PIOTR WDOWIŃSKI

Wstęp do programowania i analizy danych w języku R

WYDAWNICTWO UNIWERSYTETU ŁÓDZKIEGO

PIOTR WDOWIŃSKI

Wstęp do programowania i analizy danych w języku R

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Łódź 2020

Piotr Wdowiński – Uniwersytet Łódzki, Wydział Ekonomiczno-Socjologiczny
Instytut Ekonometrii, Katedra Ekonometrii, 90-214 Łódź, ul. Rewolucji 1905 r. nr 37/39

RECENZENT
Michał Rubaszek

REDAKTOR INICJUJĄCY
Beata Koźniewska

SKŁAD I ŁAMANIE
Piotr Wdowiński

KOREKTA TECHNICZNA
Leonora Gralka

PROJEKT OKŁADKI
Agencja Reklamowa efectoro.pl

Zdjęcie wykorzystane na okładce: © Depositphotos.com/designbyihor@gmail.com

Wydrukowano z gotowych materiałów dostarczonych do Wydawnictwa UŁ

Wydane przez Wydawnictwo Uniwersytetu Łódzkiego
Wydanie I. W.09953.20.0.M

Ark. druk. 11,5

ISBN 978-83-8220-278-6
e-ISBN 978-83-8220-279-3

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
tel. 42 665 58 63

Książkę dedykuję
mojej żonie Magdalenie i dzieciom
– Aleksandrze, Julii i Janowi

Spis treści

Spis rysunków

Przedmowa

Badania i analizy naukowe wiążą się z wyborem oprogramowania obliczeniowego. Jest ono niezbędne przy realizacji zarówno badań symulacyjnych w obszarze teorii statystyki i ekonometrii, jak i badań empirycznych. Ważne jest, aby badacz znał zasady algorytmiki i programowania, które ułatwiają planowanie eksperymentów.

Każdy złożony eksperyment obliczeniowy wymaga zastosowania odpowiedniego programu. Na mniejszą skalę możliwe jest zastosowanie oprogramowania, w którym metody i modele wybiera się z *menu okienkowego*. Jeśli eksperyment jest złożony, np. wymaga utworzenia dużego zbioru danych statystycznych, pochodzących z różnych źródeł, a następnie uporządkowania wyników w tablicach i na wykresach, to niezbędne jest zastosowanie skryptu obliczeniowego zapisanego w wybranym *meta-języku* wyższego poziomu. W językach skryptowych wyższego poziomu do dyspozycji są typowe narzędzia programistyczne, jak: deklaracja stałych, zmiennych, wyrażenia logiczne, pętle i funkcje. Jednak najważniejsze w wymienionych językach skryptowych jest to, że występują w nich również funkcje, które za pomocą jednego polecenia realizują np. wykonanie szacunków modelu ekonometrycznego. Jeśli potrzebna jest daleko idąca modyfikacja np. estymatora, to należy od podstaw zaprogramować etapy obliczeniowe, korzystając z wbudowanych narzędzi rachunku wektorowo-macierzowego. W sytuacji, gdy dany program nie

zawiera odpowiednich metod wbudowanych, można z niego przekazać dane do innego programu, a następnie pobrać wyniki i dalej je wykorzystać.

Można przyjąć, że każdy eksperyment obliczeniowy wymaga sporządzenia raportu z badań. Wydaje się to proste. Wystarczy sporządzić tablice oraz rysunki i zamieścić je w dokumencie *Word*. Problem komplikuje się, jeśli proces tworzenia raportu ma być powtarzalny, próba statystyczna jest często aktualizowana lub na ostateczną decyzję o wyborze modelu składa się wiele podobnych prób obliczeniowych. Jeśli dodatkowo raport ma obejmować wiele formatów – `HTML`, `PDF`, `DOCX` – to niezbędne jest wielofunkcyjne narzędzie, nie tylko obliczeniowe. Program `R` łączy wszystkie wspomniane cechy, a ponadto można w nim wykonać aplikację internetową, zasilaną danymi i wykonującą obliczenia w czasie rzeczywistym. Do tego służy serwer `R/Shiny`. Możliwość skorzystania z wielu funkcji wymaga od badacza znajomości zaawansowanego aparatu pojęciowego i narzędziowego obejmującego:

- znajomość zasad algorytmiki i programowania komputerów,
- znajomość języków `LaTeX`, `HTML` i `JavaScript` oraz arkuszy stylów `CSS` (*kaskadowych arkuszy stylów*),
- znajomość organizacji baz danych `SQL`,
- umiejętność pracy w środowisku `Linux`,
- znajomość organizacji pracy serwerów:
  - zdalnych `VPS` (*Virtual Private Server*),
  - `Apache` do obsługi stron internetowych,
  - baz danych `MySQL` i programów towarzyszących, np. `phpMyAdmin`,
- ogólną znajomość systemów `CMS` do budowy stron internetowych: `Joomla` i `Wordpress`,
- umiejętność pracy z domenami internetowymi, programami administracyjnymi dla serwerów `Apache/PHP/HTML`, np. `DirectAdmin`,
- znajomość konstrukcji rekordów `DNS` do zaawansowanej obsługi domen internetowych.

Jak widać, lista umiejętności jest długa i wymaga specjalistycznej wiedzy. Część zagadnień zostanie poruszona w tej książce. Zapraszam do lektury. Będę wdzięczny za wszelkie uwagi, które pozwolą na przygotowanie poprawionych i uzupełnionych wydań. Książka ukazuje się w kilku wersjach: papierowej i internetowej w postaci strony `HTML` oraz pliku `PDF`. W przyszłości planuje się wydanie książki w formatach `EPUB` i `MOBI` dla czytników. Mam nadzieję, że dzięki takiej formule wydania trafi ona do wielu Czytelników.

Chciałbym podziękować *Narodowemu Bankowi Polskiemu* za zainteresowanie mnie programowaniem w języku `R` oraz sfinansowanie zaawansowanych szkoleń. Składam serdeczne podziękowania Recenzentowi, dr. hab. Michałowi Rubaszkowi, prof. SGH, za wnikliwe uwagi krytyczne, które pozwoliły nadać książce ostateczny kształt. Dziękuję również mojej rodzinie za udzielone wsparcie podczas pracy nad książką.

Piotr Wdowiński
Uniwersytet Łódzki
piotr.wdowinski@uni.lodz.pl

Rozdział 1

Wprowadzenie

Program R jest programem obliczeniowym wykorzystywanym w wielu dziedzinach nauki[1]. Jego zastosowanie jest bardzo szerokie. Program R wykorzystuje się m.in. w badaniach i analizach dotyczących:

- statystyki i ekonometrii[2],

[1] R Core Team, *R: A Language and Environment for Statistical Computing*, (2014), http://www.R-project.org/.

[2] P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*, Wydawnictwo Naukowe PWN, Warszawa 2013; Y. Cohen, J.Y. Cohen, *Statistics and Data with R: An Applied Approach Through Examples*, Wiley 2008; Y. Croissant, G. Millo, *Panel Data Econometrics with R*, Wiley 2018; J.D. Cryer, K.-S. Chan, *Time Series Analysis: With Applications in R*, Springer-Verlag, New York 2008; E. Gatnar, M. Walesiak (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009; P. Gerrard, R.M. Johnson, *Mastering Scientific Computing with R*, Packt Publishing 2015; R. Kaas i in., *Modern Actuarial Risk Theory: Using R*, Springer-Verlag, Berlin Heidelberg 2008; B. Kamiński, M. Zawisza, *Receptury w R: Podręcznik dla ekonomistów*, Oficyna Wydawnicza SGH, Warszawa 2012; C. Kleiber, A. Zeileis, *Applied Econometrics with R*, New York 2008; K. Kopczewska, T. Kopczewski, P. Wójcik, *Metody ilościowe w R. Aplikacje ekonomiczne i finansowe (wyd. II)*, CeDeWu.pl 2016; B. Makhabel, *Learning Data Mining with R*, Packt Publishing 2015; J.-M. Marin, C. Robert, *Bayesian Essentials with R*, Springer-Verlag, New York 2014; L. Pace, *Beginning R: An Introduction to Statistical Programming*, Apress 2012; C. Robert, G. Casella, *Introducing Monte Carlo Methods with R*, Springer-Verlag, New York

- algorytmiki i programowania[3],
- medycyny i biologii[4],
- przetwarzania dużych zbiorów danych[5],
- finansów[6],
- lingwistyki[7].

R jest programem Open Source (otwartym), opartym na licencji *GNU General Public License* i dlatego jego zastosowanie jest powszechne. Program R stanowi język skryptowy, który pozwala na korzystanie

2010; M. Rubaszek, *Modelowanie polskiej gospodarki z pakietem R*, Oficyna Wydawnicza SGH, Warszawa 2012; R. Schumacker, S. Tomek, *Understanding Statistics Using R*, Springer-Verlag, New York 2013; E.A. Suess, B.E. Trumbo, *Introduction to Probability Simulation and Gibbs Sampling with R*, Springer-Verlag, New York 2010; R.S. Tsay, *An Introduction to Analysis of Financial Data with R*, Wiley 2012; M.D. Ugarte, A.F. Militino, A.T. Arnholt, *Probability and Statistics with R*, Chapman & Hall/CRC 2015; J. Verzani, *Using R for Introductory Statistics*, Chapman & Hall/CRC 2014.

[3] J.M. Chambers, *Software for Data Analysis: Programming with R*, Springer-Verlag, New York 2008; M.J. Crawley, *The R Book, 2nd Edition*, Wiley 2012; M. Gągolewski, *Programowanie w języku R: analiza danych, obliczenia, symulacje*, Wydawnictwo Naukowe PWN, Warszawa 2016; C. Gillespie, R. Lovelace, *Efficient R Programming: A Practical Guide to Smarter Programming*, O'Reilly Media 2017; G. Grolemund, *Hands-On Programming with R: Write Your Own Functions And Simulations*, O'Reilly Media 2014; T. Mailund, *Advanced Object-Oriented Programming in R: Statistical Programming for Data Science, Analysis and Finance*, Apress 2017; K. Ren, *Learning R Programming*, Packt Publishing 2016; P. Spector, *Data Manipulation with R*, Springer-Verlag, New York 2008; P. Teetor, *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*, O'Reilly Media 2011; H. Wickham, G. Grolemund, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly Media 2016; G.M. Wiley, J.F. Wiley, *Advanced R: Data Programming and the Cloud*, Apress 2016.

[4] P.D. Lewis, *R for Medicine and Biology*, Jones & Bartlett Learning 2010.

[5] S. Munzert i in., *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*, Wiley 2015.

[6] A. Arratia, *Computational Finance: An Introductory Course with R*, Atlantis Press 2014; E. Berlinger, F. Illés, i in., *Mastering R for Quantitative Finance*, Packt Publishing 2015; R. Kaas i in., *Modern Actuarial Risk Theory: Using R*...

[7] S.T. Gries, *Statistics for Linguistics with R: A Practical Introduction*, Mouton de Gruyter 2009.

z wbudowanych i zewnętrznych pakietów, tworzenie własnych funkcji i pakietów, osadzanie w kodzie funkcji z innych pakietów oraz rozbudowę istniejących pakietów. Program R pozwala również na wizualizację wyników badań w postaci graficznej a nawet na budowę interaktywnych stron internetowych w kodzie HTML. Język R jest oparty na języku S, który został stworzony przez Johna Chambersa i innych[8]. Produkt powstał w innowacyjnej organizacji Bell Laboratories (wcześniej w ramach AT&T, obecnie Lucent Technologies). W ramach Bell Laboratories powstał również system operacyjny Unix oraz języki programowania C i C++.

Jak wspomniano, program R jest oparty na idei otwartego oprogramowania Open Source w ramach licencji publicznej *GNU General Public License*. Licencja ta otwiera możliwość tworzenia i modyfikacji pakietów obliczeniowych dla programu R. Wszelkie modyfikacje muszą uznawać kod źródłowy za rozwiązanie pierwotne. Możliwość samodzielnej analizy kodu źródłowego programu R i jego pakietów ma dla użytkowników walory dydaktyczne i naukowe. Ma również walory dyscyplinujące dla samych autorów pakietów, gdyż ich rozwiązania są poddane powszechnej weryfikacji. Prowadzone przez społeczność R statystyki popularności pakietów nakładają na autorów jeszcze większą dyscyplinę. Wiele pakietów w postaci kodu źródłowego znajduje się na platformie GitHub (https://github.com/). Globalny dostęp do programu R, pakietów i dyskusji w Internecie stwarza unikalne warunki dla upowszechniania nauki, w tym w krajach o niskim dochodzie, gdzie dostęp do programów komercyjnych jest znacząco utrudniony. Dzięki dyskusji w Internecie dotyczącej szczegółowych rozwiązań algorytmicznych dla programu R na platformach takich jak Stack Overflow (https://stackoverflow.com/), użytkownicy z całego świata mogą zgłaszać problemy i prawie natychmiast je rozwiązywać. Taka sytuacja stwarza bodźce do efektywnej alokacji wiedzy i umiejętności.

[8] R.A. Becker, J.M. Chambers, A.R. Wilks, *The S Language – A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole, Pacific Grove, CA 1988; J.M. Chambers, *Programming with Data – A Guide to the S Language*, Springer-Verlag, New York 1998; J.M. Chambers, T.J. Hastie, *Statistical Models in S*, Chapman & Hall/CRC 1991.

Rozwojem programu R zajmuje się grupa statystyków, skupiona wokół projektu *R Development Core Team* i fundacja *The R Foundation* (https://www.r-project.org/foundation/) założona przez członków wspomnianego zespołu. Zewnętrzne pakiety są tworzone przez ogromną grupę entuzjastów z całego świata. Wśród nich są naukowcy o znaczącym dorobku, pracujący w najlepszych uniwersytetach. Wszystkie ścieżki dla programu R prowadzą do zasobów zgromadzonych na stronie internetowej *The R Project for Statistical Computing* – https://www.r-project.org/. Na tej stronie znajdują się:

- program R,
- baza pakietów,
- dokumentacja,
- materiały dodatkowe.

Bardzo istotnym ogniwem systemu programu R jest oprogramowanie towarzyszące. Zaliczają się do niego cztery systemy:

- RStudio,
- RStudio Server,
- Shiny Server,
- shinyapps.io.

Program RStudio jest interfejsem graficznym (GUI – *graphical user interface*) ułatwiającym obsługę programu R. RStudio zawiera kilka obszarów roboczych obejmujących:

- pozycje menu,
- kod skryptu,
- konsolę wyników,
- środowisko zmiennych,
- obszar prezentacji grafiki.

W programie RStudio możliwa jest kompilacja tekstu i wyników algorytmów obliczeniowych do wielu formatów dokumentów, w tym: HTML, DOCX i PDF. System R Markdown, rozwijany w projekcie RStudio, daje możliwość tworzenia artykułów, prezentacji i innych form

tekstowych uwzględniających obliczenia w programie R. Jest to szczególnie ważny element w pracy naukowej, gdyż kod skryptowy programu R można umieścić bezpośrednio w tekście artykułu (chunks) i wykonać go podczas kompilacji tekstu do postaci wynikowej. W ten sposób można przygotowywać artykuły naukowe, a nawet książki. Niniejsza książka powstała w programach R i RStudio w oparciu o pakiet bookdown (https://bookdown.org/) autorstwa Y. Xie[9].

Program RStudio Server jest, podobnie jak RStudio, interfejsem graficznym służącym do pracy zdalnej w programie R. Oznacza to, że RStudio Server jest instalowany na serwerze w środowisku Linux, na którym również jest zainstalowany program R. W ten sposób można pracować w programie R z dostępem zdalnym, podobnie jak w oparciu o program RStudio instalowany lokalnie na komputerze. Praca zdalna daje ogromne możliwości współpracy zespołów z różnych części świata, gdyż skrypt obliczeniowy i wspólny artykuł są umieszczone na serwerze i na nim odbywa się praca całego zespołu. Możliwości RStudio Server są wykorzystywane również do celów dydaktycznych i szkoleniowych, gdyż studenci uzyskują dostęp poprzez Internet do jednego serwera z wielu terminali komputerowych. Niepotrzebne są wówczas lokalne instalacje programów R i RStudio. RStudio Server obejmuje dwie wersje: Open Source i komercyjną.

Program Shiny Server, rozwijany w projekcie RStudio, jest serwerem języka HTML opartym na pakiecie shiny dla programu R[10]. Program ten instaluje się w środowisku Linux po pobraniu z lokalizacji:

```
https://www.rstudio.com/products/shiny/download-server/
```

[9] Y. Xie, *Dynamic Documents with R and knitr*, Chapman & Hall/CRC 2015; Y. Xie, *Bookdown: Authoring Books and Technical Documents with R Markdown*, Chapman & Hall/CRC 2016; Y. Xie, J.J. Allaire, G. Grolemund, *R Markdown: The Definitive Guide*, Chapman & Hall/CRC 2018.

[10] W. Chang i in., *Shiny: Web Application Framework for R*, (2018), https://CRAN.R-project.org/package=shiny.

Pozwala on na budowanie interaktywnych stron internetowych. Odbywa się to poprzez umieszczenie w kodzie algorytmu strony internetowej, zapisanego za pomocą poleceń pakietu `shiny`, standardowego kodu obliczeniowego `R`. W ten sposób można zbudować stronę internetową z zamieszczonym algorytmem obliczeniowym, powstałym w programie `R` i przekazywać parametry do kodu za pomocą standardowych elementów formularzy stron internetowych. W ten sposób za pomocą `R`, `RStudio` i `Shiny Server` można tworzyć bardzo rozbudowane systemy eksperckie.

Strona `shinyapps.io` zawiera projekty utworzone za pomocą języka pakietu `shiny` w programie `R`. W systemie tym można utworzyć różne rodzaje kont i umieścić na serwerze kod stron aplikacji `shiny`. Zatem strona `shinyapps.io` jest miejscem, gdzie zainstalowany jest `Shiny Server`. Jak powiedziano wcześniej, `Shiny Server` można zainstalować samodzielnie w systemie `Linux` na dowolnym serwerze zdalnym, np. VPS (*Virtual Private Server*), udostępnianym, zwykle komercyjnie, przez wielu dostawców infrastruktury internetowej.

W następnym rozdziale zostaną przedstawione szczegóły instalacji programów:

- `R` wraz z pakietami,
- `RStudio`.

Instalacja będzie dotyczyć środowiska systemu `Windows`.