Rozdział 1

Wstęp

O czym jest ta książka?

Nasza książka jest nowoczesnym podręcznikiem wprowadzającym do analizy i prognozowania szeregów czasowych, który przedstawia najważniejsze metody i modele z punktu widzenia zastosowań.

Opieramy się na darmowym systemie **R** (`http://www.r-project.org`), który jest standardem współczesnej statystyki oraz powszechnie stosowanym narzędziem praktycznej analizy danych.

Czytelnik pozna wszystkie etapy analizy szeregów czasowych, począwszy od graficznej prezentacji danych, niezbędnych przekształceń wstępnych, poprzez identyfikację tendencji długoterminowych i sezonowych, dopasowanie i diagnostykę modeli, a kończąc na konstrukcji prognoz i ocenie ich dokładności. W podręczniku w przystępny sposób przedstawiono podstawy i praktyczne aspekty tych zagadnień. Książka zawiera wiele przykładów opartych na rzeczywistych szeregach czasowych z różnych obszarów zastosowań. Można w niej też znaleźć informację o tym, jak nauczyć się korzystać z systemu **R** oraz szczegółowy opis ważnych funkcji i bibliotek związanych z analizą szeregów czasowych. Zamieściliśmy również fragmenty kodów pozwalających na wykonanie opisywanych analiz.

Staraliśmy się, aby książka nie była wyłącznie przeglądem metodologii analizy szeregów czasowych, który można znaleźć w klasycznych podręcznikach. Chcemy, żeby odpowiadała na najważniejsze pytania praktyków. Czytelnik dowiaduje się więc między innymi, jak odpowiednio przygotować dane do analizy, jak wybrać optymalny model czy metodę dla określonych danych oraz w jaki sposób ocenić i porównać wiarygodność skonstruowanych prognoz. Dużą uwagę poświęcamy prawidłowej interpretacji wyników przeprowadzanych analiz.

„Analiza i prognozowanie...” jest uniwersalnym podręcznikiem. Nie ograniczamy się do jednego obszaru zastosowań. Nie zamieszczamy również *case studies*, prezentujących rozwiązanie jedynie specyficznych problemów bizne-

sowych. W książce można jednak znaleźć przykłady zastosowań określonych metod analizy szeregów czasowych dla danych makroekonomicznych, finansowych, demograficznych i innych. Pozostawiamy zatem Czytelnikowi swobodę w doborze przedstawionych narzędzi do rozwiązywania określonych zagadnień, z którymi spotyka się w swojej pracy zawodowej. Mamy jednocześnie nadzieję, że przedstawione w podręczniku podstawy metodologiczne analizy szeregów czasowych i wskazówki praktyczne ułatwią wybór właściwych metod i modeli oraz pomogą poprawnie zinterpretować uzyskane wyniki.

Dla kogo jest ta książka i jak może być wykorzystywana?

Książka jest przeznaczona dla wszystkich zainteresowanych poznaniem praktyki analizy i prognozowania szeregów czasowych.

Będzie ona przydatna dla praktyków podejmujących ważne decyzje biznesowe na podstawie analizy wielkości zależnych od czasu. Podręcznikiem mogą być zainteresowane m.in. osoby pracujące w departamentach analiz ekonomicznych, controllingu, sprzedaży, marketingu i innych prognozujących zachowanie szeregów czasowych związanych z gospodarką, ekonomią, produkcją przemysłową, rynkiem energii czy sprzedażą.

Książka może też pomóc osobom prowadzącym badania naukowe w dziedzinie ekonomii, demografii, socjologii oraz nauk przyrodniczych, w których analizuje się szeregi czasowe opisujące dynamikę różnych zjawisk. „Analiza i prognozowanie...” może być wykorzystana również jako podręcznik dla studentów kierunków matematycznych, ekonomicznych, informatycznych, zarządzania i marketingu oraz wybranych kierunków humanistycznych.

Z podręcznika mogą korzystać zarówno osoby nieposiadające jeszcze żadnego doświadczenia w zakresie analizy szeregów czasowych, jak i osoby bardziej zaawansowane, które będą mogły uzupełnić i usystematyzować swoją wiedzę. Bardziej doświadczeni Czytelnicy mogą wykorzystywać podręcznik jako dokumentację pomagającą w analizowaniu szeregów czasowych w środowisku **R**, do której zagląda się, aby wyszukać potrzebną funkcję i poznać przykłady jej użycia.

Korzystanie z podręcznika nie wymaga od Czytelnika znajomości statystyki, rachunku prawdopodobieństwa czy modelowania matematycznego. Podstawowa wiedza w tym zakresie pomoże jednak głębiej zrozumieć bardziej zaawansowane zagadnienia. Jest to możliwe również dzięki temu, że nie unikamy podawania wzorów i przedstawiania precyzyjnych wyjaśnień oraz opisów omawianych metod i modeli.

Co ważne, po każdym rozdziale znajduje się seria ćwiczeń do samodzielnego wykonania. Ułatwia to zdobycie praktycznych umiejętności.

Oprogramowanie – pakiet R

Wprowadzenie do analizy i prognozowania szeregów czasowych oparto na przykładach przygotowanych dla środowiska **R**. O wyborze **R** zdecydowały głównie olbrzymie możliwości tego środowiska w zakresie analizy danych (w tym szeregów czasowych), bogaty zestaw narzędzi graficznych oraz jego ogromna i wciąż rosnąca popularność wśród praktyków. Dodatkowo, pakiet **R** jest darmowy do wszelkich zastosowań, w tym komercyjnych.

Od Czytelnika nie wymaga się znajomości pakietu **R** przed rozpoczęciem korzystania z książki. W dodatku do podręcznika znajduje się krótki rozdział „Jak nauczyć się **R**?", zawierający najważniejsze wskazówki i zalecenia praktyczne, które ułatwią Czytelnikowi rozpoczęcie pracy z **R** i przyspieszą poznanie możliwości tego środowiska.

Do przykładów prezentowanych w podręczniku używana była wersja **R** 3.2.0. Dla chcących powtórzyć analizy przedstawione w książce, ważne jest, jakie wersje pakietów **R** były wykorzystane. Poniżej podajemy informacje o pakietach dostarczane przez funkcję `sessionInfo()`.

```
> sessionInfo()
R version 3.2.0 (2015-04-16)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 7 x64 (build 7601) Service Pack 1

locale:
[1] LC_COLLATE=Polish_Poland.1250
[2] LC_CTYPE=Polish_Poland.1250
[3] LC_MONETARY=Polish_Poland.1250
[4] LC_NUMERIC=C
[5] LC_TIME=Polish_Poland.1250

attached base packages:
[1] stats     graphics  grDevices utils     datasets
[6] methods   base

other attached packages:
 [1] MASS_7.3-40       xtable_1.7-4      tempdisagg_0.24.0
 [4] lattice_0.20-31   tseries_0.10-34   quantmod_0.4-4
 [7] TTR_0.22-0        xts_0.9-7         expsmooth_2.3
[10] forecast_6.1      timeDate_3012.100 zoo_1.7-12
[13] TSAFBook_0.1      devtools_1.8.0    knitr_1.10.5
[16] stringr_1.0.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.11.6       xml2_0.1.1        magrittr_1.5
 [4] roxygen2_4.1.1    colorspace_1.2-6  quadprog_1.5-5
 [7] tools_3.2.0       nnet_7.3-9        parallel_3.2.0
[10] grid_3.2.0        git2r_0.10.1      rversions_1.0.1
[13] digest_0.6.8      formatR_1.2       codetools_0.2-11
[16] curl_0.8          memoise_0.2.1     evaluate_0.7
[19] fracdiff_1.4-2    stringi_0.4-1
```

Zawartość

W książce znaleźć można informacje na temat klasycznych modeli statystycznych oraz metod algorytmicznych stosowanych do dekompozycji i prognozowania szeregów czasowych. Omówiono także najważniejsze przekształcenia wstępne szeregów, poprzedzające właściwą analizę. Bardziej zaawansowany lub dociekliwy Czytelnik znajdzie w podręczniku informacje, jaka literatura pomoże mu w pogłębianiu wiedzy w zakresie zaawansowanych metod analizy szeregów czasowych.

Najważniejsze zagadnienia omówione w książce:

1. Wczytywanie i podstawowe operacje na danych w środowisku R.
2. Graficzna prezentacja danych:
   - ✏ wykresy zwykłe i sezonowe,
   - ✏ wykresy autokorelacji,
   - ✏ wybrane wykresy specjalistyczne.
3. Przekształcenia wstępne szeregów:
   - ✏ przekształcenia szeregu ułatwiające analizę,
   - ✏ korekty kalendarzowe,
   - ✏ agregowanie danych,
   - ✏ różnicowanie.
4. Dekompozycja szeregów – identyfikacja regularnych tendencji w danych:
   - ✏ składowe szeregu czasowego: trend, cykliczność i sezonowość,
   - ✏ metody wygładzania i dekompozycji szeregu,
   - ✏ eliminacja trendu i sezonowości.
5. Modele ARIMA:
   - ✏ modele stacjonarne i niestacjonarne (AR, MA, ARMA, ARIMA, SARIMA),
   - ✏ identyfikacja modelu i estymacja jego parametrów,
   - ✏ analiza poprawności dopasowania modelu – diagnostyka,
   - ✏ wybór optymalnego modelu dla danych.
6. Prognozowanie szeregów:
   - ✏ najprostsze (naiwne) metody prognozowania,
   - ✏ prognozowanie na podstawie modeli ARIMA,
   - ✏ algorytmy wygładzania wykładniczego,
   - ✏ prognozy oparte na dekompozycji,
   - ✏ ocena i porównanie dokładności prognoz.

Dane wykorzystywane w przykładach

Omawiane w podręczniku metody analizy i prognozowania szeregów ilustrujemy, wykorzystując przykładowe dane. Są to przede wszystkim rzeczywiste

szeregi czasowe, wybrane z różnych obszarów zastosowań i zróżnicowane pod względem występujących regularności, siły zależności czasowej, częstotliwości próbkowania oraz długości. Aby zaprezentować idee poszczególnych metod oraz ułatwić proste pokazanie różnych – mogących wystąpić w praktyce – wariantów, wykorzystujemy także dane symulowane.

Techniczne aspekty związane z wykorzystaniem poszczególnych funkcji środowiska **R** (takie jak: parametry wejściowe danej funkcji, postać wyników i możliwość ich prezentacji graficznej) przedstawiamy głównie opierając się na kilku typowych szeregach czasowych:

- ✏ `AirPass` – historyczne dane zawierające informacje o miesięcznej liczbie pasażerów linii lotniczych,
- ✏ `pkb` – wartości kwartalnego produktu krajowego brutto w Polsce,
- ✏ `usgdp` – szereg zawierający kwartalne wartości produktu krajowego brutto w Stanach Zjednoczonych.

W podręczniku zdecydowaliśmy się bazować głównie na tych zbiorach danych, wierząc, że ułatwi to Czytelnikowi zrozumienie złożonego (wieloetapowego) schematu analizy szeregów czasowych oraz zwiększy przejrzystość prezentacji poszczególnych metod. W razie potrzeby w przykładach odwołujemy się także do innych danych.

Materiały uzupełniające

Książce towarzyszy biblioteka (pakiet) `TSAFBook`, opracowana dla środowiska **R**, zawierająca szeregi czasowe wykorzystywane w przykładach. Znalazły się tutaj przede wszystkim dane dotyczące Polski, w tym szeregi makroekonomiczne, finansowe oraz dotyczące sytuacji gospodarczej. Biblioteka `TSAFBook` dostępna jest w repozytorium CRAN (`http://cran.r-project.org/`) i może być zainstalowana za pomocą komendy wydanej w konsoli **R**'a: `install.packages("TSAFBook")` lub z poziomu GUI.

Pakiet `TSAFBook` oraz dodatkowe materiały, w szczególności pliki z danymi i fragmenty **R**-kodów, można znaleźć na towarzyszącej książce stronie `http://TSAFBook.quantup.pl`.

Uwagi od Czytelników

Zachęcamy wszystkich Czytelników do dzielenia się wszelkimi uwagami na temat książki, pomysłami usprawnień i uzupełnień oraz informacjami o powstałych wątpliwościach. Z Adamem można się skontaktować korzystając z adresu `a.zagdanski@gmail.com`, z Arturem – z `artur@quantup.pl`.

Jak korzystać z książki?

Niecierpliwych Czytelników, którzy chcą jak najszybciej rozpocząć analizowanie szeregów z wykorzystaniem pakietu **R**, zachęcamy do zapoznania się w pierwszej kolejności z podrozdziałem 2.4, w którym przedstawiona jest możliwie kompletna analiza wybranego szeregu czasowego, z uwzględnieniem najbardziej popularnych metod i modeli.

Przy pierwszym czytaniu niektóre podrozdziały bądź ich fragmenty można pominąć i wrócić do nich (w razie potrzeby) później. Poniżej przedstawiamy nasze sugestie dla kolejnych rozdziałów książki.

- ✏ **Rozdział 2: Wprowadzenie** – zalecamy przeczytanie w całości.
- ✏ **Rozdział 3: Dane** – można pominąć podrozdziały: 3.5 i 3.6.
- ✏ **Rozdział 4: Wykresy i analiza opisowa** – można pominąć podrozdział 4.1.2.
- ✏ **Rozdział 5: Przekształcenia wstępne szeregów** – można pominąć podrozdziały: 5.1, 5.4 i 5.5.
- ✏ **Rozdział 6: Dekompozycja szeregów czasowych** – można pominąć podrozdziały: 6.1.5, 6.2.2, 6.5.
- ✏ **Rozdział 7: Modele ARIMA** – można pominąć podrozdział 7.4. Aby dopasowywać modele ARIMA do danych (np. na potrzeby konstrukcji prognoz), można, w pierwszym kroku, opierać się na automatycznym wyborze optymalnego modelu (podrozdział 7.8.5), a w przyszłości wrócić do bardziej zaawansowanych zagadnień dotyczących: identyfikacji modelu (podrozdział 7.5), estymacji parametrów (podrozdział 7.6) i diagnostyki (podrozdział 7.7). Uwaga: rozdział 7 jest najbardziej zaawansowany pod względem metodologicznym!
- ✏ **Rozdział 8: Prognozowanie** – przy pierwszym czytaniu można pominąć podrozdział 8.4.7, a także bardziej teoretyczne fragmenty dotyczące poszczególnych metod. Dodatkowo podrozdziały: 8.3, 8.4 i 8.5 poświęcone odpowiednio konstrukcji prognoz na podstawie: modeli ARIMA, algorytmów wygładzania wykładniczego oraz dekompozycji mogą być w zasadzie czytane niezależnie. Uwaga: rozdział 8 jest najbardziej obszernym rozdziałem w książce!

Aby ułatwić Czytelnikowi korzystanie z podręcznika, pewne fragmenty zostały wyróżnione. Wykorzystujemy w tym celu następujące oznaczenia:

 – fragmenty zasługujące na szczególną uwagę, często mające postać ważnych zaleceń i uwag praktycznych,

 – bardziej zaawansowane lub mniej standardowe zagadnienia, do zrozumienia których może być potrzebne sięgnięcie do dodatkowej literatury.

Podziękowania

Autorzy pragną podziękować wszystkim osobom, które bezpośrednio lub pośrednio przyczyniły się do powstania tej książki i nadania jej obecnego kształtu. Podręcznik powstał w dużej mierze na podstawie naszych wieloletnich doświadczeń, związanych z pracą dydaktyczną na Politechnice Wrocławskiej, pracą konsultantów biznesowych oraz prowadzeniem, we współpracy z firmą QuantUp, komercyjnych szkoleń i warsztatów. Dziękujemy więc wszystkim naszym współpracownikom, dyplomantom, stażystom, studentom i uczestnikom szkoleń, którzy zainspirowali nas do napisania tego podręcznika i których uwagi w jakikolwiek sposób wpłynęły na jego aktualną postać.

Adam pragnie w szczególny sposób podziękować panu dr. hab. inż. Romanowi Różańskiemu (prof. nadzw. Politechniki Wrocławskiej) za zainspirowanie tematyką analizy i prognozowania szeregów czasowych oraz za długoletnią współpracę naukową i dydaktyczną.

Na koniec, najserdeczniejsze podziękowania kierujemy do naszych najbliższych, bez których wsparcia i wyrozumiałości podręcznik by po prostu nie powstał.

Artur dziękuje wyjątkowo ciepło swojej żonie Agnieszce, która od wielu lat wspiera go w jego wszystkich zawodowych (oczywiście nie tylko) działaniach oraz inspiruje do podejmowania nowych wyzwań.

Adam Zagdański, Artur Suchwałko, Wrocław 2015

O autorach

Adam Zagdański

Jest pracownikiem naukowo-dydaktycznym Wydziału Matematyki Politechniki Wrocławskiej. Ukończył matematykę stosowaną na Wydziale Podstawowych Problemów Techniki Politechniki Wrocławskiej (specjalność statystyka matematyczna). Doktor nauk matematycznych. Odbył dwuletni staż podoktorski na Uniwersytecie w Toronto, uczestnicząc w projekcie badawczym związanym z zastosowaniami nowoczesnych metod statystycznych i *data mining* w analizie danych genetycznych.

Jest współautorem kilkunastu artykułów naukowych z zakresu statystyki i bioinformatyki. Brał aktywny udział w kilkunastu zagranicznych i krajowych konferencjach naukowych. Jego aktualne zainteresowania naukowe to zastosowanie metod statystyki wielowymiarowej i *data mining* w analizie danych biologicznych (m.in. danych mikromacierzowych i spektrometrycznych), metody integracji danych genomicznych oraz analiza i prognozowanie szeregów czasowych.

Posiada ponad piętnastoletnie doświadczenie dydaktyczne. Prowadzi wykłady i laboratoria komputerowe z zakresu *data mining* i statystyki stosowanej (w tym m.in.: metody nieparametryczne statystyki, analiza i prognozowanie szeregów czasowych i modelowanie stochastyczne). Jest promotorem kilkunastu prac dyplomowych z informatyki i statystyki.

Uczestniczył w komercyjnych projektach związanych z zastosowaniem nowoczesnych metod *data mining* oraz modelowaniem i prognozowaniem szeregów czasowych. Od blisko 10 lat jest również konsultantem w dziedzinie analizy danych. Prowadził komercyjne szkolenia z zakresu analizy danych i prognozowania szeregów czasowych we współpracy z firmą QuantUp. Od kilku lat współpracuje także ze szwedzką firmą bioinformatyczną MedicWave.

Artur Suchwałko

Posiada blisko dwudziestoletnie doświadczenie w różnorodnych projektach komercyjnych i naukowych związanych z analizą danych. Pracował dla różnych firm, od start-upów do międzynarodowych korporacji, i w różnych rolach, od pracownika przez konsultanta, po właściciela. Jest doświadczonym programistą oraz menedżerem projektów. Kierował zespołami do kilkunastu osób i brał udział w tworzeniu firm bazujących na analizie danych.

Od samego początku swojej drogi zawodowej łączy stosowanie matematyki, pracę naukową i dydaktyczną.

Przez kilkanaście lat był statystykiem, a później ekspertem w Departamencie Ryzyka Kredytowego i Analiz Lukas Banku. Zdobył tam duże doświadczenie w praktycznym modelowaniu statystycznym, także w tworzeniu oprogramowania służącego do tego celu.

Jest doktorem matematyki oraz autorem i współautorem kilkunastu prac naukowych. Kilkanaście lat uczył statystyki, *data miningu* i programowania na Politechnice Wrocławskiej. Był promotorem ponad pięćdziesięciu prac dyplomowych magisterskich i inżynierskich z matematyki i informatyki.

Od roku 2007 uczy analityków, jak analizować dane. Przeprowadził wiele komercyjnych szkoleń z dziedziny budowy i walidacji modeli predykcyjnych, innych obszarów analizy danych oraz **R**, spędzając w salach szkoleniowych blisko półtora tysiąca godzin.

Od paru lat rozwija z sukcesem swoją firmę QuantUp (`http://quantup.pl`) zajmującą się analizą danych, modelowaniem statystycznym i tworzeniem oprogramowania oraz szkoleniami z tych dziedzin.

Kilka lat temu został dyrektorem naukowym (Chief Science Officer) szwedzkiej firmy bioinformatycznej MedicWave. Od roku 2012 jest dodatkowo Vice CEO tej firmy.

Jest fanem systemu **R**. Od kilkunastu lat używa **R** i uczy, jak go używać. Popularyzuje także analizę danych i system **R** uczestnicząc w konferencjach biznesowych oraz działaniach non profit.

Więcej informacji o nim można znaleźć na jego profilu LinkedIn: `http://www.linkedin.com/in/artursuchwalko`.