

Marek Gągolewski

Programowanie w języku R

Analiza danych,
obliczenia,
symulacje



Projekt okładki i stron tytułowych **Agnieszka Łydzba**

Ilustracja na okładce **Itana/Shutterstock**

Wydawca **Bianka Piwowarczyk-Kowalewska**

Produkcja **Anna Bączkowska**

Łamanie **FixPoint**

Książka, którą nabyłeś, jest dziełem twórcy i wydawcy. Prosimy, abyś przestrzegał praw, jakie im przysługują. Jej zawartość możesz udostępnić nieodpłatnie osobom bliskim lub osobiście znanym. Ale nie publikuj jej w internecie. Jeśli cytujesz jej fragmenty, nie zmieniaj ich treści i koniecznie zaznacz, czyje to dzieło. A kopiując jej część, rób to jedynie na użytek osobisty.

Szanujemy cudzą własność i prawo

Więcej na www.legalnakultura.pl

Polska Izba Książki

Copyright © by Wydawnictwo Naukowe PWN SA
Warszawa 2014

ISBN 978-83-01-17461-3

Wydanie pierwsze

Wydawnictwo Naukowe PWN SA
tel. 22 69 54 321, faks 22 69 54 288
infolinia 801 33 33 88
e-mail: pwn@pwn.com.pl, www.pwn.pl

Druk i oprawa: Totem, Inowrocław

Spis treści

Przedmowa	XIII
Podstawy R	
1. Środowisko R i program RStudio	3
1.1. Cechy języka R	3
1.2. Organizacja pracy w R i RStudio	5
1.2.1. Konsola R	5
1.2.2. Program RStudio	6
1.2.3. Pierwsze kroki w trybie interaktywnym	8
1.2.4. Edytor skryptów	10
1.2.5. System pomocy	12
2. Podstawowe typy atomowe: wektory i NULL	14
2.1. Podstawowe i złożone typy danych w R	14
2.2. Wektory atomowe	15
2.2.1. Wektory wartości logicznych	15
2.2.2. Wektory liczbowe i zespolone	18
2.2.3. Wektory napisów	22
2.2.4. Hierarchia typów dla wektorów atomowych	23
2.3. Tworzenie obiektów nazwanych	25
2.4. Braki danych, wartości nieskończone i nie-liczby	29
2.5. Typ pusty (NULL)	31
3. Operacje na wektorach	34
3.1. Podstawowe operatory	34
3.1.1. Operatory arytmetyczne	35
3.1.2. Operatory logiczne	38
3.1.3. Operatory relacyjne	40
3.1.4. Priorytety operatorów	41

3.2.	Indeksowanie wektorów. Filtrowanie danych	43
3.2.1.	Rodzaje indeksatorów	43
3.2.2.	Modyfikowanie wybranych elementów	46
3.3.	Przegląd funkcji wbudowanych	47
3.3.1.	Zwektoryzowane funkcje matematyczne	47
3.3.2.	Agregacja danych	51
3.3.3.	Operacje na sąsiadujących elementach wektorów	54
3.3.4.	Wyszukiwanie indeksów elementów wektorów	54
3.3.5.	Operacje oparte na permutowaniu elementów wektora	55
3.3.6.	Operacje na zbiorach	57
3.3.7.	Podstawowe operacje na napisach	58
4.	Listy	61
4.1.	Tworzenie list	62
4.2.	Indeksowanie list	64
4.2.1.	Operator „[”	64
4.2.2.	Operator „[[”	65
4.2.3.	Modyfikowanie zawartości list	66
4.3.	Wybrane operacje	68
4.3.1.	Łączenie, rozwijanie i powielanie list	68
4.3.2.	Wywoływanie funkcji na wszystkich elementach listy	70
5.	Funkcje	72
5.1.	Tworzenie obiektów typu funkcja	73
5.1.1.	Bloki wyrażeń	75
5.1.2.	Sprawdzanie poprawności argumentów	78
5.1.3.	Kilka uwag dla projektantów funkcji	80
5.1.4.	Biblioteki funkcji w plikach .R	82
5.1.5.	Odwoływanie się do funkcji z pakietów R	82
5.2.	Zasięg nazw w funkcjach	84
5.3.	Parametry i argumenty	86
5.3.1.	Przekazywanie argumentów przez wartość	86
5.3.2.	Parametry z argumentami domyślnymi	87
5.3.3.	Leniwa ewaluacja	88
5.3.4.	Parametr specjalny „...”	90
6.	Modyfikacja przepływu sterowania	92
6.1.	Wyrażenia warunkowe if i if...else	93
6.1.1.	Określanie testowanego warunku	96
6.1.2.	Wartości zwracane przez wyrażenia warunkowe	98
6.2.	Pętle	101
6.2.1.	Pętla while	101
6.2.2.	Pętla repeat	106
6.2.3.	Pętla for	107
6.3.	Uwagi na temat wydajności	110
6.3.1.	Wydajność pętli w R	111
6.3.2.	Złożoność czasowa algorytmów	116

7. Atrybuty obiektów	120
7.1. Nadawanie i odczytywanie atrybutów	120
7.2. Atrybuty specjalne	122
7.2.1. Atrybut <i>comment</i>	123
7.2.2. Atrybut <i>names</i>	123
7.2.3. Atrybut <i>class</i> . Wstęp do programowania obiektowego w S3	128
7.3. O zachowywaniu i gubieniu atrybutów przez funkcje	133
8. Typy złożone	136
8.1. Macierze i tablice	136
8.1.1. Tworzenie macierzy	136
8.1.2. Indeksowanie macierzy	140
8.1.3. Wybrane operacje na macierzach	141
8.1.4. Tablice jako uogólnienie macierzy	143
8.1.5. Atrybut <i>dimnames</i>	145
8.2. Szeregi czasowe	146
8.3. Czynniki	147
8.3.1. Tworzenie czynników	147
8.3.2. Dlaczego czasem warto stosować czynniki?	149
8.3.3. Wybrane operacje na czynnikach	150
8.4. Ramki danych	152
8.4.1. Operatory indeksowania. Filtrowanie danych	156
8.4.2. Wybrane operacje na ramkach danych	158

Przygotowanie danych

9. Przetwarzanie napisów	165
9.1. Standardy kodowania znaków	165
9.1.1. Kodowanie ASCII	165
9.1.2. 8-bitowe kodowania polskich znaków diakrytycznych	168
9.1.3. Kodowanie UTF-8	169
9.1.4. Konwersja kodowań	171
9.2. Podstawowe operacje na napisach	175
9.2.1. Wyznaczanie długości napisów	176
9.2.2. Porównywanie napisów	176
9.2.3. Łączenie i powielanie napisów	176
9.2.4. Białe znaki, znaki specjalne i wypełnianie	177
9.2.5. Formowanie napisów na podstawie innych obiektów języka R	179
9.2.6. Zmiana pojedynczych znaków	181
9.2.7. Wyznaczanie podnapisów	182
9.3. Wyszukiwanie wzorca	183
9.3.1. Wzorce ustalone	183
9.3.2. Wyrażenia regularne	186
9.3.3. Dopasowywanie dokładne i częściowe: jedna opcja z wielu	197
9.4. Data i czas	198
9.4.1. Reprezentacja dat	198
9.4.2. Reprezentacja czasu	199
9.4.3. Operacje arytmetyczne	201
9.4.4. Konwersja daty i czasu	202

10. Przetwarzanie plików	204
10.1. Podstawowe operacje na plikach i katalogach	204
10.1.1. Ścieżki dostępu do plików i katalogów	204
10.1.2. Aktualny katalog roboczy. Ścieżki względne	208
10.1.3. Informacje o plikach i katalogach	209
10.1.4. Wybrane działania na plikach i katalogach	211
10.1.5. Wyszukiwanie plików i katalogów	212
10.2. Serializacja i deserializacja obiektów	214
10.3. Dane tabelaryczne	215
10.3.1. Ładowanie danych tabelarycznych	215
10.3.2. Zapisywanie danych tabelarycznych	218
10.4. Pliki tekstowe	220
10.4.1. Odczyt plików tekstowych	222
10.4.2. Zapis plików tekstowych	223
10.5. Połączenia	224
10.5.1. URL, czyli ujednolicone lokalizacje zasobów	224
10.5.2. Tworzenie połączeń	226
10.5.3. Odczyt z i zapis do połączeń	227
10.5.4. Zarządzanie otwartymi połączeniami	232
10.5.5. Nota o plikach binarnych	233
Wizualizacja wyników	
11. Niskopoziomowe operacje graficzne	237
11.1. Grafika: po co, kiedy i jak?	237
11.2. Podstawy użycia pakietu <code>graphics</code>	239
11.2.1. Strona i rysunki	239
11.2.2. Parametry graficzne	241
11.2.3. Rysowanie punktów i odcinków (łamanych)	245
11.2.4. Barwy	247
11.2.5. Rysowanie wielokątów	251
11.2.6. Wypisywanie tekstu	252
11.2.7. Układ współrzędnych	254
11.2.8. Tworzenie wielu rysunków na jednej stronie	256
11.3. Urządzenia graficzne	257
11.3.1. Urządzenia <code>pdf()</code> i <code>postscript()</code>	259
11.3.2. Urządzenia <code>png()</code> i <code>jpeg()</code>	261
11.3.3. Urządzenie <code>tikz()</code>	261
12. Wysokopoziomowe operacje graficzne	265
12.1. Rysowanie układu współrzędnych	265
12.2. Adnotacje i legenda	266
12.3. Wizualizacja danych jednowymiarowych	268
12.4. Wizualizacja danych dwuwymiarowych	271
12.5. Wizualizacja danych wielowymiarowych	274
12.6. Dodatek: przetwarzanie obrazów rastrowych	276
13. Generowanie raportów przy użyciu pakietu <code>knitr</code>	284
13.1. Języki znaczników	284

13.1.1. Język HTML5	285
13.1.2. Język \TeX i pakiet \LaTeX	286
13.1.3. Podstawy HTML5 i \LaTeX -a	287
13.2. Pakiet knitr	295
13.2.1. knitr i HTML5	297
13.2.2. knitr i \LaTeX	298
13.3. Dostosowywanie ustawień pakietu knitr i wstawek	299
13.3.1. Ustawienia wstawek	299
13.3.2. Ustawienia pakietu	303

Zastosowania R

14. Metody numeryczne i obliczenia naukowe	307
14.1. Wprowadzenie	307
14.1.1. Matematyczny formalizm a metody numeryczne	308
14.1.2. Własności arytmetyki zmiennopozycyjnej	308
14.2. Algebra wektorów i macierzy	316
14.2.1. Normy i metryki	317
14.2.2. Wektory i wartości własne	320
14.2.3. Rozkład Choleskiego	322
14.2.4. Rozkład QR	323
14.2.5. Rozkład SVD	324
14.3. Różniczkowanie i całkowanie	325
14.3.1. Różniczkowanie numeryczne	325
14.3.2. Całkowanie numeryczne	326
14.4. Optymalizacja	328
14.4.1. Algorytmy programowania matematycznego	330
14.4.2. Algorytmy optymalizacji ciągłej ogólnego zastosowania	331
14.5. Interpolacja, wygładzanie i aproksymacja	334
14.5.1. Interpolacja jednowymiarowa	334
14.5.2. Interpolacja dwuwymiarowa	336
14.5.3. Wygładzanie	337
14.5.4. Aproksymacja	339
14.6. Rozwiązywanie (układów) równań (nie)liniowych	340
14.6.1. Wyznaczanie miejsc zerowych funkcji jednej zmiennej	340
14.6.2. Rozwiązywanie układów równań liniowych	342
14.6.3. Rozwiązywanie układów równań nieliniowych	342
15. Symulacje i wnioskowanie statystyczne	344
15.1. Generowanie liczb (pseudo)losowych	344
15.1.1. Źródła (pseudo)losowości	344
15.1.2. Określanie ziarna generatora	346
15.1.3. Szczegóły działania generatora	347
15.2. Rozkłady prawdopodobieństwa w R	350
15.2.1. Przedrostki nazw funkcji związanych z rozkładami	351
15.2.2. Wybrane jednowymiarowe rozkłady prawdopodobieństwa	352
15.2.3. Zmienne losowe wielowymiarowe	360
15.3. Wnioskowanie statystyczne	362
15.3.1. Estymacja punktowa	362

15.3.2. Testowanie hipotez i estymacja przedziałowa	366
15.4. Przykładowe eksperymenty losowe	369
15.4.1. Własności estymatorów parametrów rozkładu Gamma	369
15.4.2. Badanie mocy testu Shapiro–Wilka	373
15.4.3. Testowanie generatora liczb z rozkładu Pareto	375
15.4.4. Całkowanie Monte Carlo	376
15.4.5. Analiza gry <i>Memo</i>	378

Zagadnienia zaawansowane

16. Zarządzanie środowiskiem R	385
16.1. Podstawowe informacje	385
16.1.1. Informacje o R	385
16.1.2. Informacje o systemie	386
16.1.3. Uruchamianie i zamykanie środowiska R	387
16.1.4. Historia poleceń	388
16.2. Opcje globalne	388
16.3. Ustawienia lokalizacyjne	391
16.4. Rozszerzanie możliwości R	393
16.4.1. Instalacja i aktualizacja pakietów	393
16.4.2. Wywoływanie innych programów	397
16.5. Zarządzanie pamięcią	398
16.5.1. Informacja o rozmiarze obiektów	399
16.5.2. Kopiowanie na żądanie	400
16.5.3. Automatyczne odświeżanie pamięci	401
16.6. Typ podstawowy, tryb a klasa obiektów	402
17. Środowiska	405
17.1. Środowiska jako zbiory obiektów	406
17.1.1. Podstawowe operacje na obiektach w środowisku	406
17.1.2. Środowiska a listy	408
17.1.3. Przekazywanie środowisk funkcjom	410
17.2. Wskaźniki na środowiska otaczające	412
17.2.1. Przypadek ręcznie tworzonych środowisk	412
17.2.2. Ścieżka wyszukiwania	413
17.2.3. Przestrzenie nazw i środowiska eksportowane przez pakiety	418
18. Syntaktyka i semantyka języka R	421
18.1. Obiekty reprezentujące wyrażenia języka R	421
18.1.1. Parser	422
18.1.2. Cytowanie	425
18.1.3. Wywołania, czyli wyrażenia złożone	426
18.2. Środowiskowy model obliczeń	430
18.2.1. Ewaluacja wyrażeń	430
18.2.2. Bieżące środowisko ewaluacyjne	432
18.3. Ewaluacja wyrażeń w obrębie funkcji	433
18.3.1. Lokalne środowiska ewaluacyjne	434
18.3.2. Środowiska otaczające lokalne środowiska ewaluacyjne	436
18.3.3. Środowiska wywołujące	441

18.3.4. Ewaluacja argumentów	442
18.4. Formuły	445
18.4.1. Przykłady funkcji stosujących argumenty typu formuła	445
18.4.2. Formuły jako wywołania	447
19. Pielęgnowanie kodu	450
19.1. Obsługa wyjątków	451
19.1.1. Błędy i inne wyjątki	451
19.1.2. Obsługa ostrzeżeń i komunikatów diagnostycznych	452
19.1.3. Obsługa błędów	453
19.2. Testowanie oprogramowania	455
19.3. Debugowanie kodu	456
19.4. Pomiar wydajności kodu	458
19.4.1. Badanie krótkich fragmentów kodu	458
19.4.2. Profilowanie aplikacji	459
20. Programowanie zorientowane obiektowo	463
20.1. Paradygmaty programowania obiektowego a R	463
20.2. Klasy S3	466
20.2.1. Określanie klasy obiektu	466
20.2.2. Ekspediowanie metod	467
20.2.3. Przeciążanie metod	471
20.3. Klasy S4	473
20.3.1. Definiowanie klas i tworzenie obiektów	475
20.3.2. Definiowanie funkcji generycznych i metod	478
20.3.3. Klasy S4 a klasy S3	482
20.3.4. Klasy referencyjne	483
20.4. Specjalne rodzaje funkcji	483
20.4.1. Funkcje podstawieniowe	483
20.4.2. Przeciążanie operatorów	485
20.4.3. Wbudowane grupy funkcji generycznych	486
21. Co dalej?	489
Bibliografia	492
Skorowidz	495