 alkalommal jelenteti meg az *R Journal* (korábbi nevén *R News*) folyóiratot, amelyben a legújabb R csomagokról, továbbá a legfrissebb verzió újdonságairól olvashatunk. Az R csomagok részletesebb leírásainak hagyományosan a *Journal of Statistical Software* biztosít helyet, amely folyóirat elismertségét jól példázza, hogy évek óta igen magas (3 fölötti) az impakt faktora.

A lektorált szövegeken túl számos további ingyenes forrást találhatunk az Interneten, például a GitHub-on közzétett tutorialok, ingyenesen elérhető könyvfejezetek, R feladatgyűjtemények vagy akár blog posztok formájában. Az R-bloggers.com oldal célja a világhálón sok száz különböző helyen megjelenő, de hasonló témával foglalkozó cikk összegyűjtése, hogy egy naponta néhányszor frissülő központi helyen olvashassuk az R relevanciájú online írásokat.

Mindezeken túl az írott médiával szemben az audiovizuális tartalmakat előnyben részesítő érdeklődők számára is kínál kész megoldásokat az R közösség: többek között a coursera.org vagy a datacamp.com oldalon is találhatunk angol nyelvű előadásokat, online teszteket és komplett (ingyenes és fizetős) tanfolyamokat, amelyek akár iskola, akár főállás mellett is végezhetők heti néhány órában.

Irodalom

ABARI K. [2008]: *Bevezetés az R-be*. http://psycho.unideb.hu/munkatarsak/abari_kalman/szamitastechnika_II/bevezetes_az_R_be_2008_04.pdf

BOSTOCK, M. – OGIEVETSKY, V. – HEER, J. [2011]: D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*. Vol. 17. Issue 12. pp. 2301–2309. http://dx.doi.org/10.1109/TVCG.2011.185

CHAMBERS, J. M. [1980]: Statistical Computing: History and Trends. *The American Statistician*. Vol. 34. Issue 4. pp. 238–243. http://dx.doi.org/10.1080/00031305.1980.10483038

DARÓCZI G. – TÓTH G. [2013]: Felhőtlen statisztika a felhőben. *Statisztikai Szemle*. 91. évf. 11. sz. 1118–1142. old.

DARÓCZI, G. [2015]: *Mastering Data Analysis with R*. Packt. London.

FRANCIS, I. [1981]: *Statistical Software: A Comparative Review*. Elsevier North-Holland. New York.

GARETH, J. – WITTEN, D. – HASTIE, T. – TIBSHIRANI, R. [2013]: *An Introduction to Statistical Learning*. Springer Science+Business Media. New York. http://dx.doi.org/10.1007/978-1-4614-7138-7

HORNIK, K. [2012]: *The R FAQ*. http://cran.r-project.org/doc/FAQ/R-FAQ.html

LEEUW, J. DE [2011]: Statistical Software: An Overview. In: *Lovric, M.* (ed.): *International Encyclopedia of Statistical Science*. Springer-Verlag. Berlin. pp. 1470–1473. http://dx.doi.org/10.1007/978-3-642-04898-2_553

MUENCHEN, R. A. [2011]: *R for SAS and SPSS Users*. Springer. New York. http://dx.doi.org/10.1007/978-0-387-09418-2

MÜNNICH Á. – NAGY Á. – ABARI K. [2006]: *Többváltozós statisztika pszichológus hallgatók számára*. Bölcsész Konzorcium. Debrecen. http://psycho.unideb.hu/statisztika/

R CORE TEAM [2016]: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna. https://www.R-project.org

REICZIGEL J. – HARNOS A. – SOLYMOSI N. [2014]: *Biostatisztika nem statisztikusoknak*. Pars Kft. Budapest.

ROUTH, D. A. [2007]: Statistical Software Review. *British Journal of Mathematical and Statistical Psychology*. Vol. 60. Issue 2. pp. 429–432. http://dx.doi.org/10.1348/000711007X241151

SMITH, D. [2016]: *The history of R's predecessor, S, from co-creator Rick Becker*. Revolution Analytics Blog. http://blog.revolutionanalytics.com/2016/07/rick-becker-s-talk.html

SOLYMOSI N. [2005]: *R…erre, erre…! Bevezetés az R-nyelv és környezet használatába*. https://cran.r-project.org/doc/contrib/Solymosi-Rjegyzet.pdf

THE R FOUNDATION FOR STATISTICAL COMPUTING [2012]: *R: Regulatory Compliance and Validation Issues. A Guidance Document for the Use of R in Regulated Clinical Trial Environments*. Vienna. http://www.r-project.org/doc/R-FDA.pdf

TÓTH D. [2016]: *Bevezetés az R statisztikai programnyelv használatába*. ELTE PPK. Budapest. https://tdeenes.gitbooks.io/rintro_ma/content/

VIZVÁRI B. – BACSI ZS. [1997]: A magyar burgonyapiac kaotikus viselkedése. In: *Fokasz N.* (szerk.): *Rend és káosz*. Replika könyvek 4. Budapest. pp. 205–214.

VIZVÁRI B. [2002]. Dinamikus piacok és irányítás. *Magyar Tudomány*. XLVII. kötet. 10. sz. 1284–1297. old.

WILKINSON, L. [1999]: *The Grammar of Graphics*. Springer. New York. http://dx.doi.org/10.1007/978-1-4757-3100-2

ZEILEIS, A. [2005]: CRAN Task Views. *R News*. Vol. 5. No. 1. pp. 39–40.

Summary

First the paper gives a brief summary on the most important features of the R programming language and software environment for statistical computing, including the milestones towards its general popularity; then lists some of the most well-known R packages, tools, educational resources, conferences and other events related to the R community as a guideline for the novice user of R.