Využití R v oblasti financí

(Diskuse k využití statistického prostředí R ve financích)

Jiří Sedláček[*]

Úvodem

R se v současnosti stalo celosvětově jedním z nejpoužívanějších statistických balíků. Především na vysokých školách v USA, Kanadě, západní Evropě, v Austrálii a na Novém Zélandu je to pravděpodobně nejpoužívanější nástroj v oblasti statistiky vůbec. Stále více se rozšiřuje i v praxi. V ČR je zatím používán bohužel mnohem méně a také povědomí o jeho přednostech a možnostech je poměrně slabé.

Tento článek si proto klade *následující cíle*:

- Uvést *základní charakteristiku R* a také hlavní důvody, proč jej využívat (nejen) v oblasti financí.
- Stručně charakterizovat hlavní možná *uživatelská rozhraní* pro R, ač už jde o tzv. IDE (obecná, specializovaná), pluginy pro populární editory nebo grafická uživatelská rozhraní.
- Analyzovat hlavní *specializované datové struktury*, určené právě (ale nejen) pro data finanční povahy: ať už se jedná o datové struktury, které jsou již v základní instalaci R nebo jsou implementovány ve specializovaných balíčcích.
- Charakterizovat „*hlavní balíčky*“ (packages) pro R, které jsou zaměřeny na statistické zpracování finančních dat.
- Rozebrat některé základní možnosti přístupu k (nejen) finančním *datům na Internetu*: opět s využitím obecných nástrojů v R nebo s podporou specializovaných balíčků.

1. Základní charakteristika R a hlavní důvody pro jeho využití

R je statistický balík, autoři preferují pojem *prostředí pro statistickou analýzu dat a jejich grafické zobrazení* (The R Project for Statistical Computing and Graphics) a současně také označení *programovacího jazyka*, používaného v rámci toho prostředí (R Core Team, 2013).

Hlavní výhody a argumenty pro využití R:

- R dnes celosvětově (včetně komerční sféry) patří mezi *tři vůbec nejpoužívanější* statistické balíky (vedle SPSS a SAS), v akademické sféře je pak jeho pozice ve většině vyspělých zemí téměř dominantní. I když různá kvantitativní i kvalitativní srovnání je nutné brát s rezervou, určitou vypovídací schopnost přece jen mají, viz zejména obsáhlá analýza podle mnoha kritérií a z mnoha zdrojů dat (Muenchen, 2013).
- *Nejnovější statistické metody* jsou v současnosti zpravidla nejprve implementovány v R a teprve potom (a mnohdy se značným zpožděním) se objevují v komerčních balících jako je SPSS nebo SAS (o dalších, méně rozšířených programech ani nemluvě).
- Vedle základní instalace R, je jen na CRAN: Comprehensive R Archive Network (2013), což je hlavní repositář balíčku pro R, koncem října 2013 ke stažení *téměř 5000 tzv. balíčků* (packages: 26. října je to přesně 4959), které pokrývají široké spektrum statistických metod (včetně velmi specializovaných) a nejrůznější oblasti aplikace

[*] Ing. Jiří Sedláček, PhD. – odborný asistent; Katedra mezinárodního obchodu, Fakulta mezinárodních vztahů, Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 130 67 Praha 3; <sedlacek@vse.cz>.

statistiky. Minimálně dalších přes 2000 balíčků nebo projektů souvisejících s R lze najít v dalších repositářích nebo individuálně na jednotlivých webech (vedle CRAN např. R-Forge (2013) a Bioconductor (2013): jak již název napovídá, ten se zaměřuje na statistické metody v oblasti bioinformatiky, zejména na genetiku).

- R je *open source* projekt (šířený pod licencí GNU GPL). To znamená nejen obrovskou úsporu nákladů (cena jediné komerční licence velkých statistických balíků často přesahuje 1000 USD, i cena akademických licencí bývá dosti vysoká), ale nemusíte ani řešit správu licencí nebo to, zda můžete program instalovat též na domácím počítači. V neposlední řadě je tím podporována reprodukovatelnost výsledků (např. u řady časopisů je dnes běžné dát k dispozici i data a programový kód).
- R je *multiplatformní*: k dispozici pro Windows, Apple: (Mac) OS X a Linux, vždy v 32bitové a 64bitové verzi.
- Podpora široké škály *grafů* (včetně velmi specializovaných) v *publikační kvalitě*.
- Použití *jednotného jazyka R* ve všech oblastech (u jiných statistických programů se často musíte naučit různé jazyky pro vlastní statistickou analýzu, pro výstupy nebo pro grafy). Programovací jazyk je relativně snadný a je (svou syntaxí i sémantikou) orientován na vývoj statistických aplikací.
- Za významnou výhodu je podle našeho názoru třeba považovat i to, že program je *hardwarově poměrně nenáročný* (samozřejmě pokud nezpracováváte obrovská data nebo nepoužíváte velmi komplikované algoritmy). Instalátor R pro Windows v aktuální verzi 3.0.2 má pouhých 52 MB (a to obsahuje jak 32bitovou, tak 64bitovou verzi). Samozřejmě většina uživatelů bude instalovat ještě další balíčky, jejich velikost se ale typicky pohybuje od několika set KB po několik MB, takže ani to nepředstavuje problém (pokud jich nebudete mít stovky). Při kapacitách dnešních disků a výkonů procesorů se může zdát nepodstatné, ale mnoho studentů i učitelů dnes z důvodu snadné přenositelnosti používá netbook nebo sice výkonnější notebook, avšak s diskem typu SSD.
- *Reprodukovatelný výzkum* a *automatické generování publikačních výstupů* z „živých dat“ prostřednictvím nástrojů jako je R Markdown, R Sweave/knitr, R HTML, Shiny, do formátů LaTeX, PDF, Markdown, HTML (s využitím PanDoc nebo specializovaných balíčků i další formáty).
- To, že R dnes představuje hlavního konkurenta i nejvýznamnějším komerčním balíkům s mnohaletou tradicí dokládá např. i to, že SPSS v posledních verzích umožňuje ve vlastním prostředí vkládat příkazy R.

Je však třeba poctivé uvést i možné **nevýhody R**:

- Samotný fakt, že R je primárně orientováno na práci s příkazy, za nevýhodu nepovažujeme (spíše naopak) a je dáno samotnou povahou zejména složitějších statistických analýz. Koneckonců SPSS nebo Stata jsou také hodně „příkazově orientovány“. *Neexistence „oficiálního jednotného GUI“* však určitou nevýhodou asi je a zejména pro ty začátečníky, kteří na příkazově orientovaný způsob práce nejsou zvyklí, budou začátky v R těžší. (Pro R přesto existuje několik rozšiřujících GUI, které se instalují jako standardní balíčky nebo jako samostatný program a budou stručně popsány v další kapitole.)
- *Oficiální dokumentace* R je velmi obsáhlá a kvalitní, domnívám se však, že v nápovědě pro jednotlivé příkazy se většina začátečníků bude orientovat spíše s obtížemi (byť v ní zpravidla existují i jednoduché příklady). Myslím, že zde je ještě velký prostor pro zlepšení: při zachování detailů pro pokročilé zpřístupnit nápovědu více i pro začátečníky.

- U určité části příkazů existuje poměrně značná *nekonzistence* ve formálních pravidlech jejich zápisů (např. srovnejte příkazy *save, save.image, savehistory, savePlot*).
- Další možné nevýhody jsou již hodně specifické a týkají se většinou až pokročilých uživatelů. Např. R standardně načítá veškerá data do paměti (výhodou je mimo jiné rychlejší zpracování). Ale může to být problém při zpracování velmi rozsáhlých dat (lze řešit např. použitím specializovaných balíčků, které data uloží na disk, ale pro ostatní funkce se chovají transparentně jako by byla v RAM). Obdobně existuje několik balíčků pro podporu paralelismu a obecně situace, kdy zpracování velmi rozsáhlých dat velmi složitými algoritmy trvá příliš dlouho, lze řešit více způsoby (samozřejmě vedle využití výkonnějšího hardware).

2. Stručný přehled různých uživatelských rozhraní pro R

Jak jsme uvedli již v úvodním přehledu, R je primárně orientováno na práci s příkazy, což je třeba považovat spíše za výhodu. Základní (nativní) prostředí v R má však relativně omezené možnosti a v neposlední řadě se také liší podle platformy: ve Windows (kde obecně podpora práce na příkazové řádce je slabá) je implementováno programem RGui, na Linuxu R využívá nativní terminál (a bude se tedy lišit podle distribuce, resp. podle toho jaký emulátor terminálu máte nainstalován defaultně), v (Mac) OS X je opět poněkud odlišné.

Pro dlouhodobější práci v R lze proto velmi doporučit instalaci některého z mnoha nabízených uživatelských rozhraní, přitom určité možnosti se vzájemně nevylučují a někdy je naopak výhodné je kombinovat. Nejde o vyčerpávající přehled všech možných rozhraní a samozřejmě ne všechny jsem osobně vyzkoušel.

S jistým zjednodušením (některá rozhraní současně patří do více kategorii) je lze rozdělit následovně:

- Rozhraní, která se primárně snaží zjednodušit práci začátečníkům (nebo i pokročilejším uživatelům, zejména pokud danou funkci použijete jen občas, tudíž si příslušné příkazy nepamatujete). Sem patří především *R Commander*, *Deducer* a zčásti také *RKWard* (současně plní i funkci IDE).
- Rozšíření (pluginy) pro populární „programátorské“ editory a IDE (která lze dále ještě rozdělit na „lehká“ nebo „těžká“ a podle jiného kritéria na „univerzální“ a „specializovaná“).
- Některá další specializovaná rozhraní, kterými se zde nebudeme zabývat.

R Commander (Fox, 2005) je podle našeho názoru nejlepší řešení v kategorii „GUI pro R založené na menu a dialogových boxech“. Jedná se o standardní balíček pro R (název balíčků je *Rcmdr*), dostupný prostřednictvím CRAN, je tedy multiplatformní. Je průběžně vyvíjen a aktualizován (jak z hlediska nabídky funkcí, tak z hlediska podpory nových verzí R). V současnosti je aktuální verze 2.0 ze srpna 2013. Po volbách v menu a dialogových boxech se zobrazí i kód příslušných příkazů jazyka R, může tedy plnit i funkci určitého výukového prostředku. Hlavní omezení vyplývá ze samostatné podstaty takového řešení: prostřednictvím menu lze v zásadě provádět jen nejběžnější příkazy (i když v současnosti existuje více než 20 dalších pluginů pro další úkoly nebo balíčky).

Deducer (2013) je jiný obdobně zaměřený nástroj, poskytující systém menu pro základní manipulaci s daty a nejběžnější statistické analýzy, současně poskytuje jednoduchý spreadsheet pro prohlížení a editace dat. Je primárně navržen pro práci v dále popsané konzoli JGR (na všech platformách), ale lze použít např. i v rámci standardní Windows RGui. Deducer lze opět stáhnout z CRAN. Také pro Deducer existuje několik dalších rozšíření, přehled s odkazy na ně najdete na hlavní stránce produktu.

Rattle GUI (2013) je další graficky orientované rozhraní, tentokrát specializované na oblast data mining (dolování dat). Jedná se opět o standardní balíček pro R, který je dostupný na CRAN, ale nemáme s ním žádné osobní zkušenosti.

RKWard (2013) se od předchozích nástrojů odlišuje tím, že jde o klasický program, který v sobě kombinuje GUI rozhraní založené na menu a dialogových boxech, s rozšiřitelným integrovaným vývojovým prostředí (IDE), které zahrnuje obvyklé editační funkce, zvýrazňování syntaxe, dokončování kódu či code folding. Dále editor dat, jednodušší import dat a práci s grafy ad. RkWard je obecně multiplatformní (Linux, Windows, (Mac) OS X, ale protože je založen na KDE, nejvhodnější a nejméně problémová je jeho instalace právě na Linuxu (zejména pokud vaše distribuce KDE již využívá). Osobně s RKWard nemám zkušenosti, v diskusích si ho řada uživatelů pochvaluje právě pro výbornou kombinaci GUI a IDE, na druhé straně ve Windows si řada uživatelů stěžovala na problémy s instalací, což uvádí i oficiální dokumentace.

Java Gui for R, často označované též zkratkou **JGR** (což je i název balíčku, který lze stáhnout opět z CRAN) je multiplatformní rozhraní pro R, napsané (jak napovídá již název) v jazyce Java. Využitím Javy je přirozeně dosaženo multiplatformnosti, na druhé straně řadu uživatelů (zejména mají-li netbook či notebook se SSD diskem o omezené kapacitě) může odradit právě nutnost instalace Java JRE (pokud již nepoužívají jiné Java aplikace). JGR zahrnuje pokročilou R konzoli, editor dat, skript editor se zvýrazňováním kódů, dokončováním kódu a dalšími funkcemi. Neobsahuje menu systém, lze však vhodně doplnit výše popsaným nástrojem *Deducer*.

Pokud primárně nehledáte GUI rozhraní založené na menu, ale integrované vývojové prostředí (IDE) a dosud nemáte své oblíbené IDE nebo svůj oblíbený editor, na prvním místě bych doporučil **RStudio** (2013). Na rozdíl od některých „obecných IDE“ (které lze případně také využít pro R), RStudio je *specializované IDE*, vytvářené na míru právě s ohledem na specifické vlastnosti a požadavky vývoje statistických aplikací v systému R.

Stejně jako samotné R, RStudio je vydáno pod open source licencí a je k dispozici pro všechny hlavní platformy (Windows, Linux, Mac OS X). Navíc je možné RStudio instalovat na webový server a přistupovat k němu jako k webové aplikaci z libovolného moderního webového prohlížeče (serverovou verzí R Studia se zde nebudeme zabývat). Autoři R Studia jsou rovněž tvůrci několika významných a hojně používaných R balíčků, jako je *ggplot2*, *plyr*, *lubridate* a též webové aplikace *Shiny*, která umožňuje snadnou tvorbu interaktivních webových aplikací v R.

Mezi **hlavní výhody Rstudia** patří:

- *Jednotné uživatelské rozhraní* pro všechny platformy (Windows, Mac OS, Linux).
- *Poměrně pokročilý textový editor* s řadou funkcí usnadňujících psaní, úpravu a spouštění kódu jako je zvýrazňování syntaxe, párování závorek, (automatické) dokončování kódu, navigace v kódu ad.
- Snazší spouštění *libovolné části kódu*, přitom zůstává možnost i přímého vkládání jednotlivých příkazů do R konzole.
- Velké množství *klávesových zkratek*.
- *Integrace standardní nápovědy* R do prostředí RStudia.
- *Prohlížeč* objektů, tabulek (data frames).
- Snazší a jednodušší *práce s grafy* (přístup k více grafům, zvětšování/zmenšování, export do různých formátů).
- *Správa projektů*: vytváření projektů, práce s více projekty.
- Integrace *version control systems* jako je Git nebo SVN.
- Další specifická podpora pro *tvorbu R balíčků* (včetně tvorby jejich dokumentace).

- Elegantní a poměrně snadná tvorba *kvalitních reprodukovatelných publikačních* výstupů (ve formátech HTML, LaTex, PDF ad.) s využitím R Markdown, R Sweave nebo knitr či R HTML.
- *Přímá publikace* výstupů a skriptů on-line na webu RPubs.com a další.

Pokud již máte svůj oblíbený editor nebo IDE, případně potřebujete pracovat s více jazyky, může být vhodnější zjistit, zda váš program podporuje také R. Zde uvedeme jen několik nejznámějších nástrojů, pokud váš oblíbenec není zmíněn, prověřte si sami (vzhledem k rostoucí popularitě R se přidávají stále další programy).

Eclipse + StatET: Eclipse je velmi pokročilé a široce přizpůsobitelné multiplatformní IDE, které nativně nebo s pomocí pluginů podporuje velké množství jazyků (mimo jiné Java, Ada, C, C++, Fortran, Perl, PHP, Ruby ad.). Pro práci s R je určen plugin StatET, v současnosti verze 3.3 z června 2013 (WalWare, 2013). Vedle víceméně obvyklých funkcí jako mocný editor se zvýrazňováním syntaxe a dokončováním kódu bývá zdůrazňován především pokročilý debugger a pravděpodobně nejlepší prohlížeč objektů pro R.

Na druhou stranu, jde o prostředí skutečně „těžkotonážní“: jen samotné instalačky (Eclipse + StatET + Java JRE + R balíček rj) mají několik set MB a pro hladký běh samotného IDE spolu s R je potřeba dostatečně výkonný procesor a několik GB RAM. Na většině dnešních PC by to nemělo představovat zásadní problém, ale instalaci na slabší notebooky nebo dokonce netbooky rozhodně nedoporučujeme. Pro většinu běžných uživatelů R je toto IDE možná až zbytečně rozsáhlé a komplikované.

Emacs Speaks Statistics (označováno též zkratkou **ESS**). Emacs je jeden ze dvou „hlavních“ pokročilých editorů pro Unix/Linux, je však dostupný též pro Windows a další platformy. Je vyvíjen již desítky let, o čem svědčí mimo jiné číslo aktuální verze (24.3). Jeho základním rysem je mimořádně vysoká přizpůsobivost a možnost příkazy kombinovat do maker (používá vestavěný Emacs Lisp).

Emacs Speaks Statistics je potom rozšíření pro Emacs zaměřené na práci se statistickými programy (kromě R jsou podporovány také např. SAS, Stata, S-Plus nebo specializované programy pro určité typy Bayesovské statistiky: BUGS či JAGS). ESS je vyvíjen řadu let a podílí se na něm několik osob, které jsou též členy R Core Team. Ke stažení je k dispozici přímo z podwebu v rámci hlavního projektu R (ESS, 2013).

Vim je „hlavní konkurent Emacs“ mezi pokročilými editory pro Unix/Linux a pochopitelně je také dostupný pro Windows, (Mac) OS X a další platformy. Ve srovnání s Emacs je méně hardwarově náročný. Na rozdíl od ESS se na oficiálních stránkách R o Vim nemluví (alespoň my jsme nic takového nenašli). Bylo nám ale divné, že by Vim v současnosti R nepodporoval.

Po krátkém hledání na oficiálních stránkách editoru Vim jsme našli **Vim-R-plugin** (Aquino, 2013). Velikost tohoto pluginu je v aktuální verzi pouhých 163 KB. Na platformě Windows je však kromě toho nutné nainstalovat Python, jehož prostřednictvím jsou posílány příkazy do prostředí R (v Linuxu se pro totéž používá Tmux). Tento plugin není vyvíjen tak dlouho jako ESS, ale dle popisu na jeho domovské stránce by jeho funkcionalita měla být srovnatelná s ESS i jinými, zde popsanými IDE. Zahrnuje komunikaci s R, dokončování příkazů, prohlížení R dokumentace, prohlížeč objektů a zvýrazňování syntaxe nejen pro R, ale také pro R Markdown a R reStructuredText.

Geany (Wendling aj., 2013) je editor a lehké IDE pro Linux, Windows, (Mac) OS X a další platformy. Pro svou malou velikost bývá v řadě Linux distribucí již předinstalován, přitom (minimálně zvýrazňováním syntaxe, příp. také jednoduchou šablonou) podporuje velké množství (několik desítek) jazyků včetně R a také Markdown. Geany je proto vhodný i

pro starší notebooky a netbooky. Instalačka pro Windows je sice trochu větší, protože se současně instalují knihovny GTK2, které program využívá, ale jinak jsou jeho hardwarové nároky i zde malé. Pro uživatele Windows je určitě přínosem i to, že ovládání je (ve srovnání s Emacs nebo Vim) mnohem více podobné jiným Windows programům.

Tinn-R je na R specializovaný editor (dnes podporuje i jiné jazyky), dostupný pouze pro Windows (na CRAN k němu existuje i standardní R balíček TinnR s doplňkovými funkcemi). V rámci editorů/IDE podle mě nyní existují lepší (navíc multiplatformní) možnosti, Tinn-R však zůstává oblíben řadou autorů. Mimo jiné ho používají a/nebo doporučují některé neoficiální příručky (tzv. Contributed Documentation), které jsou ke stažení v rámci webu R-project.

Jiné editory/IDE: aspoň nějakou podporu pro R (minimálně zvýrazňování syntaxe) nabízí celá řada další programů. Např. ve Windows je dosti oblíbený **Notepad++** (již z názvu je zřejmé, že jde o výrazně vylepšenou náhradu standardního Poznámkového bloku) obsahuje mimo jiné plugin **NppToR** (2013) s podporou zvýrazňování kódu, dokončování kódu, code folding a zajišťuje též komunikaci mezi editorem a RGui. **PSPad** (Fiala, 2013) je i ve světě oblíbený volně šiřitelný (freeware) editor pro Windows českého původu. V současnosti je pro něj k dispozici rozšíření **accessR**, které zatím nabízí jen vybrané funkce pro R. **UltraEdit** (a UltraEdit Studio) jsou komerční (placené) programy. Pokud ho náhodou již vlastníte, je dobré vědět, že pro něj nyní též existuje (uživateli vytvořený) „wordfile“, tedy konfigurační soubor pro R. Wordfiles obecně podporují nejen zvýrazňování syntaxe, ale i párování, nápovědu k funkcím a code folding. Co z toho je k dispozici i pro R, si z pochopitelných důvodů již případně musíte vyzkoušet sami.

Shrnutí aneb jaké prostředí si tedy vybrat? Jak je vidět i z tohoto neúplného přehledu, nabídka je poměrně široká a různorodá a zdaleka ne všechny nástroje mám vyzkoušené. Přesto bych si dovolil uvést několik doporučení.

- Většině uživatelů doporučujeme *RStudio*. Jeho jednotlivé výhody už byly uvedeny výše. Za hlavní přednost však lze považovat to, jak jsou tyto funkce společně integrovány do jednoho programu, který přitom není nijak extra složitý, ani hardwarově náročný. Za další výhody lze považovat i to, že program je intenzivně vyvíjen týmem zkušených vývojářů, jsou nejen opravovány dílčí nedostatky a chyby, ale jsou přidávány i nové funkce, včetně reakcí na (smysluplné) připomínky uživatelů.
- Pokud hledáte GUI založené na menu a dialogových boxech, z několika možností bychom doporučili *R Commander*.
- Chcete-li prostředí, které v sobě kombinuje GUI s menu a dialogovými boxy spolu s vývojářským IDE, vyzkoušejte *RKWard*. Ve Windows je však nutná určitá opatrnost: pokud budete mít problémy s instalací nebo se stabilitou již nainstalovaného programu, bude vhodnější zvolit nějaká jiné řešení.
- Pokud již používáte nějaké jiné IDE nebo editor, asi bude nejvhodnější vyzkoušet, jak dobře podporuje i R. Možnosti jsou velmi odlišné a zahrnují např. lehké IDE *Geany* nebo velmi komplexní, ale taky velmi náročný (na hardware i uživatele) produkt *Eclipse*.
- Obdobná situace nastává i v případě, že potřebujete souběžně pracovat ve více jazycích.
- Za zvláštní zmínku stojí i editory *Emacs* a *Vim* (s příslušnými rozšířeními). Může se vyplatit investovat čas a úsilí do dobrého zvládnutí jednoho z těchto editorů a získáte tak velmi efektivní nástroj pro nejrůznější účely. Je však třeba upozornit, že pro uživatele Windows je ovládání těchto editorů velmi atypické (a oba editory se i navzájem značně liší).

3. Datové struktury důležité pro finanční data

Finanční data mají často charakter časových řad a pro časové řady existuje v R celá řada datových struktur (ať už implementovaných v základní instalaci R nebo prostřednictvím specializovaných balíčků).

3.1 Třída ts: time series

Základní datová struktura (třída, objekt) pro analýzu časových řad se podle očekávání jmenuje **ts:** což je zkratka ze slov *time series*. Tato třída je navržena pro zpracování časových řad s pravidelnými rozestupy: je tedy vhodná např. pro roční, čtvrtletní nebo měsíční data. Tato třída je implementována již v základní instalaci R, stejně jako řada příkazů (funkcí) pro práci s takovými datovými řadami. Ve stručnosti zmíníme jen několik základních.

Data časové řady bychom v realitě pochopitelně typicky načetli ze souboru, ale pro ilustraci si vytvoříme jednoduchou čtvrtletní časovou řadu, která obsahuje přirozená čísla 1 až 14 a začíná ve 3. čtvrtletí roku 2009. Základní prozkoumání časové řady je možné např. pomocí příkazů *start*, *end*, *frequency* nebo *deltat*. Pro zpoždění časové řady slouží příkaz *lag*, pro vytvoření první (nebo druhé, či sezónní) diference příkaz *diff*.

```
(radaQ = ts(1:14, start = c(2009, 3), frequency = 4))
##      Qtr1 Qtr2 Qtr3 Qtr4
## 2009                1    2
## 2010    3    4    5    6
## 2011    7    8    9   10
## 2012   11   12   13   14

start(radaQ)
## [1] 2009    3

end(radaQ)
## [1] 2012    4

frequency(radaQ)
## [1] 4

deltat(radaQ)
## [1] 0.25

lag(radaQ, -1)
##      Qtr1 Qtr2 Qtr3 Qtr4
## 2009                   1
## 2010    2    3    4    5
## 2011    6    7    8    9
## 2012   10   11   12   13
## 2013   14

diff(radaQ, 1)
##      Qtr1 Qtr2 Qtr3 Qtr4
## 2009                   1
## 2010    1    1    1    1
## 2011    1    1    1    1
## 2012    1    1    1    1

diff(radaQ, 1, 2)
##      Qtr1 Qtr2 Qtr3 Qtr4
## 2010    0    0    0    0
## 2011    0    0    0    0
## 2012    0    0    0    0

diff(radaQ, 4)
##      Qtr1 Qtr2 Qtr3 Qtr4
## 2010                4    4
## 2011    4    4    4    4
## 2012    4    4    4    4
```

Pro úplnost uveďme, že i pro analýzu časových řad se v R poměrně často používají datové rámce (*data.frame*), případně vektory. Důvodů může být více, jedním z hlavních je ten, že pro danou statistickou analýzu použijete příkazy (resp. balíček), které pro práci s časovými

řadami nebyly navrženy. Obvyklý postup je ten, že z dat dočasně odstraníte časové údaje (nebo ze souboru načtete jen část dat), tato data patřičně zpracujete a k výsledkům (např. pro grafy) opět doplníte časové údaje. V R pochopitelně existují též datové struktury matice (*matrix*), seznamy (*list*), které jsou rovněž implementovány již v základní instalaci, ale těmi se zde nebudeme blíže zabývat.

3.2 Třídy zoo a xts

Třída **zoo** (viz Zeileis, Grothendieck, 2005), která je implementována ve stejnojmenném balíčku *zoo*, je poměrně velmi abstraktní a je určena zejména pro *časové řady s nepravidelnými rozestupy* mezi jednotlivými hodnotami ukazatele. Údaje daného ukazatele jsou svázaný s množinou celočíselných hodnot „Z“, odtud i původ jména „zoo“. Pokud při vytváření konkrétního objektu typu zoo nespecifikujete explicitně hodnoty jednotlivých indexů, implicitně se použije vzestupná (pravidelná) sekvence 1, 2, 3, jak ukazuje následující ukázka. Jako zdroj dat použijeme generátor náhodných čísel normálního rozdělení. (Parametr width=100 je použit proto, aby se výpis vešel do jednoho řádku, standardně je 80.) Uložení dat ve struktuře zoo možné ještě lépe ukáže (jak jinak) standardní příkaz str. Pro výpis nebo pro modifikaci indexů lze použít příkaz index.

```
options(width = 100)
library(zoo)

(radaZOO = zoo(rnorm(10)))
##       1       2       3       4       5       6       7       8       9      10
## -0.4534 -0.8579  0.2803 -1.3464  0.2323 -1.6849  1.0327 -0.4264  1.5831 -0.2476

str(radaZOO)
## 'zoo' series from 1 to 10
##   Data: num [1:10] -0.453 -0.858 0.28 -1.346 0.232 ...
##   Index:  int [1:10] 1 2 3 4 5 6 7 8 9 10

index(radaZOO)
##  [1]  1  2  3  4  5  6  7  8  9 10
```

Ale smyslem třídy zoo jsou právě *nepravidelné* časové řady, jak alespoň trochu ilustruje následující příkaz. Vždy je možné data uložená v objektu typu ts převést na objekt typu zoo, ale obráceně to pochopitelně obecně neplatí, resp. pokud se takto pokusíte převést časovou řadu s nepravidelnými rozestupy, dojde ke kompletnímu zničení (odstranění) časových údajů (zůstanou jen samotné hodnoty ukazatele).

```
(radaZOO2 = zoo(rnorm(10), c(1, 3, 5, 7, 11, 13, 17, 19, 23, 27)))
##       1       3       5       7      11      13      17      19      23      27
##  1.0805  0.0068  1.0207 -0.7765 -1.4495 -0.3263  0.3182  1.9966  0.7345 -0.1844
```

Kromě třídy zoo se často používá z ní odvozená třída **xts** (*extensible time series*), definovaná opět ve stejnojmenném balíčku *xts* (Ryan – Ulrich, 2013). Rozestupy mezi ukazateli mohou opět být nepravidelné, ale musí to být uspořádané a unikátní hodnoty typu datum nebo čas (přesněji některá ze tříd „Date“, „POSIXct“, „timeDate“, „yearmon“, „yearqtr“) a na rozdíl od třídy zoo je nutné je vždy explicitně uvést. Do objektu lze rovněž uložit různá metadata (dvojice jméno=hodnota), např. datum poslední aktualizace, údaje o zdroji dat ap. Kromě metod obdobných jako u třídy zoo, balíček xts také upravuje metodu (příkaz) plot, tak aby po vykreslení grafu byly ihned zřejmé nepravidelné rozestupy mezi jednotlivými hodnotami ukazatelů. Kromě toho třída implementuje celou řadu specifických metod, včetně možností převodu z různých ostatních tříd pro časové řady (podrobněji viz oficiální dokumentace).

V ilustrativním příkladu vytvoření časové řady typu *xts* jsme použili malý trik. Využili jsme stejnou nepravidelnou řadu čísel 1, 3, 5, 7, 11… jako v předchozím příkladu pro třídu *zoo*, ale tato čísla jsme převedli na datum. A protože tato čísla jsou interpretována podle pravidel Unix (Linux) time, resp. POSIX time, kde první den epochy je 1. leden 1970 a ten

má pořadové číslo 0, získali jsme tak nepravidelnou časovou řadu od 2. ledna (pořadové číslo 1) do 28. ledna (pořadové číslo 27).

```
library(xts)
(radaXTS = xts(rnorm(10),
order.by = as.Date(c(1, 3, 5, 7, 11, 13, 17, 19, 23, 27))))
##                 [,1]
## 1970-01-02  2.0402
## 1970-01-04  0.6980
## 1970-01-06 -1.2118
## 1970-01-08  0.6063
## 1970-01-12 -2.0101
## 1970-01-14 -0.4862
## 1970-01-18 -1.3568
## 1970-01-20 -0.2059
## 1970-01-24  1.3974
## 1970-01-28  0.6761
```

3.3 Třída timeSeries

Tato třída je implementována v balíčku **timeSeries** (Wuertz – Chalabi, 2013), který je součástí poměrně velkého souboru balíčků souhrnně označovaných **Rmetrics**. Tento balíček závisí na dalším balíčků *timeDate*, který je používán v rámci celého souboru balíčků Rmetrics (souhrnný popis Rmetrics najdete v následující kapitole věnované charakteristice vybraných balíčků důležitých pro oblast financí).

Třída interně ukládá datové údaje dle *POSIXct*, ale navíc podporuje různé doplňkové funkce, které jsou důležité právě při zpracování (některých) finančních dat, především práce s parametry, které jsou vázány na jednotlivá finanční centra (viz níže) jako jsou časové zóny, dny svátků, ap.

```
library(timeSeries)
data = rnorm(9)
charvec = paste("2013-0", 1:9, "-01", sep = "")
(radaTimeSeries = timeSeries(data, charvec))

## GMT
##                TS.1
## 2013-01-01  0.9508
## 2013-02-01 -0.2408
## 2013-03-01 -0.2607
## 2013-04-01 -1.1273
## 2013-05-01 -1.8575
## 2013-06-01 -0.2176
## 2013-07-01  1.4053
## 2013-08-01  0.3486
## 2013-09-01 -1.4667
```

3.4 Další třídy (irts, its, fts; ti, tis)

Pravděpodobně ne tak často používané (pro řadu uživatelů mohou být užitečné, osobně však s nimi nemáme zkušenosti) jsou třída **irts**, implementovaná v balíčku *tseries*, třída **its** ze stejnojmenného balíčku a třída **fts** opět v balíčku stejného jména. Jsou opět určeny pro nepravidelné časové řady a všechny tyto tři třídy ukládají časové údaje ve formát *POSIXct*. Liší se však některými dalšími vlastnostmi a také pořadím argumentů.

Třídy **ti** (Time Index) a **tis** (Time Index Series) jsou implementovány v balíčku *tis*. Jejich koncepce má blíže ke standardní třídě ts, jsou však flexibilnější a také jsou kompatibilní s databázovým systém FAME. V rámci balíčku je implementována celá řada užitečných metod. Time Index má dvě části: *tif* (Time Index Frequency) a periodu. Time Index Series je vektor (příp. matice), indexované pomocí *ti*.

Jakou třídu tedy zvolit? Na tuto otázku není jednoduchá odpověď, lépe řečeno volba závisí hlavně na dvou faktorech:

- Jaký je charakter dat v časové řadě, kterou chcete zpracovávat?
- Jaké třídy podporuje balíček, který chcete nebo potřebujete využít. Některé balíčky podporují i více tříd, jiné jen jednu.

Pokud je daná časová řada pravidelná a nepotřebujete nějaké speciální doplňkové vlastnosti nebo příkazy, základní třída *ts* je sázkou na jistotu: podporuje ji velké množství balíčků. Pokud je časová řada nepravidelná, třída *zoo* a/nebo *xts* může být dobrou volbou, protože obě jsou dosti rozšířené. Všechny balíčky v kolekci Rmetrics používají vlastní třídu *timeSeries*, která navíc obsahuje zajímavé doplňkové vlastnosti, užitečné právě pro zpracování finančních dat. Je ale docela možné, že budete muset použít některou další (zde jen krátce zmíněnou) třídu nebo dokonce třídu zde ani neuvedenou.

4. Vybrané balíčky R pro oblast financí

Jak již bylo uvedeno v úvodní charakteristice statistického prostředí R, jedním z jeho význačných rysů je jeho *extrémní rozšiřitelnost*: vedle základní instalace jsou k dispozici tisíce balíčků, které pokrývají téměř každou oblast aplikace statistiky a nejrůznější statistické metody (a další balíčky jsou zaměřeny na specifické pomocné úlohy).

Je zřejmé, že při takovém množství balíčků nelze ani zmínit (natož se jim blíže věnovat) všechny balíčky, které se potenciálně týkají oblasti financí. Kromě toho např. jen pro zpracování časových řad existují desítky různě specializovaných balíčků. Zaměřím se proto jen na několik, podle mého názoru klíčových a široce využitelných balíčků.

4.1 Balíček quantmod: Quantitative Financial Modelling Framework

Balíček **quantmod** (Ryan, 2013) je nepříliš velký (v aktuální verzi pro Windows 433 KB) primárně určený pro import vybraných finančních dat a rychlou analýzu takovýchto dat (především různé varianty grafů). Vyžaduje balíček *xts* (který implementuje stejnojmennou datovou strukturu, popsanou v předchozí kapitole). Balíček xts zase vyžaduje rovněž výše popsaný balíček *zoo*, ale při standardním způsobu instalace se tím nemusíte zabývat. R používá ve svém správci balíčků obdobný princip jako většina Linux distribucí a tyto závislosti jsou řešeny automaticky.

Balíček nabízí velmi jednoduchý import dat z vybraných zdrojů (viz následující kapitola). Z dalších příkazů je zdaleka nejpoužívánější chartSeries, který již v základní variantě zobrazí použitelný graf. Přidáním dalších parametrů lze doplnit např. Bollingerovy pásy, objem obchodů, klouzavé průměry, RSI index ap.

4.2 Balíček RQuantLib

Tento balíček (Eddelbuettel – Nguyen, 2013) je poněkud neobvyklý v tom, že vlastně jde o R rozhraní k poměrně rozsáhlé *C++ knihovně QuantLib* (open source projekt, který je vyvíjen nezávisle na R). Instalace balíčku ve Windows (cca 6 MB) je stejně snadná jako u jiných balíčků, protože binárka obsahuje již vše potřebné. V Linuxu to však může být komplikovanější (v závislosti na tom, jakou používáte distribuci). Obvykle je totiž nutné nejprve ze zdrojových kódů zkompilovat samotnou knihovnu *QuantLib* a také knihovnu *Boost*, teprve pak lze již obvyklým způsobem instalovat samotný balíček.

Balíček obsahuje řadu nástrojů pro kvantitativní analýzu, modelování, obchodování a řízení rizik. Konkrétně lze zmínit např. různé funkce pro evropské či americké opce, funkce pro různé druhy dluhopisů (s pevným, proměnlivým či nulovým úročením). Součástí jsou i doplňkové funkce, např. široká paleta funkcí pro práci s kalendářem (zjišťování obchodních dnů a svátků, zda je víkend, konec měsíce, počet dní: všech nebo jen obchodních atd.)

4.3 Soubor balíčků Rmetrics

Rmetrics je souhrnné označení pro poměrně velký soubor (téměř 20) vzájemně provázaných balíčků, které společně pokrývají široké spektrum funkcí z oblasti financí. Označení Rmetrics pochází z toho, že jsou vyvíjeny v rámci *Rmetrics Association* (2013), což je nezisková organizace se sídlem ve Švýcarsku. Na webových stránkách této organizace jsou mimo jiné nabízeny též různé knihy v elektronickém formátu. Většina je placená, ale několik lze stáhnout zdarma. Jedna z nich se zabývá editorem *tinn-R*, který byl krátce zmíněn výše, jiná se věnuje diskusi o datových strukturách pro časové řady.

Tato kolekce balíčků byla původně vyvíjena primárně pro výukové účely (v dokumentaci většiny balíčků stále ještě najdete označení Environment for teaching „Financial Engineering and Computational Finance“), ale dnes je to v podstatě již plnohodnotná kolekce, i když zejména u některých balíčků autoři důrazně varují, že názvy funkcí nebo jejich argumenty, stejně jako default hodnoty se mohou změnit. V tomto stručném přehledu se nebudu zabývat všemi balíčky, ale většinu jich zmíním.

timeDate: Rmetrics – Chronological and Calendar Objects (Wuertz – Chalabi, 2013). Jak napovídá již název, tento balíček implementuje celou řadu důležitých funkcí pro práci s kalendářem a časem a je využíván všemi ostatními balíčky Rmetrics. Balíček podporuje různé formáty pro prací s kalendářovými a časovými údaji, implementována je i funkce *whichFormat*, která dokáže automaticky identifikovat většinu způsobů zápisu kalendářových a časových údajů. Důležitá je samozřejmě také práce s časovými zónami (např. když potřebujete synchronizovat data z různých zdrojů).

Vedle celé řady „obvyklých“ funkcí typu první nebo poslední den v měsíci (čtvrtletí), je-li daný den pracovní či svátek ap., za zmínku stojí skupina funkcí začínající slovem *holiday*, které pro zvolené finanční centrum vypíše dny svátků. Příkaz *listFinCenter* vypíše (průběžně rozšiřovaný) seznam finančních center. Např. pro Evropu je to v aktuální verzi 55 položek, včetně Prahy. Pro každý finanční centrum balíček „zná“ základní informace, jako je časová zóna a geografické umístění.

```
library(timeDate)
(listFinCenter("Europe"))
## [1] "Europe/Amsterdam"  "Europe/Andorra"  "Europe/Athens"      "Europe/Belgrade"
## [5] "Europe/Berlin"  "Europe/Bratislava"  "Europe/Brussels"    "Europe/Budapest"
...
## [33]"Europe/Podgorica"  "Europe/Prague"   "Europe/Riga"        "Europe/Rome"
...
## [53] "Europe/Zagreb"    "Europe/Zaporozhye"  "Europe/Zurich"

(datum1 = timeDate("2013-10-15", Fin = "Europe/London"))
## Europe/London
## [1] [2013-10-15 01:00:00]

(datum2 = timeDate("2013-10-15", Fin = "Europe/Prague"))
## Europe/Prague
## [1] [2013-10-15 02:00:00]

(datum3 = timeDate("2013-10-15", Fin = "America/New_York"))
## America/New_York
## [1] [2013-10-14 20:00:00]
```

timeSeries: Rmetrics – Financial Time Series Objects. Druhý základní pomocný balíček ze skupiny Rmetrics především implementuje třídu *timeSeries*, která již byla popsána výše v celkovém přehledu datových struktur pro časové řady. Kromě toho je pochopitelně implementována celá řada příkazů pro operaci s daty uloženými v této datové struktuře. Za zmínku určitě stojí příkaz *isRegular*, který umožňuje otestovat, zda se jedná o pravidelnou časovou řadu a tedy, zda je např. možné data převést na standardní objekt *ts*. Příkaz *readSeries* je

analogií standardního příkazu *read.table*, avšak data jsou rovnou načtena do objektu typu *timeSeries*. Příkaz používá i stejné základní parametry jako jsou *header* nebo *sep*, navíc lze specifikovat i volitelné parametry *zone, FinCenter, format*.

fBasics: Rmetrics – Markets and Basic Statistics (Wuertz et al., 2013a). Jak opět napovídá už název, tento balíček implementuje některé základní funkce pro zpracování finančních dat. První skupina funkcí počítá převážně základní ukazatele deskriptivní statistiky, druhá implementuje funkce pro různá statistická rozdělení nebo zjišťování jejich parametrů, třetí skupina obsahuje velké množství testů pro testování hypotéz, čtvrtá obsahuje příkazy pro kreslení grafů, pátá zahrnuje různé maticové operace a poslední šestá zahrnuje různé pomocné funkce.

fImport: Rmetrics – Economic and Financial Data Import (Wuertz et al., 2013a). Další důležitý pomocný balíček implementuje metody pro import finančních dat (zatím) z několika webových zdrojů (Yahoo, Oanda, the Federal Reserve). Dále je zde několik obecných funkcí pro čtení a zápis dat, za zvláštní zmínku stojí funkce, které umožňují spolupráci s několika prohlížeči (např. Links, Lynx) a dále je zde několik předdefinovaných CSV textových souborů, které popisují strukturu dat vybraných zdrojů.

fOptions implementuje různé metody pro oceňování základních druhů evropských a amerických opcí. **fAsianOptions, fExoticOptions** se obdobně zabývá asijskými a dalšími druhy opcí. **fBonds** se samozřejmě zabývá dluhopisy, avšak v současnosti zřejmě patří mezi nejméně dokončené balíčky, soudě alespoň podle toho, že PDF dokumentace má jen tři stránky a obsahuje jedinou metodu (Wuertz, Diethelm et al., 2012). **fAssets** implementuje různé metody pro práci s aktivy. **fPortfolio**: implementuje řadu metod pro výběr a optimalizaci portfolia. K tomu balíčků existuje i (placená) eletronická kniha *Portfolio Optimization with R/Rmetrics*.

fRegression: Regression Based Decision and Prediction. Balíček implementuje poměrně širokou škálu regresních metod (např. klasický lineární model, robustní lineární model, generalized linear model, generalized additive model ad.), které byly adaptovány pro práci s časovými řadami ve výše popsaném objektu *timeSeries*. Součástí jsou pochopitelně i regresní testy a upravené metody pro kreslení grafů z těchto datových struktur.

4.4 Další vhodné balíčky

Tím samozřejmě nabídka balíčků v oblasti financí zdaleka nekončí, přičemž jejich funkce se často aspoň zčásti překrývají s balíčky výše uvedenými. Jen krátce bez bližšího popisu bych zmínil např. balíček *portfolio* (nezaměňovat s výše uvedeným *fPortfolio* z Rmetrics) a související *porfolioSim* pro simulace. Balíček *PerformanceAnalytics* je další balíček obdobného zaměření. *OptHedging* se opět zabývá opcemi a vytvářením strategie optimálního zajištění, opcím se věnuje i balíček *AmericanCallOpt*. Dále existuje mnoho balíčků, které implementují jednu nebo dvě specializované metody či testy, např. *schwartz97, BenfordTest, FinAsym* ad.

5. Zdroje (nejen) finančních dat na Internetu a jejich zpracování

Obecně existuje několik základních variant, jakým způsobem jsou (nejen finanční) data na Internetu k dispozici a podle toho se liší i výchozí přístup k jejich získání pro následné statistické zpracování.

- Data nejsou ke stažení, je nutné je extrahovat z HTML stránek.
- Data ke stažení v nějakém rozšířeném formátu.
- Existuje veřejné API, návazně bývá přímá podpora v R.

Pokud data *nejsou* vůbec ke stažení, je nutné je **extrahovat z HTML** stránek. Obecně jde o tu nejhorší variantu, protože je nutné řešit „případ od případu“, podle konkrétní struktury data a struktury HTML stránek. Vzhledem k nutnosti individuálního řešení zde zmíníme jen základní možné nástroje.

Data stáhneme externími nástroji (mimo R), nám se např. osvědčil open source program *GNU Wget*. Ten je primárně vyvíjen pro Linux, ale existuje i Windows a (Mac) OS X verze. Obdobný charakter má i program *cURL*. Pokud hledáte nějaký program s GUI rozhraním, můžete vyzkoušet např. *HTTrack*. Dříve než se pustíte do zpracování samotných HTML stránek, je často nutné řešit různé nestandardnosti: např. ručně doplnit seznam stránek ke stažení o atypická URL, příp. řešit „problémový obsah“ některých stažených stránek. Obvykle též není možné stáhnout větší počet HTML stránek najednou, mezi jednotlivá stažení je nutné vkládat umělé pauzy.

Pro extrahování vybraných dat je nejprve nutné prostudovat interní strukturu HTML stránek. V jednodušších případech lze data vybrat vhodně sestaveným *regulárním výrazem*, dále lze použít např. programy *AWK* nebo *sed*. Jedná se o mocné nástroje, ale pro většinu lidí mohou být dosti obtížně. Pokud aspoň trochu umíte některý „klasický“ programovací jazyk, může být pro vás vhodnější napsat si jednoduchý jednoúčelový program.

Pokud chcete využít nástroje přímo v R, přichází v úvahu např. příkaz *scan*, který je dostupný již v základní instalaci. Ten lze přizpůsobit velkých množstvím (volitelných) parametrů, takže jeho osvojení také zabere nějaký čas. Zajímavou možnost nabízí balíček *fImport*, který (a) obsahuje příkaz *read-downloads* a (b) umožňuje systémová volání pro několik textově orientovaných WWW prohlížečů. Tyto příkazy však zatím nemáme vyzkoušené.

Lepší (a zejména pro finanční data naštěstí četnější) varianta je, že data jsou **ke stažení v souboru** (jiném než HTML stránka). Pokud se jedná o textový soubor a potřebuje ho číst po řádcích, v R tomu slouží standardní příkaz *readLines*, který mimo jiné umožňuje specifikovat, že se má maximálně přečíst jen určitý počet řádků.

Poměrně hodně často bývají data k dispozici v *textových souborech typu CSV*. Doslovný překlad zkratky je „čárkou oddělené hodnoty“, ale oddělovačem může být a často také bývá i jiný znak. Základním příkazem v R pro tyto účely je příkaz *read.table*, který má opět mnoho parametrů: mezi nejdůležitější patří *header* (logický: zda v prvním řádku jsou názvy polí), *sep* umožňuje zadat libovolný oddělovač, *dec* určit zda čísla používají desetinnou čárku nebo tečku a *skip* určit kolik řádků na začátku se má přeskočit. Příkazy *read.csv*, *read csv2*, *read.delim* a *read.delim2* vlastně nejsou samostatné příkazy, ale varianty příkazu *read.table* s různým implicitním nastavením parametrů *header*, *sep* a *dec*.

V případě zpracování časových řad potom takto načtená data můžete převést do vámi požadované varianty datové struktury. Některé typy tříd z rozšiřujících balíčků obsahují i obdobu příkazu *read.table*, který data načte rovno do požadovaného objektu, jak už jsme uvedli výše.

Nejen (finanční) data bývají často ke stažení také ve *formátu Excel* (ve starší verzi XLS nebo novější XLSX). Hodně renomovaných autorů v R doporučuje (máte-li tuto možnost, což předpokládám většina čtenářů tohoto článku má) nejprve tyto soubory otevřít v Excelu, případně upravit názvy sloupečků pro lepší další zpracování, soubory uložit do CSV a dále použít výše popsaný příkaz *read.table*. Jinak v R pochopitelně existuje poměrně hodně balíčků (např. i balíček *fImport*), které se přímému importu dat z různých verzí formátu Excel (v různém rozsahu a různým způsobem) zabývají. Některé z těchto balíčků však nejsou multiplatformní, příp. vyžadují instalaci dalších komponent.

Databázové a jiné formáty: pro celou řadu dalších formátů opět existují specializované balíčky, příp. opět můžete zvolit strategii nejprve soubor převést do jiného, pro zpracování snazšího formátu.

Nejlepší varianta je, pokud data jsou dostupná (pokud možné v jednotné struktuře) prostřednictví **veřejného API**. V těchto případe obvykle existuje též další **přímá podpora v R** prostřednictvím specializovaného balíčku (balíčků).

5.1 Web Quandl a balíček Quandl

V současnosti asi nejzajímavějším zdrojem mimořádně širokého spektra nejen finančních dat je projekt firmy se sídlem v Torontu: *Quandl*, někdy označovaného jako „wikipedia pro data v časových řadách“ nebo „vyhledávač pro numerická data“. Firma byla založena v roce 2011, samotný web je v provozu asi rok. V současnosti na své hlavní stránce uvádí, že k dispozici je přes 7 miliónů datových souborů z oblasti finanční, ekonomické a sociologické (Quandl, 2013). Quandl čerpá data ze širokého spektra zdrojů prostřednictvím vlastního „univerzálního parseru“ a všechna je zpřístupňuje v jednotném formátu (nejvíce, cca 2,1 mil. pochází ze zdrojů OSN, kolem 900 tisíc finančních dat je z projektu Open Financial Data Project, téměř 800 tisíc je ze Světové banky, využívá se asi 12 dalších mezinárodních organizací, dále centrální banky jednotlivých zemí, řada statistických a jiných institucí USA a spousta dalších institucí a zdrojů).

Představu o tom, jak zajímavé a široké spektrum dat je k dispozici lze získat již jakýmkoliv běžným WWW prohlížečem. Jen pro ilustraci: např. pokud jde o měnové kurzy, jsou vůči USD k dispozici zhruba pro 200 měn, obdobná data jsou nyní též pro digitální měnu Bitcoin. V případě komodit je v současnosti sledováno 86 druhů komodit a 12 komoditních indexů. Zdaleka nejvíce dat je pochopitelně k dispozici pro USA (např. data pro více 15 000 druhů akcií), pro ČR se zpočátku nabízelo jen několik základních makroekonomických řad, ale v současnosti je i kolekce dat pro naši republiku velmi slušná a rovněž je nabízeno srovnání (v daném ukazateli) s ostatními zeměmi.

Pochopitelně lze data prohlížet, nechat si zobrazit graf a přímo z WWW prohlížeče lze zvolená data také stáhnout ve formátech CSV, JSON, XML a Excel. O současném postavení R svědčí i to, že poslední páté tlačítko „R“ zobrazí malé okénko, do kterého je vygenerován jednořádkový kód pro R, který lze zkopírovat do standardní R konzole (pokud nepoužíváte jiné uživatelské rozhraní), do editoru v R Studiu nebo jiného, vámi oblíbenému editoru (viz popis uživatelských rozhraní v úvodu článku) a příkaz se tak snadno může stát součást vašeho skriptu. Příklad (doplnili jsme pouze přiřazení „usdeur =„ načtených dat, aby se nevypisovala všechna data a poté příkazem *head* vypsali jen prvních 10 řádků časové řady):

```
usdeur = read.csv("http://www.quandl.com/api/v1/datasets/QUANDL/
USDEUR.csv?&trim_start=1999-09-06&trim_end=2013-11-01&sort_order=asc",
colClasses = c(Date = "Date"))
head(usdeur, 10)
##          Date   Rate High..est. Low..est.
## 1  1999-09-06 0.9410     0.9527    0.9295
## 2  1999-09-07 0.9455     0.9564    0.9348
## 3  1999-09-08 0.9444     0.9559    0.9330
## 4  1999-09-09 0.9437     0.9541    0.9334
## 5  1999-09-10 0.9480     0.0000    0.0000
## 6  1999-09-13 0.9598     0.9714    0.9484
## 7  1999-09-14 0.9649     0.9762    0.9536
## 8  1999-09-15 0.9654     0.9771    0.9538
## 9  1999-09-16 0.9631     0.9739    0.9525
## 10 1999-09-17 0.9620     0.0000    0.0000
```

Co činí Quandl ještě zajímavějším pro statistické zpracování dat v R je (stejnojmenný) *balíček Quandl*: Quandl Data Connection (McTaggart – Daroczi (2013), který lze v současnosti stáhnout přímo z hlavního repositáře CRAN, na stránkách Quandl (resp. na GitHub,

odkud z nich vede odkaz) je však zpravidla novější verze, zvláště pokus si zvolíte větev „develop“, někdy je však rozdíl i ve větvi „master“ (každý balíček zařazený do CRAN prochází poměrně důkladnou kontrolou, která zabere nějaký čas). (V současnosti je obdobná funkcionalita dostupná i pro další jazyky a statistické nebo matematické programy, např. Python, Stata či Matlab, R však bylo první a tento balíček je často aktualizován.)

Po instalaci balíčku standardně můžete jeho prostřednictví provést 50 volání Quandl API denně, po bezplatné registraci (je vyžadováno jméno, uživatelské jméno, e-mail a pochopitelně heslo; alternativně se lze zaregistrovat prostřednictvím účtů na Linkedin, Google, GitHub nebo Twitter) získáte autentizační token a 500 volání API denně. Pokud byste potřebovali ještě vyšší limity, lze je dohodnout prostřednictvím e-mailu connect@quandl.com. Kromě toho, po registraci máte možnost vytvářet vlastní tzv. *supersets* (pojmenované uživatelské kolekce dat), záložky na oblíbená data, příp. také můžete nahrát vlastní data.

5.2 Data dostupná prostřednictvím balíčku quantmod a fImport

Výše stručně popsaný balíček **quantmod** podporuje přímý import dat z webu Yahoo! Finance, Google Finance, FRED (Federal Reserve Economic Data, zahrnuje i mezinárodní data) a OANDA (Forex Trading and Exchange Rates Services). Tyto zdroje (snad s výjimkou OANDA) jsou nyní dostupné i prostřednictvím výše popsaného webu *Quandl*, kde najdete mnohem více dalších dat. Většinou proto bude výhodnější výše popsaný postup.

Přesto má import prostřednictvím balíčku *quantmod* smysl. Jednak je možné, že se zcela neshoduje konkrétní množina dostupných dat, ale hlavně pokud budete data zpracovávat prostřednictvím tohoto balíčku, je jednodušší provést i import zde. Základní příkaz pro tyto účely je příkaz *getSymbols*, implicitní zdroj je Yahoo. Zde stačí jako parametr uvést čtyřpísmennou zkratku, pod kterou jsou akcie dané firmy obchodovány na burze. Syntax pro ostatní weby je stejný, pouze je nutné doplňkový parametr „src“ (např. src = "google").

Obdobnou funkcionalitu nabízí i balíček **fImport** (Yahoo, FRED, OANDA), který můžete využít samostatně, hlavní smysl však dává jeho využití spolu s dalšími balíčky z kolekce *Rmetrics*, protože data načítá přímo do třídy *timeSeries*.

5.3 České projekty otevřených dat

Jak je zřejmé již z výše uvedeného, zejména v USA a ve Velké Británii je dostupné velké množství nejrůznějších dat (ze státní správy, z univerzit, ale i z firem) *prostřednictvím veřejného API*. Bohužel v ČR jsou tyto snahy teprve v počátcích. Na stránkách *OpenData.cz* (2013), což je mimochodem společný projekt FIS VŠE a MFF UK, najdete kromě užitečných informací především *katalog dat ČR*.

Ten obsahuje odkazy na datové sady zveřejňované veřejnou správou publikované v různých formátech (strukturovaných i nestrukturovaných) jako je csv, xml, excel, word, pdf, html ad. V současnosti je to zhruba 200 položek a dalších 160 sdílených z UEP. Bohužel u naprosté většiny položek je aspoň zatím uvedeno „nemá otevřenou licenci“. Přes tato omezení je web již nyní poměrně zajímavých zdrojem, kde hledat vhodná data a jeho význam, stejně jako rozsah katalogu, nepochybně v budoucnu poroste.

Doplňkem tohoto webu je projekt *Společně otevíráme data* (2013), což je soutěž pro aplikace a projekty nad otevřenými daty. 25. října byl ukončen příjem přihlášek a do 28. listopadu by měli být vyhlášeni vítězové.

Závěr

R je v současnosti jedním z hlavních nástrojů pro statistickou analýzu dat, zejména v akademické oblasti. V komerční sféře je jeho postavení malinko slabší, ale i zde jeho význam roste, mimo jiné tím, že z většiny světových univerzit vychází mnoho absolventů, kteří během studia používali právě R. Bohužel v ČR je R zatím méně rozšířené.

V úvodní kapitole článku jsme proto shrnuli hlavní důvody pro využití R. V druhé kapitole jsme porovnali hlavní možná uživatelská rozhraní pro R, která mohou práci zjednodušit, zpříjemnit a zefektivnit. Jedna skupina rozhraní rozšiřuje prostředí R v tom smyslu, že poskytuje základní GUI pro práci s nejběžnějšími funkcemi. Další skupina nahrazuje původní, poměrně spartánské rozhraní, pokročilým editorem a obvykle i kompletním integrovaným vývojářským prostředí (IDE). Podle našeho názoru velmi vydařené RStudio je příkladem IDE, které bylo vytvořeno „na míru R“, navíc v sobě integruje celou řadu užitečných doplňkových funkcí. Řada dalších nástrojů integruje práci v R do prostředí obecnějšího editoru/IDE.

Protože článek je zaměřen na využití R v oblasti financí, ve 3. kapitole jsme analyzovali datové struktury, vhodné pro zpracování takových dat, od základní třídy ts až po několik variant tříd, které podporují i nepravidelné časové řady, umožňují ukládat metadata ap. Dospěli jsme k závěru, že primárně bychom měli volit podle specifického charakteru dat. V některých případech se však naopak musíme přizpůsobit tomu, jaké třídy podporuje implementace zvolené statistické metody.

Na 3. kapitolu navazuje čtvrtá kapitola, která se věnuje stručnému rozboru několika základních balíčků zajímavých pro zpracování finančních dat. Quantmod je spíše jednodušší a menší balíček zaměřený zejména na graficky orientovanou analýzu finanční dat (vhodné zejména pro osoby inklinující k technické analýze), integrální součástí jsou i nástroje na import dat z několika důležitých zdrojů.

Balíček RQuantLib je vlastně R rozhraní k poměrně rozsáhlé C++ knihovně QuantLib, která je vyvíjena nezávisle na R. Balíček implementuje poměrně širokou funkcí příkazů pro evropské a americké opce, pro různé druhy dluhopisů a mnoho doplňkových funkcí (např. kalendářových).

Rmetrics neoznačuje jeden balíček, ale ucelenou kolekci téměř 20 balíčků, z nichž zhruba tři poskytují základní funkce pro zbylé, které se pak zaměřují na různé oblasti financí (např. na opce, dluhopisy, aktiva, správu portfolia, regresní a jiné metody).

Závěrečné kapitola potom zkoumá, jak a kde hledat a získat vhodná data na Internetu. Nejprve jsou analyzovány základní možnosti a nástroje s ohledem na to, v jaké podobě a struktuře jsou data ke stažení. Potom se soustřeďujeme na nejpříznivější variantu, kdy data v definované struktuře jsou dostupné prostřednictvím veřejného API. V těchto případech obvykle existuje též korespondující balíček v R, který zpracování takových dat dále usnadňuje.

V současnosti asi nejvýznamnějším zdrojem (nejen) finančních dat je web Quandl. Ten již prostřednictvím webového rozhraní nabízí specifickou podporu pro R tím, že vygeneruje příslušný příkaz pro stažení konkrétního datového souboru. Možnosti práce s daty tohoto webu dále usnadňuje stejnojmenný R balíček. Stručně jsou rozebrány i možnosti importu v rámci dalších probíraných balíčků a na závěr jsou charakterizovány (zatím počáteční) české snahy o otevřená data on-line.

Literatura:

[1] Aquino, J. (2013): *Vim-R-plugin: Plugin to work with R*. [on-line], Vim.org, c2013, [cit. 31. 10. 2013] <http://www.vim.org/scripts/script.php?script_id=2628>.

[2] Bioconductor (2013): *Bioconductor: Open Source Software for Bioinformatics.* [on-line], Bioconductor, c2013, [cit. 26. 10. 2013], <http://www.bioconductor.org/>.

[3] Deducer (2013): *An R Graphical User Interface (GUI) for Everyone*. [on-line], Deducer, c2013, [cit. 26. 10. 2013], <http://www.deducer.org/pmwiki/pmwiki.php?n=Main.DeducerManual>.

[4] Eddelbuettel, D. – Nguyen, K. (2013): *RQuantLib: R interface to the QuantLib library*. R package version 0.3.10. <http://CRAN.R-project.org/package=RQuantLib>.

[5] ESS (2013): *ESS – version 13.09. Emacs Speaks Statistics*. [on-line], ESS Developers, c2013, [cit. 26. 10. 2013], <http://ess.r-project.org/>.

[6] Fiala, J. (2013): *Textový editor PSPad.* [on-line], Slavkov u Brna, Jan Fiala, c2013, [cit. 26. 10. 2013], <http://www.pspad.com/cz/>.

[7] Fox, J. (2005): *The R Commander: A Basic Statistics Graphical User Interface to R*. Journal of Statistical Software, 2005, roč. 14, č. 9, s. 1-42.

[8] Wendling, C. aj. (2013): *Geany*. [on-line], Geany.org, c2013, [cit. 26. 10. 2013], <http://www.geany.org/>.

[9] McTaggart, R. – Daroczi, G. (2013): *Quandl: Quandl Data Connection*. R package version 2.1.2. <http://CRAN.R-project.org/package=Quandl>.

[10] Muenchen, R. A. (2013): *The Popularity of Data Analysis Software*, poslední aktualizace 28. 5. 2013. [on-line], r4stats.com, c2013, [cit. 26. 10. 2013] , <http://r4stats.com/articles/popularity/>.

[11] NppToR (2013): *NppToR: R in Notepad++. NppToR provides R auto-completion and code passing between Notepad++ and R*. [on-line], SourceForge, c2013, [cit. 26. 10. 2013], <http://sourceforge.net/projects/npptor/files/>.

[12] OpenData.CZ (2013): *Iniciativa za otevřenou datovou infrastrukturu*. [on-line], Iniciativa OpenData.cz, c2013, [cit. 26. 10. 2013], <http://www.opendata.cz/>.

[13] OtevrenaData.CZ (2013): *Společně otevíráme data.* [on-line], Praha, Fond Otakara Motejla, c2013, [cit. 26. 10. 2013] <http://www.otevrenadata.cz/>.

[14] Quandl (2013): *Quandl: Find, Use and Share Numerical Data*. [on-line], Toronto, Quandl, c2013, [cit. 26. 10. 2013], <http://www.quandl.com/>.

[15] R Core Team (2013): R: *A Language and Environment for Statistical Computing*. Wien, R Foundation for Statistical Computing, c2013, <http://www.R-project.org/>.

[16] Rattle GUI (2013): *A Graphical User Interface for Data Mining using R.* [on-line], Canberra, Togaware Pty, c2013, [cit. 26. 10. 2013], <http://rattle.togaware.com/>.

[17] R-Forge (2013): *R-Forge.* [on-line], R-Forge Administration and Development Team, c2013, [cit. 26. 10. 2013], <http://r-forge.r-project.org/>.

[18] RKWard (2013): *Welcome to RKWard*. 2. 4. 2013. [on-line], SourceForge, c 2013, [cit. 26. 10. 2013], <http://sourceforge.net/apps/mediawiki/rkward/index.php?title=Main_Page>.

[19] *Rmetrics Association* (2013): [on-line], Zurich, Rmetrics Association, c2013, [cit. 26. 10. 2013], <https://www.rmetrics.org/RmetricsAssociation>.

[20] RStudio (2013): *RStudio IDE*. [on-line], Boston, RStudio, c2013, [cit. 26. 10. 2013], <http://www.rstudio.com/ide/>.

[21] Ryan, J. A. – Ulrich, J. M. (2013): *xts: eXtensible Time Series*. R package version 0.9-7. <http://CRAN.R-project.org/package=xts>.

[22] Ryan, J. A. (2013): *quantmod: Quantitative Financial Modelling Framework*. R package version 0.4-0. [on-line] [cit. 26. 10. 2013], <http://CRAN.R-project.org/package=quantmod>.

[23] *The Comprehensive R Archive Network* (2013): [on-line], c2013, [cit. 26. 10. 2013] , <http://cran.r-project.org/>.

[24] WalWare (2013): *StatET*. [on-line], c2013, [cit. 26. 10. 2013] <http://www.walware.de/?page=/it/statet/>.

[25] Wuertz, D. – Chalabi, Y. – Maechler, M. aj. (2013): *timeDate: Rmetrics – Chronological and Calendar Objects*. R package version 3010.98. <http://CRAN.R-project.org/package=timeDate>.

[26] Wuertz, D. – Chalabi, Y. (2013): *timeSeries: Rmetrics – Financial Time Series Objects*. R package version 3010.97. <http://CRAN.R-project.org/package=timeSeries>.

[27] Wuertz, D. aj. (2012): *fBonds: Bonds and Interest Rate Models*. R package version 2160.76. <http://CRAN.R-project.org/package=fBonds>.

[28] Wuertz, D. aj. (2013a): *fBasics: Rmetrics – Markets and Basic Statistics*. R package version 3010.86. <http://CRAN.R-project.org/package=fBasics>.

[29] Wuertz, D. aj. (2013b): *fImport: Rmetrics – Economic and Financial Data Import*. R package version 3000.82. <http://CRAN.R-project.org/package=fImport>.

[30] Zeileis, A. – Grothendieck, G. (2005): *zoo: S3 Infrastructure for Regular and Irregular Time Series*. Journal of Statistical Software, 2005, roč. 14, č. 6, s. 1-27.

Český finanční a účetní časopis, 2013, roč. 8, č. 4, s. 145-163.

Využití R v oblasti financí

Jiří Sedláček

ABSTRAKT

R je v současnosti jedním z hlavních nástrojů pro statistickou analýzu dat, zejména v akademické oblasti. Nejprve proto zkoumáme hlavní výhody a důvody pro využití R a srovnáváme ho s ostatními (komerčními) programy. Dále se zabýváme výběrem vhodného uživatelského rozhraní pro R (typu GUI nebo typu editor či integrované vývojové prostředí) a srovnáme jednotlivé produkty v každé kategorii. Následuje rozbor hlavních datových struktur, vhodných pro zpracování časových řad, včetně specifických vlastností finančních dat. Volba datové struktury by primárně měla být určována charakterem dat, např. pravidelné versus nepravidelné časové řady, dále zde potřebujeme ukládat i metadata, u finančních dat jsou důležité i další vlastnosti (podpora časových zón ap.). V praxi je však nutné zohlednit i to, které datové struktury jsou podporovány zvolenými rozšiřujícími balíčky. Následuje analýza základních balíčků užitečných pro zpracování finančních dat (quantmod, RQuantLib, kolekce Rmetrics ad.): jakými oblastmi se zabývají, implementované metody a podporované datové struktury. Na závěr srovnáváme různé způsoby získání finančních dat z Internetu a blíže rozebíráme nejvhodnější zdroje včetně jejich přímé podpory v R (zejména web Quandl a balíček Quandl, stručně další balíčky). Charakterizujeme též české snahy o otevřená data.

Klíčová slova: Statistika; Finance; R; Program; Datové struktury; Balíčky; Import dat.

Using R in Finance

ABSTRACT

R is open source software environment (and language) for statistical computing and graphics. Different surveys are showing R's popularity has increased substantially in recent years, especially in academic environment. Therefore, at the beginning main advantages and comparison to commercial statistical software are presented. Second, selection of the best interface for given tasks is important. In each category (GUI or editors/IDEs) several product are compared. Data structure for time series in base installation is suitable for regular time series only. Therefore, several other data structures in different packages are compared: almost all support irregular time series, but differ in other attributes (often important for financial data). In following section, analysis of some well-known packages for financial data (quantmod, RQuantLib, Rmetrics collection and others) are performed. At the beginning of the last section, different ways of downloading data from Internet are shortly presented. Then the relevant sources of financial data are more deeply investigated (in particular web Quandl and corresponding Quandl package for R). Czech projects for open data (still in initial phase) are also shortly described.

Key words: Statistics; Finance; R; Program; Data structures; Packages; Data import.

JEL classification: C10, G10.