Adam Zagdański | Artur Suchwałko

Analiza i prognozowanie szeregów czasowych

Praktyczne wprowadzenie na podstawie środowiska R

PWN

Spis treści

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`.

Rozdział 2

Wprowadzenie

Analiza szeregów czasowych zyskuje ostatnio coraz bardziej na znaczeniu i jest z niesłabnącym powodzeniem stosowana w wielu obszarach nauki, biznesu czy przemysłu. Podstawowym celem analizy szeregów czasowych jest zbudowanie modelu, który będzie dobrze opisywał dynamikę czasową obserwowanego zjawiska i który może być następnie wykorzystany do prognozowania przyszłych (nieznanych) wartości. Przed właściwym modelowaniem koniecznie jest oczywiście odpowiednie przygotowanie danych, a w szczególności uwzględnienie na tym etapie sposobu gromadzenia danych i wykonanie przekształceń, które mogą ułatwić dopasowanie modelu. Duże znaczenie ma również prawidłowa identyfikacja regularności występujących w analizowanym szeregu, takich jak tendencje długoterminowe (trendy) oraz wahania sezonowe (sezonowość).

Podobnie jak w przypadku innych nowoczesnych metod analizy ilościowej, duże znaczenie dla rozwoju metod analizy i prognozowania szeregów czasowych miał rozwój technologii informatycznych. Istotny dla upowszechnienia metod analizy szeregów jest łatwy dostęp do specjalistycznego oprogramowania, dającego praktykom możliwość zastosowania zarówno standardowych, jak i nowoczesnych algorytmów oraz przeprowadzenia zaawansowanych obliczeń. Szczególną rolę odegrał w tym przypadku rozwój darmowego oprogramowania, takiego jak system statystyczny **R** (`www.r-project.org`). Dzięki temu dostęp do metod analizy i prognozowania szeregów stał się powszechny, a nowoczesne metody opracowywane w ośrodkach naukowych są często w krótkim czasie udostępniane w postaci dodatkowych pakietów (bibliotek), rozszerzających standardowe możliwości środowiska **R**.

Pierwszy rozdział książki ma charakter wprowadzający i omawiamy w nim najważniejsze zagadnienia związane z analizą szeregów czasowych. Będzie mowa m.in. o tym, czym jest szereg czasowy i jakie są podstawowe zadania analizy szeregów. Szczególną uwagę poświęcamy przeglądowi etapów występujących w analizie szeregów czasowych, począwszy od niezbędnych przekształceń, poprzez identyfikację regularnych wzorców, a kończąc na prognozowaniu.

Co ważne, ten rozdział zawiera „przykład dla niecierpliwych", który pokazuje, jak wykonać kompletną analizę szeregu czasowego w **R**. Przykład jest zaprezentowany w możliwie uproszczony sposób, aby ułatwić zrozumienie procesu. Zawiera też odnośniki do rozdziałów, w których poszczególne zagadnienia są szerzej omówione.

Zapoznanie się z tym wprowadzającym rozdziałem ułatwi Czytelnikowi orientację w strukturze książki i pozwoli zrozumieć całość procesu analizy szeregu czasowego.

2.1. Czym jest szereg czasowy?

Szeregiem czasowym (ang. *time series*) nazywamy obserwacje interesującej nas wielkości zarejestrowane w kolejnych (zazwyczaj regularnych) odstępach czasu, np. kolejnych dniach, miesiącach lub kwartałach. Przykładami mogą być szeregi zawierające informacje o rocznej wielkości produkcji samochodów osobowych, stopie bezrobocia w kolejnych miesiącach czy też kwartalne dane dotyczące przyrostu produktu krajowego brutto (PKB).

Rysunek 2.1 przedstawia typowe szeregi czasowe z różnych obszarów zastosowań. Mamy następujące dane[1]:

szereg A: miesięczna liczba pasażerów linii lotniczych (w tysiącach) w USA, w latach 2002–2014,

szereg B: kwartalne wartości produktu krajowego brutto (PKB) w Polsce, zarejestrowane w okresie 1995–2014,

szereg C: kwartalna wielkość produkcji samochodów osobowych w UK w okresie 1977:1–2005:1,

szereg D: miesięczne, średnie ceny 1 litra benzyny w Polsce w okresie 2006–2014,

szereg E: rentowność 10-letnich obligacji skarbowych USA, dane miesięczne w okresie styczeń 1994–maj 2004,

szereg F: miesięczne kursy zamknięcia indeksu Dow Jones w okresie styczeń 1990–marzec 2007,

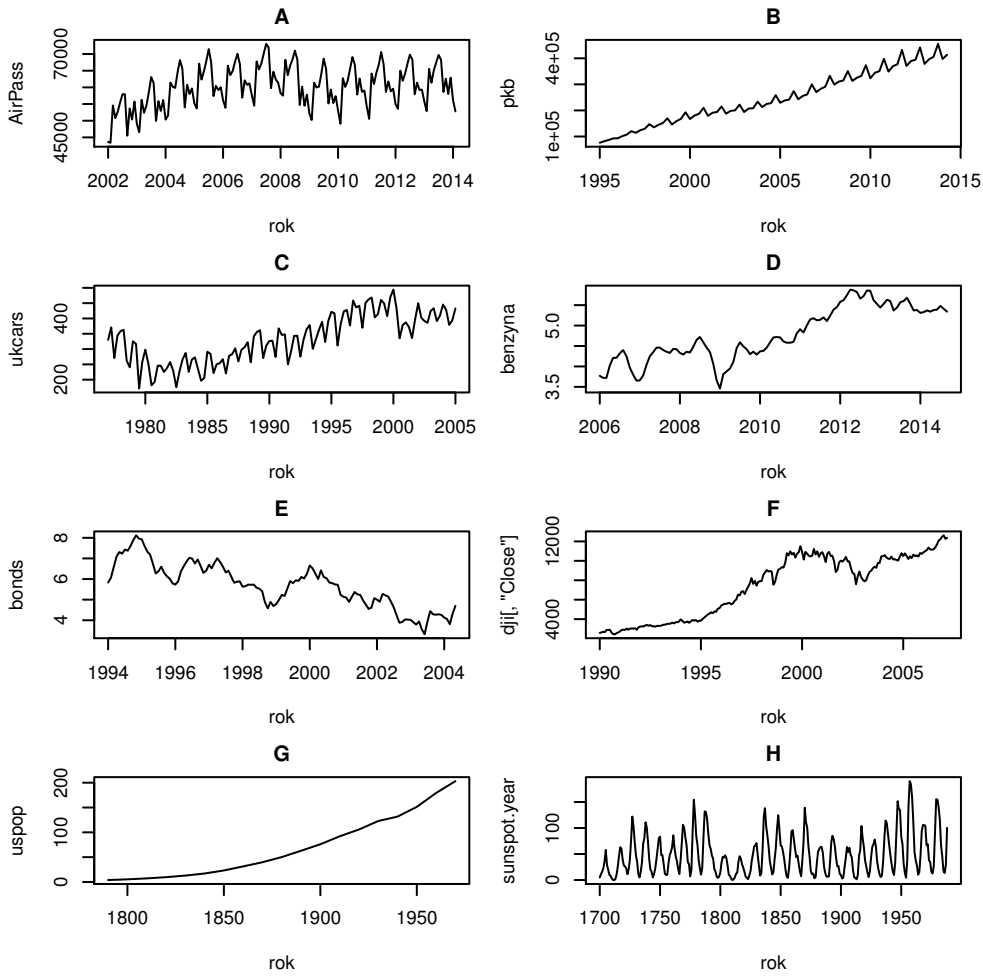
szereg G: populacja USA (w milionach) w okresie 1790–1970, dane 10-letnie,

szereg H: roczne liczby plam słonecznych w latach 1700–1988.

Jak widzimy, szeregi czasowe mogą różnić się częstotliwością próbkowania, czyli interwałami czasowymi pomiędzy kolejnymi obserwacjami (np. kolejne dni, miesiące, kwartały, lata[2] itd.).

[1] Dokładniejszą informację na temat źródła pochodzenia prezentowanych danych oraz ich dostępności w środowisku **R** przedstawimy w podrozdziale 3.1.

[2] W podręczniku ograniczamy się do tzw. regularnych szeregów czasowych, które charakteryzują się jednakowymi odstępami czasowymi pomiędzy kolejnymi obserwacjami.



Rysunek 2.1. Przykłady szeregów czasowych z różnych obszarów zastosowań

Analizując wykresy przykładowych szeregów (rys. 2.1), możemy zauważyć duże zróżnicowanie pod względem występujących regularności. W szczególności, mamy szeregi, w których obecna jest wyraźna tendencja długoterminowa (trend) oraz takie, w przypadku których łatwo można dostrzec zachowania okresowe (sezonowe).

Kolejne obserwacje szeregu czasowego charakteryzują się zatem nieprzypadkowym porządkiem i wykazują najczęściej istotną zależność (korelację). To właśnie badanie charakteru i siły tej zależności jest podstawowym zadaniem analizy szeregów czasowych. Oczywiście zależność pomiędzy obserwacjami jest często wykorzystywana do prognozowania przyszłych wartości szeregu.

Dla szeregów czasowych stosowane są zwykle oznaczenia: X_t, Y_t lub Z_t, w których argument t oznacza czas (ang. *time*) i należy do ustalonego zbioru

chwil (punktów czasowych), np. $t \in I$. Często zamiast operowania rzeczywistą skalą czasową (daty kalendarzowe), obserwacje indeksuje się kolejnymi liczbami naturalnymi. Dla przykładu szereg zawierający dane miesięczne dla kolejnych 12 lat (łącznie 144 obserwacje) zapisujemy dla uproszczenia jako $X_1, X_2, \ldots, X_{144}$.

2.2. Główne zadania analizy szeregów czasowych

Z szeregami czasowymi często się spotykamy, gdy zachodzi konieczność podejmowania ważnych decyzji biznesowych, np. dotyczących kupna/sprzedaży, produkcji, zatrudnienia czy logistyki. Na początek wymieńmy kilka typowych przykładów analiz, w których wykorzystuje się szeregi czasowe:

- ✏ Prognozowanie wielkości sprzedaży lub popytu na określony produkt/surowiec w kolejnych okresach.
- ✏ Prognozy wartości wskaźników makroekonomicznych (np. inflacji lub PKB) w kolejnych kwartałach.
- ✏ Analiza sytuacji na rynku pracy (w szczególności analiza tendencji dotyczących bezrobocia i zatrudnienia, w różnych grupach wiekowych).
- ✏ Prognozowanie wartości akcji danej spółki, cen surowców, kursów walutowych itp. w kolejnych okresach.
- ✏ Przewidywanie zmian cen danego produktu (np. paliw) w kolejnych miesiącach.
- ✏ Analiza zmian demograficznych, socjologicznych, klimatycznych i ich wpływu na koniunkturę w określonej gałęzi przemysłu.

Wymienione przykładowe zastosowania analizy szeregów czasowych cechuje więc dość duża różnorodność. Wyłaniają się tutaj jednak dwa główne zadania, tzn. identyfikacja regularnych tendencji (tzw. dekompozycja szeregu czasowego) oraz prognozowanie. Regularne tendencje to przede wszystkim trend – długoterminowa tendencja rozwojowa, np. wzrostowa lub spadkowa, oraz sezonowość – cykliczne wahania wartości szeregu wokół tendencji rozwojowej, związane najczęściej z danym miesiącem roku, porami wakacji lub zmianami pogody.

Przed analitykiem, wyciągającym wnioski na podstawie analizy szeregów czasowych, pojawia się więc wiele pytań, w tym na przykład:

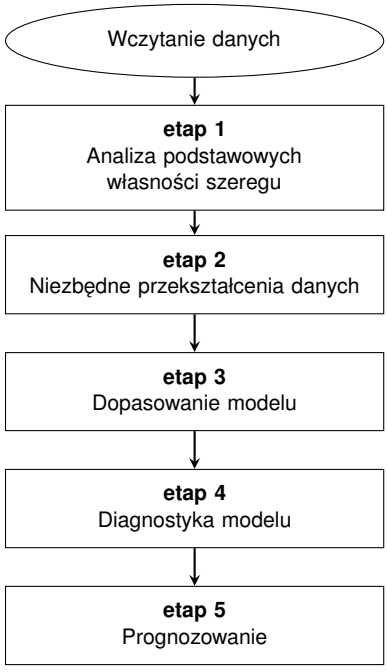
- ✏ Jak przygotować dane przed właściwą analizą?
- ✏ Jakie metody analizy i prognozowania szeregów powinny być zastosowane?
- ✏ Jak mierzyć błąd prognozy i czy uzyskane prognozy są wystarczająco dokładne?

W kolejnych rozdziałach naszego podręcznika postaramy się odpowiedzieć na te i podobne pytania.

2.3. Etapy w analizie szeregu czasowego

Analiza szeregu czasowego jest zazwyczaj procesem wieloetapowym. Rysunek 2.2 przedstawia typowe etapy w analizie szeregu czasowego. Niektóre etapy są opcjonalne i mogą być ściśle związane ze specyfiką danych, na przykład sposobem rejestrowania danych, regularnościami występującymi w analizowanym szeregu itp. Większość metod i modeli stosowanych w analizie szeregów występuje zazwyczaj w wielu wariantach lub wymaga wybrania adekwatnych parametrów. Po przeprowadzeniu określonego etapu analizy szeregu czasowego może się okazać, że konieczny będzie powrót do poprzednich etapów. Dla przykładu, jeżeli przeprowadzając diagnostykę (etap 4) stwierdzimy, że model nie jest dobrze dopasowany do danych, możemy powrócić do etapu dopasowania (etap 3) i spróbować znaleźć lepszy model dla naszego szeregu czasowego. Niekiedy konieczny może być także powrót do etapu 2 i wykonanie dodatkowych przekształceń, które mogą poprawić jakość dopasowania modelu. Oczywiście, może się również zdarzyć tak, że konieczne będzie powtórne przyjrzenie się podstawowym własnościom analizowanego szeregu, a więc powrót do etapu 1.

Wykonanie kolejnych etapów analizy szeregu czasowego wymaga zarówno dobrej znajomości samej metodologii, jak i odpowiedniego doświadczenia



Rysunek 2.2. Ogólny schemat analizy szeregów czasowych

analityka. Z pomocą przychodzą tu rozwiązania (m.in. odpowiednie funkcje dostępne w środowisku **R**) opracowane z myślą o automatyzacji pewnych etapów analizy. Stosowanie tego rodzaju narzędzi nie zwalnia oczywiście analityka z konieczności weryfikacji uzyskanych wyników.

Poszczególne etapy w analizie szeregu czasowego (w szczególności stosowane modele i metody) będziemy sukcesywnie omawiali w kolejnych rozdziałach książki. Na wstępie chcielibyśmy jednak umożliwić Czytelnikowi spojrzenie na całość tego procesu, przy okazji prezentując pokrótce (bez nadmiernych szczegółów technicznych) możliwości oferowane w tym zakresie w pakiecie **R**. Z tego względu w podrozdziale 2.4 przedstawiona jest możliwie kompletna analiza wybranego szeregu czasowego, z uwzględnieniem najbardziej popularnych metod i modeli.

2.4. Przykład dla niecierpliwych

Na potrzeby przykładu wybraliśmy szereg czasowy zawierający informacje o liczbie turystów korzystających z noclegów na terenie Dolnego Śląska. Są to dane miesięczne, zarejestrowane w okresie styczeń 2009–marzec 2014. Dane zostały pobrane ze strony GUS: Bank Danych Lokalnych, dział TURYSTYKA (`http://stat.gov.pl/bdl/`).

Analiza, którą przeprowadzimy, będzie obejmowała następujące kroki:

1. Wczytanie (import) danych do środowiska **R**.
2. Konwersja danych na format odpowiedni dla **R**.
3. Analiza podstawowych własności szeregu.
4. Dekompozycja: identyfikacja trendu i sezonowości.
5. Niezbędne przekształcenia szeregu.
6. Podział danych na zbiór uczący i testowy.
7. Dopasowanie modeli.
8. Prognozowanie: konstrukcja prognoz punktowych i przedziałów predykcyjnych.
9. Ocena i porównanie dokładności prognoz.

2.4.1. Wczytanie danych

Zakładamy, że kolejne wartości szeregu zapisane są w pliku tekstowym `hotele.txt`[3], który został umieszczony w wybranym katalogu (w naszym przypadku, w katalogu `C:/Users/adam/Desktop/Dane`). Poniższy fragment kodu pozwala wczytać dane do przestrzeni roboczej **R**'a.

[3] Plik `hotele.txt` dostępny jest na stronie `http://TSAFBook.quantup.pl`. Dane w formacie binarnym **R**'a można także znaleźć w towarzyszącym książce pakiecie `TSAFBook`.

W książce przedstawione są najważniejsze metody i modele analizy szeregów czasowych. Nie jest to jednak wyłącznie przegląd metodologii, jaki można znaleźć w klasycznych podręcznikach, ale znajdują się w niej konkretne wskazówki dla praktyków, 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.

Publikacja ma charakter uniwersalny. Nie ogranicza się do konkretnego obszaru zastosowań. Są w niej przykłady analizy szeregów makroekonomicznych, finansowych, demograficznych czy też szeregów związanych z wielkością sprzedaży lub ceną usług i produktów. Ogromną zaletą książki jest wykorzystanie środowiska R, popularnego i powszechnie stosowanego darmowego narzędzia praktycznej analizy danych. Czytelnik znajdzie w niej również fragmenty kodów pozwalających na samodzielne wykonanie opisywanych analiz.

Jest to przydatne źródło wiedzy dla praktyków podejmujących ważne decyzje biznesowe oraz naukowców prowadzących badania w dziedzinie ekonomii, demografii, socjologii oraz nauk przyrodniczych. Jest to również niezbędny podręcznik dla studentów kierunków ścisłych, technicznych oraz wybranych humanistycznych.

ADAM ZAGDAŃSKI jest pracownikiem naukowo-dydaktycznym Wydziału Matematyki Politechniki Wrocławskiej. Jego aktualne zainteresowania naukowe skupiają się wokół prognozowania szeregów czasowych oraz zastosowania metod statystycznych w analizie danych biologicznych. Prowadził wiele kursów akademickich oraz szkolenia związane z zastosowaniami metod statystyki i pozyskiwania wiedzy. Brał także udział w komercyjnych projektach z zakresu analizy danych i prognozowania. Jest entuzjastą analizy danych w środowisku R, które wykorzystuje w pracy zawodowej od kilkunastu lat.

ARTUR SUCHWAŁKO ma blisko 20 lat doświadczenia zawodowego w analizie danych, modelowaniu statystycznym i programowaniu na potrzeby analizy danych. Pracował w różnych rolach (statystyk, ekspert, trener, konsultant, naukowiec i nauczyciel akademicki, programista oraz manager zespołu programistów, zarządzający firmą, właściciel firmy) dla podmiotów różnych rodzajów (uczelnie, firmy z wielu branż) oraz różnej wielkości (od start-upów po międzynarodowe korporacje). Od roku 2011 prowadzi własną firmę analityczną QuantUp. Popularyzuje system R i analizę danych, także podczas konferencji oraz działań non-profit.

Partner wydania: QUANTUP

PWN
Wydawnictwo
Naukowe PWN SA
Infolinia: 801 33 33 88
www.pwn.pl

ISBN 978-83-01-18356-1
9 788301 183561 >