SYLLABUS: Summer, 2023
STATISTICS 626: Methods in Time Series Analysis

IMPORTANT: The exams will be given in the evenings between 7-8:15 pm (July 10 and Aug. 1) and they are Zoom-proctored. This will obviate the need for DL students to find human proctors. Any student in the 700-Section who needs alternative accommodation should contact Penny Jackson (pljackson44@tamu.edu) at least a week before the time of the Exam.

INSTRUCTOR: Mohsen Pourahmadi, Blocker 439.
TIME: MTW 10-11:15 am CST, Blocker 457
OFFICE HOURS: By Appointment (Please request by email at least two hours before your desired time for a Zoom meeting). Zoom Link (https://tamu.zoom.us/my/j5840599736). Zoom Q&A: Tu 7 pm (On Demand). Please send an email by 4 pm or earlier if you are planning to attend and have questions. Otherwise, there will be no Q&A session that evening.
PH. & FAX NOs: 979-845-3164, 979-845-3144
E-MAIL: pourahm@tamu.edu

Teaching Assistant: Mr. Jian Yan (jian@stat.tamu.edu)
ZOOM Office Hours: Wed. 6-8 pm, Zoom link (https://tamu.zoom.us/j/6944986984).

Required TEXT: *Time Series: A Data Analysis Approach Using R*, Shumway, R. and Stoffer, D., Chapman and Hall/CRC, 2019, ISBN 9780367221096 - CAT# K421312
PREREQUISITE: STAT 601 or 642/608 and a working knowledge of regression analysis. basic math such as complex numbers, trig. functions and matrices.

PLAN OF THE COURSE: STAT 626 is for a mixed group of motivated graduate students in statistics and other fields who seek an intermediate background in methods of time series analysis and forecasting. The course starts with discussing some common examples of time series datasets. Then, introduces the basic theory of stationary processes such as the covariance, autocorrelation and partial correlation functions, ARIMA models, spectral analysis, and forecasting. The **data analysis project** in the course is extremely helpful in creating a balance between theory and applications to economics, engineering and biomedical sciences or areas of interest to students in the course. In the first two weeks, students are expected to form **diverse** groups of about five (300 and 700 Sections, Stat majors, non-stat majors,...,). Students in a group should ideally be interested in similar/complementary application areas. They shall develop a project plan and identify the relevant research papers and dataset to study, analyze and present in class or via ZOOM at various indicated times in the course, see item #4 below for more details. There is a wealth of genuine datasets and R programs in the text, please take a casual look at the data examples in Chap. 1 of the text before the start of the semester.

A Tentative Course Outline: Dates, topics and sections may change.

Week	Topic	Section
	Syllabus; TS Data and Plot; Reg. Residuals	1.1-1.2
1	TS Models I: WN, RW, AR, MA	1.3
	Autocorrelation Function (ACF) and R	1.4, Append
	Stationary TS, Linear Processes	2.1
2	Estimation of autocorrelation	2.2
	Bivariate TS and cross-correlation	2.3
	5-min Project Presentation	
	A Review of Regression: $y = X\beta + e$.	3.1-3.2
3	Exploratory Data Analysis and Smoothing	3.3
	TS Models II: AR(1), MA	4.1
4	TS Models II: ARMA, $\phi(B)x_t = \theta(B)w_t$.	4.2
	ARMA : Invertibility& Causality (Roots $\|z\| > 1$,$\pi-$, $\psi-$) weights)	
	ARMA ACF and PACF	4.1-4.2
5	ARMA Forecasting	4.4
	Estimation: LS, Yule-Walker Eqs., MLE	4.3
6	**No classes during the week of July 2nd.**	
	Prep. and Study for **Exam I (75 min.), Monday July 10**,	
	Work on the 10-min Project Pres.	
7	**Exam I, Monday July 10**	
	ARIMA Models	Ch. 5
	Building ARIMA Models	5.2
8	**10-min Project Presentation**	
	SARIMA Models	5.3
	Regression with Autocorrelated Errors	5.4
9	GARCH Models	8.1
	Unit-Root Testing,	8.2
	Spectral Analysis	6.2
10	Spectral Estimation	Chap 7
	Special Topics Requested by Students	
	EXAM II (75 minutes), Tu. August 1	
10/11	**20-min Project Presentation**	
	Final PPR. (Written report due.)	
	Final Project Presentation may start on Aug. 2nd,	
	and continue to the following week till all groups are done.	
	These sessions will be on Zoom and in the evenings (Approx. 7-8:30 pm).	

LEARNING OUTCOMES: Students in the course will develop skills to:

1. Distinguish time series (dependent) data from the more standard sample (independent) data,
2. Understand the deeper aspects of stationarity concept, and reduce nonstationary time series data to stationarity,
3. Analyze time series data using graphical tools, regressions, ARIMA models and forecast future values using the R software,
4. Analyze seasonal time series data using SARIMA models, etc.
5. Deal with bivariate, multivariate and high-dimensional time series data.

GRADE POLICY:

1. **Exams**: There will be two midterms (75 minutes long, mostly multiple-choice) and no final. The midterms will constitute 25% and 35% of the grade, respectively. The midterms are closed book/notes, a formula sheet will be provided.
2. **Homeworks**: Will be assigned regularly and posted at Canvas, it will contribute 10% to your grade. **Homework solutions must be in a single PDF file and posted at Canvas**. The quality of writing and logical presentation of the arguments leading to a result, not just the correct answer, will contribute greatly to the grade for this part of the course. You may consult with other students about the homework, but always write up your solutions by yourself. You should never just copy from another person. **Do not include R programs and computer printouts in your HW, unless asked to do so.** From time to time there will be bonus HW problems/questions. **Bonus HW points will be banked**, and used only in borderline cases when assigning letter grades at the end of the summer.
3. Exam Policy: All exams are given online through Canvas, and they are Zoom-proctored and recorded.
4. Homework Policy: Homework assignments will be available in the Homework folder on Canvas. There will be regular assignments with due date posted. Homework solutions must be in a single PDF file and posted at Canvas You should be identified on the initial page with your TYPED Name, Course and Email address. Your homework solutions must be your own work, not from outside sources, consistent with the university rules on academic integrity.

5. **Missed assignments**: Each homework must be turned in by 10:00 am CST on the assigned due date. Late homework is not accepted without an excuse that is recognized as valid by the university. Likewise, you will only be allowed to make up an exam if it is missed for a valid reason. See items 11-13 for more details.

6. The final course grade will be based on the standard scale where a total of 90 to 100 percent will be an A, 80 to less than 90 percent will be a B, etc.

7. **Classroom participation and Canvas Discussion** among students are encouraged, they are integral parts of the learning process .

8. **Data Analysis Project**: Will involve a significant amount of data analysis, reading the relevant literature in the student's area of interest, computational effort and discussion. There will be bi-weekly written reports and during class times Zoom presentations (live/recorded). The project starts by each group choosing a suitable time series dataset.

   The first project presentation will be 5-min. long introducing the group members, describing the data, etc, and the first written report could be a page long, describing the data and application area, and is due 24 hours after the presentation. Subsequent project presentations and reports are 10-min. and 20-minute long, 5-page and 10-page long, respectively. The project is worth 30% of your grade, and provides a wonderful opportunity for hands-on experience, developing teamwork and organizational skills, writing and presentation skills. Your grades for this part will be determined by quality of your data analysis, Zoom/oral presentations, written reports, display of teamwork,..., interest shown by other groups and students and your group's ability to discuss and answer questions.

   The project reports should be organized and typed following the format of a research article in statistics or your area of applications, each report is due 24 hours after the classroom presentation. It should contain the names of the group members and their responsibilities, have a title, abstract, objectives,..., references. The quality of writing and Zoom presentation s will contribute greatly to the grade for this part of the course.

9. ACADEMIC INTEGRITY STATEMENT: "An Aggie does not lie, cheat, or steal or tolerate those who do." The Aggie Honor Council Rules and Procedures are available at http://www.tamu.edu/aggiehonor.

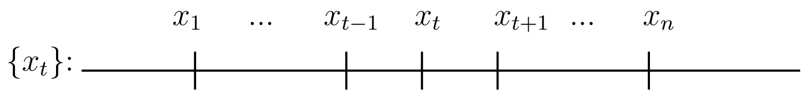
10. STATEMENT ON PLAGIARISM: As commonly defined, plagiarism consists of passing off as one's own ideas, words, writing, etc., which belong to another. In accordance with this definition, you are committing plagiarism if you copy the work of another

STAT 626: Outline of Lecture 1
Time Series Data: Examples (§1.2)

1. A Quick Review of the Syllabus,

2. **Time Series Data:** $x_1, x_2, \ldots, x_n$.

   A variable measured over time:



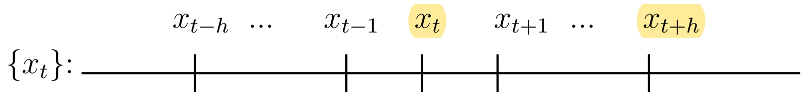
   Time Series Plot: Plot of x_t vs time should reveal some patterns over time like:

   Trend, Cycle, Variability, Dependence,

   Exploratory Data Analysis: Chapters 1-3.

   STATIONARY TIME SERIES, Chaps 4-5

   Going beyond sample data or independent and identically distributed (i.i.d.) rvs.



   Autocovariance Function (ACF) & PACF:

$$\gamma(h) = \text{cov}(x_{t+h}, x_t), \quad h = 0, 1, \ldots.$$

   Correlogram: Plot of Correlations vs Lags,

3. Autoregressive Integrated Moving Average (ARIMA) Models.

Goal(s) of Time Series Analysis: Forecasting, Detection,.....

Chapter 1

Time Series Elements

1.1 Introduction

The analysis of data observed at different time points leads to unique problems that are not covered by classical statistics. The dependence introduced by the sampling data over time restricts the applicability of many conventional statistical methods that require random samples. The analysis of such data is commonly referred to as *time series analysis*.

To provide a statistical setting for describing the elements of time series data, the data are represented as a collection of random variables indexed according to the order they are obtained in time. For example, if we collect data on daily high temperatures in your city, we may consider the time series as a sequence of random variables, $x_1, x_2, x_3, \ldots$, where the random variable x_1 denotes the high temperature on day one, the variable x_2 denotes the value for the second day, x_3 denotes the value for the third day, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a *stochastic process*. In this text, t will typically be discrete and vary over the integers $t = 0, \pm 1, \pm 2, \ldots$ or some subset of the integers, or a similar index like months of a year.

Historically, time series methods were applied to problems in the physical and environmental sciences. This fact accounts for the engineering nomenclature that permeates the language of time series analysis. The first step in an investigation of time series data involves careful scrutiny of the recorded data plotted over time. Before looking more closely at the particular statistical methods, we mention that two separate, but not mutually exclusive, approaches to time series analysis exist, commonly identified as the *time domain approach* (Chapter 4 and 5) and the *frequency domain approach* (Chapter 6 and 7).

1.2 Time Series Data

The following examples illustrate some of the common kinds of time series data as well as some of the statistical questions that might be asked about such data.

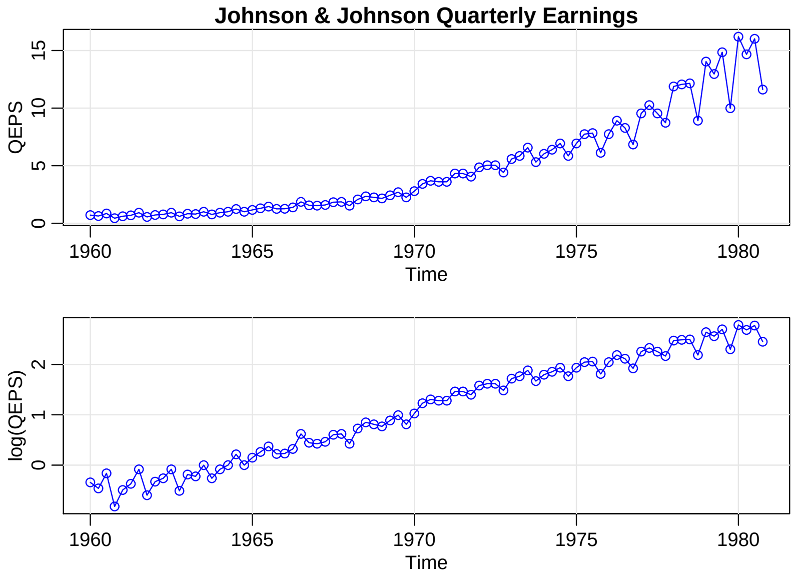


Figure 1.1 *Johnson & Johnson quarterly earnings per share, 1960-I to 1980-IV [top]. The same data logged [bottom].*

Example 1.1. Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share (QEPS) for the U.S. company Johnson & Johnson and the data transformed by taking logs. There are 84 quarters (21 years) measured from the first quarter of 1960 to the last quarter of 1980. Modeling such series begins by observing the primary patterns in the time history. In this case, note the increasing underlying trend and variability, and a somewhat regular oscillation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 3 (see Problem 3.1) using regression techniques.

If we consider the data as being generated as a small percentage change each year, say r_t (which can be negative), we might write $x_t = (1 + r_t)x_{t-4}$, where x_t is the QEPS for quarter t. If we log the data, then $\log(x_t) = \log(1 + r_t) + \log(x_{t-4})$, implying a linear growth rate; i.e., this quarter's value is the same as last year plus a small amount, $\log(1 + r_t)$. This attribute of the data is displayed by the bottom plot of Figure 1.1.

The R code to plot the data for this example is,[1]

```
library(astsa)     # we leave this line off subsequent examples
par(mfrow=2:1)
tsplot(jj, ylab="QEPS", type="o", col=4, main="Johnson & Johnson
          Quarterly Earnings")
tsplot(log(jj), ylab="log(QEPS)", type="o", col=4)
```

◊

[1] We assume astsa version 1.8.6 or later has been installed; see Section A.2

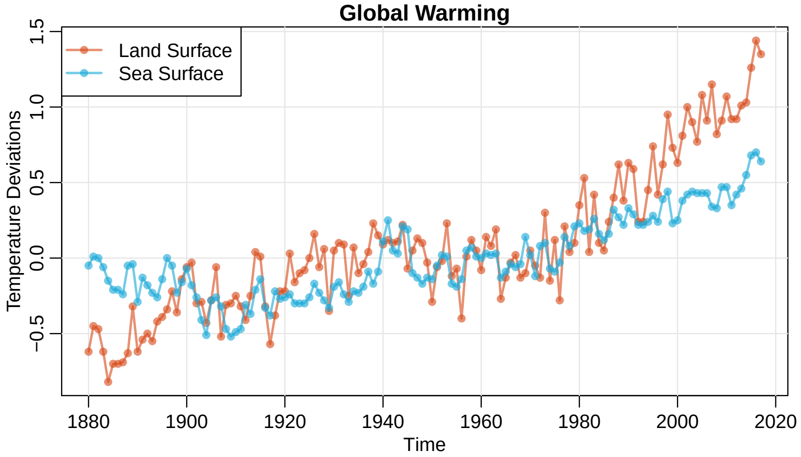


Figure 1.2 *Yearly average global land surface and ocean surface temperature deviations (1880–2017) in °C.*

Example 1.2. Global Warming and Climate Change

Two global temperature records are shown in Figure 1.2. The data are (1) annual temperature anomalies averaged over the Earth's land area, and (2) sea surface temperature anomalies averaged over the part of the ocean that is free of ice at all times (open ocean). The time period is 1880 to 2017 and the values are deviations (°C) from the 1951-1980 average, updated from Hansen et al. (2006). The upward trend in both series during the latter part of the twentieth century has been used as an argument for the climate change hypothesis. Note that the trend is not linear, with periods of leveling off and then sharp upward trends. It should be obvious that fitting a simple linear regression of the either series (x_t) on time (t), say $x_t = \alpha + \beta t + \epsilon_t$, would not yield an accurate description of the trend. Most climate scientists agree the main cause of the current global warming trend is human expansion of the *greenhouse effect*; see https://climate.nasa.gov/causes/. The R code for this example is:

```
culer = c(rgb(.85,.30,.12,.6), rgb(.12,.65,.85,.6))
tsplot(gtemp_land, col = culer[1], lwd=2, type="o", pch=20,
           ylab="Temperature Deviations", main="Global Warming")
lines(gtemp_ocean, col=culer[2], lwd=2, type="o", pch=20)
legend("topleft", col=culer, lty=1, lwd=2, pch=20, legend=c("Land
           Surface", "Sea Surface"), bg="white")
```

◊

Example 1.3. Dow Jones Industrial Average

As an example of financial time series data, Figure 1.3 shows the trading day closings and returns (or percent change) of the Dow Jones Industrial Average (DJIA) from 2006 to 2016. If x_t is the value of the DJIA closing on day t, then the return is

$$r_t = (x_t - x_{t-1})/x_{t-1} .$$

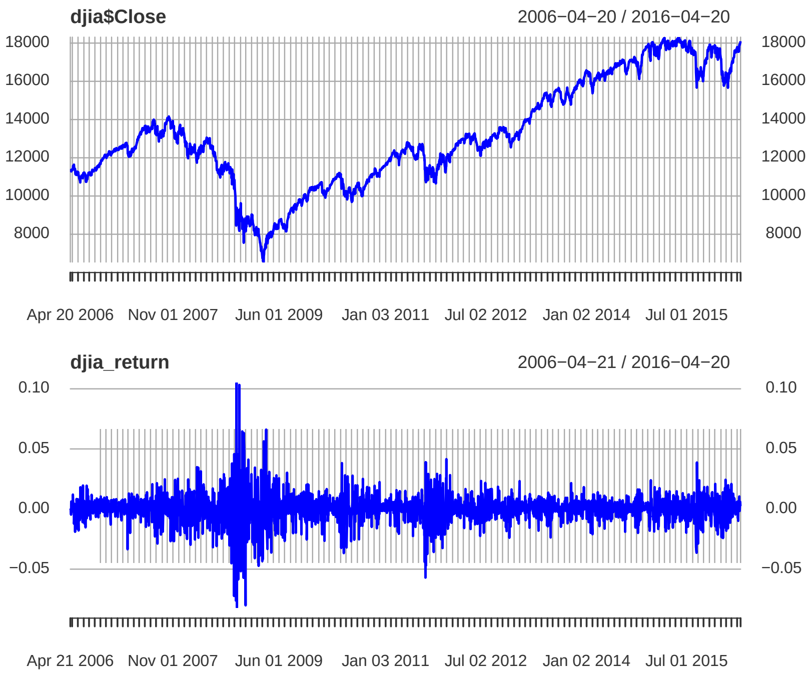


Figure 1.3 *Dow Jones Industrial Average (DJIA) trading days closings [top] and returns [bottom] from April 20, 2006 to April 20, 2016.*

This means that $1 + r_t = x_t / x_{t-1}$ and

$$\log(1 + r_t) = \log(x_t / x_{t-1}) = \log(x_t) - \log(x_{t-1}) \,,$$

just as in Example 1.1. Noting the expansion

$$\log(1 + r) = r - \tfrac{r^2}{2} + \tfrac{r^3}{3} - \cdots \qquad -1 < r \leq 1 \,,$$

we see that if r is very small, the higher order terms will be negligible. Consequently, because for financial data, $x_t / x_{t-1} \approx 1$, we have

$$\log(1 + r_t) \approx r_t \,.$$

Note the financial crisis of 2008 in Figure 1.3. The data shown are typical of return data. The mean of the series appears to be stable with an average return of approximately zero, however, the *volatility* (or variability) of data exhibits clustering; that is, highly volatile periods tend to be clustered together. A problem in the analysis of these type of financial data is to forecast the volatility of future returns. Models have been developed to handle these problems; see Chapter 8. The data set is an `xts` data file, so it must be loaded.

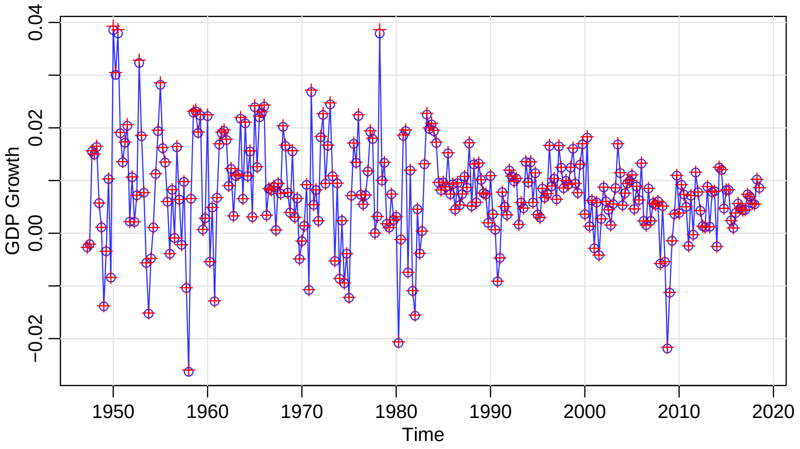


Figure 1.4: *US GDP growth rate calculated using logs (–○–) and actual values (+).*

```
library(xts)
djia_return = diff(log(djia$Close))[-1]
par(mfrow=2:1)
plot(djia$Close,  col=4)
plot(djia_return, col=4)
```

You can see a comparison of r_t and $\log(1 + r_t)$ in Figure 1.4, which shows the seasonally adjusted quarterly growth rate, r_t, of US GDP compared to the version obtained by calculating the difference of the logged data.

```
tsplot(diff(log(gdp)), type="o", col=4, ylab="GDP Growth") # diff-log
points(diff(gdp)/lag(gdp,-1), pch=3, col=2)           # actual return
```

It turns out that many time series behave like this, so that logging the data and then taking successive differences is a standard data transformation in time series analysis. ◊

Example 1.4. El Niño – Southern Oscillation (ENSO)

The Southern Oscillation Index (SOI) measures changes in air pressure related to sea surface temperatures in the central Pacific Ocean. The central Pacific warms every three to seven years due to the ENSO effect, which has been blamed for various global extreme weather events. During El Niño, pressure over the eastern and western Pacific reverses, causing the trade winds to diminish and leading to an eastward movement of warm water along the equator. As a result, the surface waters of the central and eastern Pacific warm with far-reaching consequences to weather patterns.

Figure 1.5 shows monthly values of the Southern Oscillation Index (SOI) and associated Recruitment (an index of the number of new fish). Both series are for a period of 453 months ranging over the years 1950–1987. The series show an obvious annual cycle (hot in the summer, cold in the winter), and, though difficult to see, a slower frequency of three to seven years. The study of the kinds of cycles and their

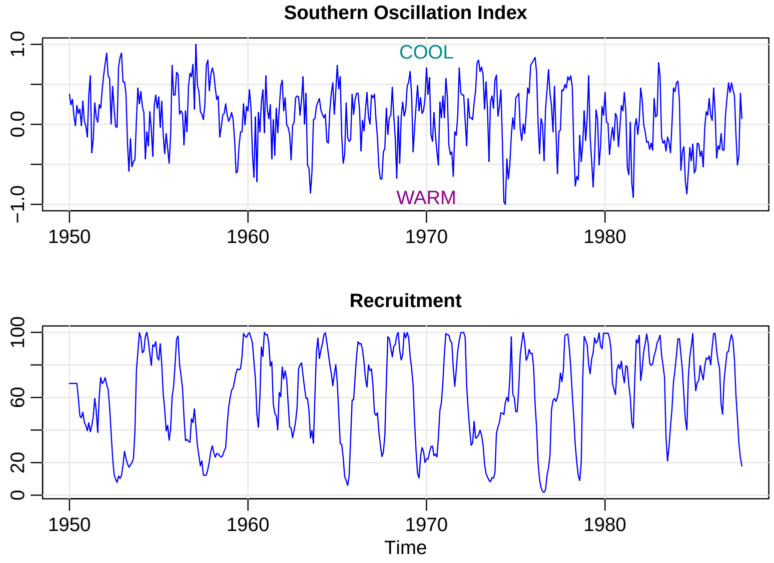


Figure 1.5: *Monthly SOI and Recruitment (estimated new fish), 1950-1987.*

strengths is the subject of Chapter 6 and 7. The two series are also related; it is easy to imagine that fish population size is dependent on the ocean temperature.

The following R code will reproduce Figure 1.5:

```
par(mfrow = c(2,1))
tsplot(soi, ylab="", xlab="", main="Southern Oscillation Index", col=4)
text(1970, .91, "COOL", col="cyan4")
text(1970,-.91, "WARM", col="darkmagenta")
tsplot(rec, ylab="", main="Recruitment", col=4)
```

◊

Example 1.5. Predator-Prey Interactions

While it is clear that predators influence the numbers of their prey, prey affect the number of predators because when prey become scarce, predators may die of starvation or fail to reproduce. Such relationships are often modeled by the Lotka–Volterra equations, which are a pair of simple nonlinear differential equations (e.g., see Edelstein-Keshet, 2005, Ch. 6).

One of the classic studies of predator-prey interactions is the snowshoe hare and lynx pelts purchased by the Hudson's Bay Company of Canada. While this is an indirect measure of predation, the assumption is that there is a direct relationship between the number of pelts collected and the number of hare and lynx in the wild. These predator-prey interactions often lead to cyclical patterns of predator and prey abundance seen in Figure 1.6. Notice that the lynx and hare population sizes are asymmetric in that they tend to increase slowly and decrease quickly (↗↓).

The lynx prey varies from small rodents to deer, with the snowshoe hare being

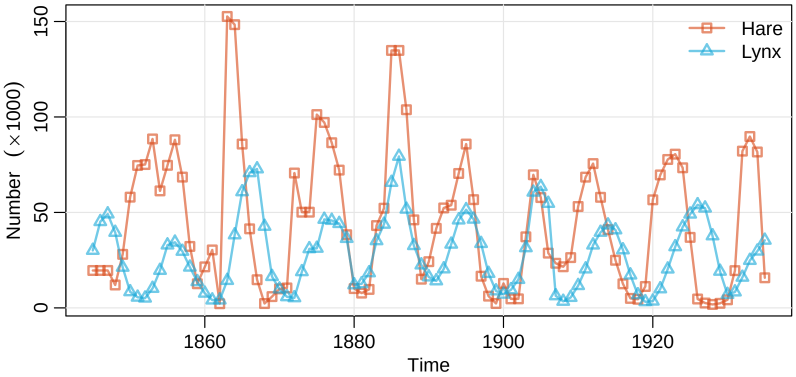


Figure 1.6 *Time series of the predator-prey interactions between the snowshoe hare and lynx pelts purchased by the Hudson's Bay Company of Canada. It is assumed there is a direct relationship between the number of pelts collected and the number of hare and lynx in the wild.*

its overwhelmingly favored prey. In fact, lynx are so closely tied to the snowshoe hare that its population rises and falls with that of the hare, even though other food sources may be abundant. In this case, it seems reasonable to model the size of the lynx population in terms of the snowshoe population. This idea is explored further in Example 5.17.

Figure 1.6 may be reproduced as follows.

```
culer = c(rgb(.85,.30,.12,.6), rgb(.12,.67,.86,.6))
tsplot(Hare, col = culer[1], lwd=2, type="o", pch=0,
          ylab=expression(Number~~~(""%*% 1000)))
lines(Lynx, col=culer[2], lwd=2, type="o", pch=2)
legend("topright", col=culer, lty=1, lwd=2, pch=c(0,2),
          legend=c("Hare", "Lynx"), bty="n")
```

◊

Example 1.6. fMRI Imaging

Often, time series are observed under varying experimental conditions or treatment configurations. Such a set of series is shown in Figure 1.7, where data are collected from various locations in the brain via functional magnetic resonance imaging (fMRI).

In fMRI, subjects are put into an MRI scanner and a stimulus is applied for a period of time, and then stopped. This on-off application of a stimulus is repeated and recorded by measuring the blood oxygenation-level dependent (BOLD) signal intensity, which measures areas of activation in the brain. The BOLD contrast results from changing regional blood concentrations of oxy- and deoxy- hemoglobin.

The data displayed in Figure 1.7 are from an experiment that used fMRI to examine the effects of general anesthesia on pain perception by comparing results from anesthetized volunteers while a supramaximal shock stimulus was applied. This stimulus was used to simulate surgical incision without inflicting tissue damage. In

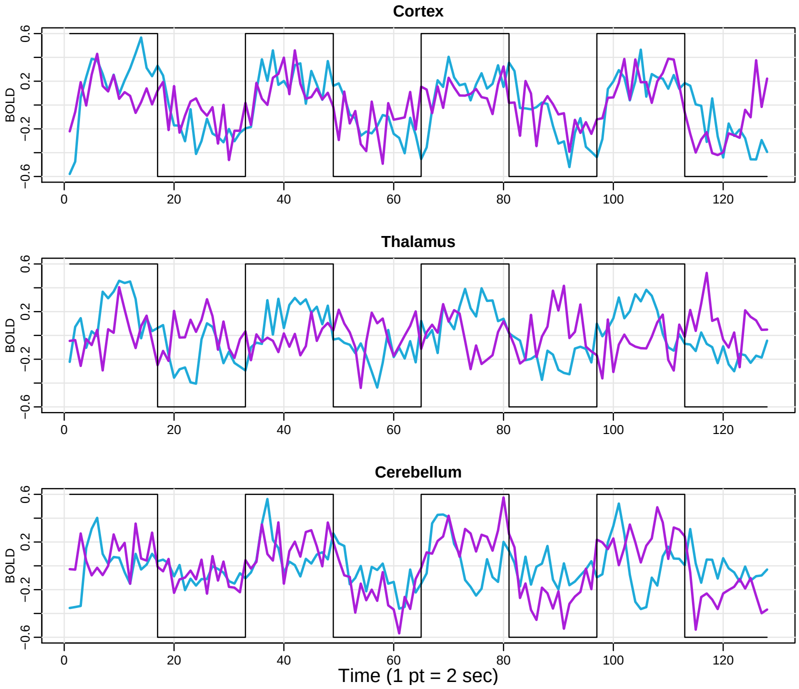


Figure 1.7 *fMRI data from two locations in the cortex, the thalamus, and the cerebellum;* $n = 128$ *points, one observation taken every 2 seconds. The boxed line represents the presence or absence of the stimulus.*

this example, the stimulus was applied for 32 seconds and then stopped for 32 seconds, so that the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$).

Notice that the periodicities appear strongly in the motor cortex series but seem to be missing in the thalamus and perhaps in the cerebellum. In this case, it is of interest to statistically determine if the areas in the thalamus and cerebellum are actually responding to the stimulus. Use the following R commands for the graphic:

```
par(mfrow=c(3,1))
culer = c(rgb(.12,.67,.85,.7), rgb(.67,.12,.85,.7))
u = rep(c(rep(.6,16), rep(-.6,16)), 4)    # stimulus signal
tsplot(fmri1[,4], ylab="BOLD", xlab="", main="Cortex", col=culer[1],
           ylim=c(-.6,.6), lwd=2)
lines(fmri1[,5], col=culer[2], lwd=2)
lines(u, type="s")
tsplot(fmri1[,6], ylab="BOLD", xlab="", main="Thalamus", col=culer[1],
           ylim=c(-.6,.6), lwd=2)
lines(fmri1[,7], col=culer[2], lwd=2)
lines(u, type="s")
```

```
tsplot(fmri1[,8], ylab="BOLD", xlab="", main="Cerebellum",
          col=culer[1], ylim=c(-.6,.6), lwd=2)
lines(fmri1[,9], col=culer[2], lwd=2)
lines(u, type="s")
mtext("Time (1 pt = 2 sec)", side=1, line=1.75)
```

◊

1.3 Time Series Models

The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data, like that encountered in the previous section.

The fundamental visual characteristic distinguishing the different series shown in Example 1.1 – Example 1.6 is their differing degrees of smoothness. A parsimonious explanation for this smoothness is that adjacent points in time are correlated, so the value of the series at time t, say, x_t, depends in some way on the past values $x_{t-1}, x_{t-2}, \ldots$. This idea expresses a fundamental way in which we might think about generating realistic looking time series.

Example 1.7. White Noise
A simple kind of generated series might be a collection of *uncorrelated* random variables, w_t, with mean 0 and finite variance σ_w^2. The time series generated from uncorrelated variables is used as a model for noise in engineering applications where it is called *white noise*; we shall sometimes denote this process as $w_t \sim wn(0, \sigma_w^2)$. The designation white originates from the analogy with white light (details in Chapter 6). A special version of white noise that we use is when the variables are independent and identically distributed normals, written $w_t \sim \text{iid N}(0, \sigma_w^2)$.

The upper panel of Figure 1.8 shows a collection of 500 independent standard normal random variables ($\sigma_w^2 = 1$), plotted in the order in which they were drawn. The resulting series bears a resemblance to portions of the DJIA returns in Figure 1.3. ◊

If the stochastic behavior of all time series could be explained in terms of the white noise model, classical statistical methods would suffice. Two ways of introducing serial correlation and more smoothness into time series models are given in Example 1.8 and Example 1.9.

Example 1.8. Moving Averages, Smoothing and Filtering
We might replace the white noise series w_t by a moving average that smoothes the series. For example, consider replacing w_t in Example 1.7 by an average of its current value and its immediate two neighbors in the past. That is, let

new time series
value at time t

$$v_t = \tfrac{1}{3}\big(w_{t-1} + w_t + w_{t+1}\big), \tag{1.1}$$

which leads to the series shown in the lower panel of Figure 1.8. This series is much smoother than the white noise series and has a smaller variance due to averaging. It should also apparent that averaging removes some of the high frequency (fast

which smoothenes the graph, by doing linear
combination of curent, pre,next -->sliding sum.
As we assumed white noise is uncorrelated, variance
of sum becomes sum of variances and we see that
varaicne becomes by 1/3. -- see notes.

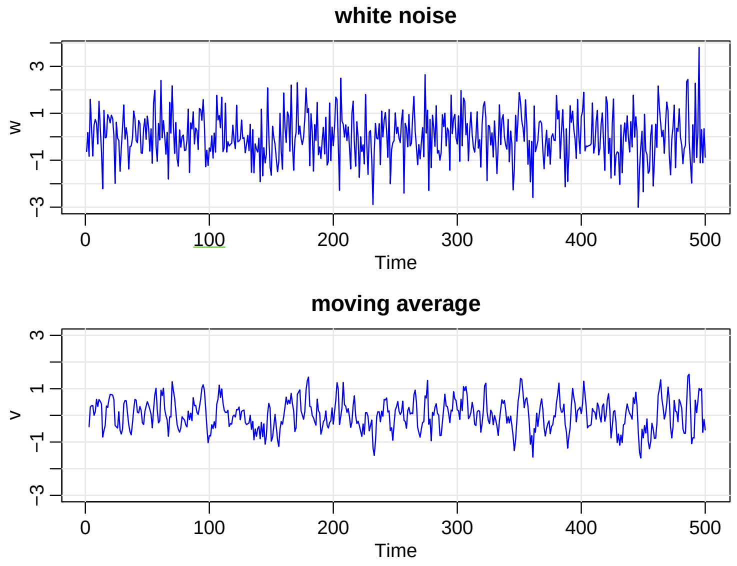


Figure 1.8 *Gaussian white noise series (top) and three-point moving average of the Gaussian white noise series (bottom).*

oscillations) behavior of the noise. We begin to notice a similarity to some of the non-cyclic fMRI series in Figure 1.7.

A linear combination of values in a time series such as in (1.1) is referred to, generically, as a filtered series; hence the command `filter`. To reproduce Figure 1.8:

```
par(mfrow=2:1)
w = rnorm(500)                            # 500 N(0,1) variates
v = filter(w, sides=2, filter=rep(1/3,3))  # moving average
tsplot(w, col=4, main="white noise")
tsplot(v, ylim=c(-3,3), col=4, main="white noise")
```

◊

The SOI and Recruitment series in Figure 1.5, as well as some of the fMRI series in Figure 1.7, differ from the moving average series because they are dominated by an oscillatory behavior. A number of methods exist for generating series with this quasi-periodic behavior; we illustrate a popular one based on the autoregressive model considered in Chapter 4.

during prediction, roughly remove the error - wt, like Linear regression.

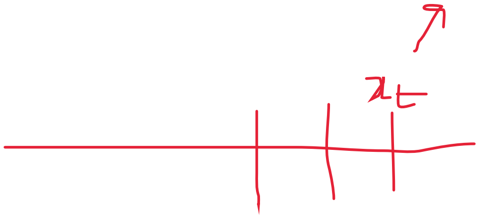
Example 1.9. Autoregressions

Suppose we consider the white noise series w_t of Example 1.7 as input and calculate the output using the second-order equation

$$x_t = 1.5x_{t-1} - .75x_{t-2} + w_t \tag{1.2}$$

how many x should we bring in ?
last 2 days?
3 days? 1 year?

successively for $t = 1, 2, \ldots, 250$. The resulting output series is shown in Figure 1.9. Equation (1.2) represents a regression or prediction of the current value x_t of a



coefficients ?

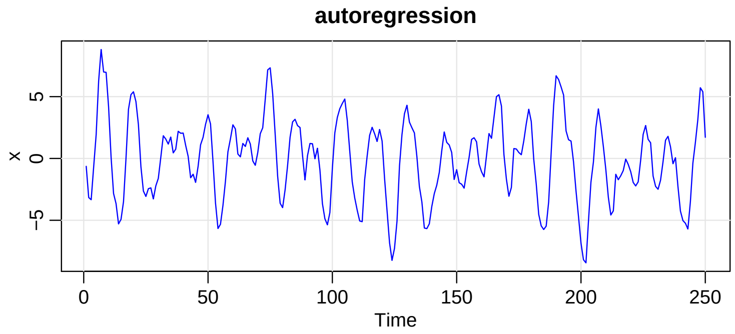


Figure 1.9: *Autoregressive series generated from model* (1.2).

time series as a function of the past two values of the series, and, hence, the term autoregression is suggested for this model. A problem with startup values exists here because (1.2) also depends on the initial conditions x_0 and x_{-1}, but for now we set them to zero. We can then generate data *recursively* by substituting into (1.2). That is, given $w_1, w_2, \dots, w_{250}$, we could set $x_{-1} = x_0 = 0$ and then start at $t = 1$:

$$\begin{aligned}
x_1 &= 1.5x_0 - .75x_{-1} + w_1 = w_1 \\
x_2 &= 1.5x_1 - .75x_0 + w_2 \quad = 1.5w_1 + w_2 \\
x_3 &= 1.5x_2 - .75x_1 + w_3 \\
x_4 &= 1.5x_3 - .75x_2 + w_4
\end{aligned}$$

and so on. We note the approximate periodic behavior of the series, which is similar to that displayed by the SOI and Recruitment in Figure 1.5 and some fMRI series in Figure 1.7. This particular model is chosen so that the data have pseudo-cyclic behavior of about 1 cycle every 12 points; thus 250 observations should contain about 20 cycles. This autoregressive model and its generalizations can be used as an underlying model for many observed series and will be studied in detail in Chapter 4.

One way to simulate and plot data from the model (1.2) in R is to use the following commands. The initial conditions are set equal to zero so we let the filter run an extra 50 values to avoid startup problems.

```
set.seed(90210)
w = rnorm(250 + 50)   # 50 extra to avoid startup problems
x = filter(w, filter=c(1.5,-.75), method="recursive")[-(1:50)]
tsplot(x, main="autoregression", col=4)
```

◊

Example 1.10. Random Walk with Drift

A model for analyzing trend such as seen in the global temperature data in Figure 1.2, is the random walk with drift model given by

$$x_t = \delta + x_{t-1} + w_t \tag{1.3}$$

Delta=0 it becomes a auto regressive model.

variance increases for randomwalk , see notes.

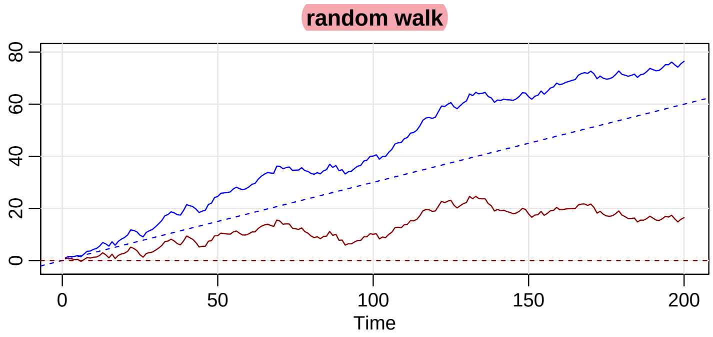


Figure 1.10 *Random walk, $\sigma_w = 1$, with drift $\delta = .3$ (upper jagged line), without drift, $\delta = 0$ (lower jagged line), and dashed lines showing the drifts.*

delta is negative, trend downward. and other.

for $t = 1, 2, \ldots$, with initial condition $x_0 = 0$, and where w_t is white noise. The constant δ is called the drift, and when $\delta = 0$, the model is called simply a random walk because the value of the time series at time t is the value of the series at time $t-1$ plus a completely random movement determined by w_t. Note that we may rewrite (1.3) as a cumulative sum of white noise variates. That is,

$$x_t = \delta\, t + \sum_{j=1}^{t} w_j \tag{1.4}$$

random walk is cummulative summation of white noise (linear combination of white noise). + drift

for $t = 1, 2, \ldots$; either use induction, or plug (1.4) into (1.3) to verify this statement. Figure 1.10 shows 200 observations generated from the model with $\delta = 0$ and .3, and with standard normal nose. For comparison, we also superimposed the straight lines δt on the graph. To reproduce Figure 1.10 in R use the following code (notice the use of multiple commands per line using a semicolon).

```
set.seed(314159265)       # so you can reproduce the results
w  = rnorm(200);  x  = cumsum(w)    # random walk
wd = w +.3;       xd = cumsum(wd)  # random walk with drift
tsplot(xd, ylim=c(-2,80), main="random walk", ylab="", col=4)
abline(a=0, b=.3, lty=2, col=4)     # plot drift
lines(x, col="darkred")
abline(h=0, col="darkred", lty=2)
```

◊

Example 1.11. Signal Plus Noise

Many realistic models for generating time series assume an underlying signal with some consistent periodic variation contaminated by noise. For example, it is easy to detect the regular cycle fMRI series displayed on the top of Figure 1.7. Consider the model

$$x_t = 2\cos(2\pi \tfrac{t+15}{50}) + w_t \tag{1.5}$$

for $t = 1, 2, \ldots, 500$, where the first term is regarded as the signal, shown in the

non linear regression.

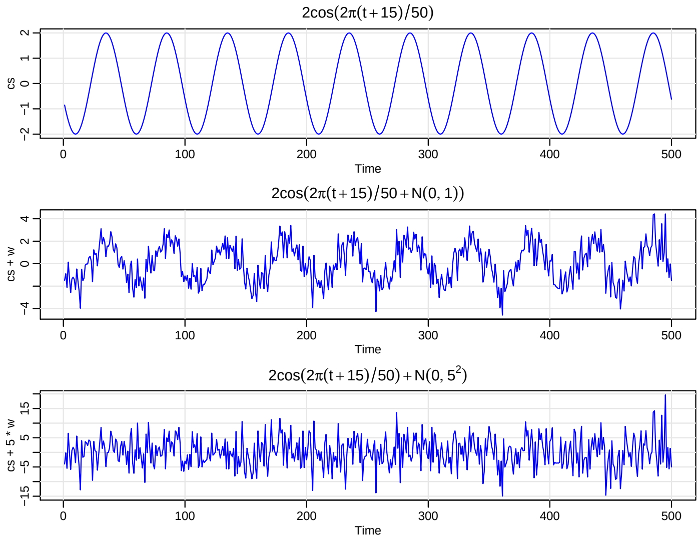


Figure 1.11 *Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise,* $\sigma_w = 1$ *(middle panel) and* $\sigma_w = 5$ *(bottom panel); see* (1.5).

upper panel of Figure 1.11. We note that a sinusoidal waveform can be written as

$$A\cos(2\pi\omega t + \phi), \tag{1.6}$$

where A is the amplitude, ω is the frequency of oscillation, and ϕ is a phase shift. In (1.5), $A = 2$, $\omega = 1/50$ (one cycle every 50 time points), and $\phi = .6\pi$.

An additive noise term was taken to be white noise with $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel), drawn from a normal distribution. Adding the two together obscures the signal, as shown in the lower panels of Figure 1.11. The degree to which the signal is obscured depends on the amplitude of the signal relative to the size of σ_w. The ratio of the amplitude of the signal to σ_w (or some function of the ratio) is sometimes called the *signal-to-noise ratio (SNR)*; the larger the SNR, the easier it is to detect the signal. Note that the signal is easily discernible in the middle panel, whereas the signal is obscured in the bottom panel. Typically, we will not observe the signal but the signal obscured by noise.

To reproduce Figure 1.11 in R, use the following commands:

```
t  = 1:500
cs = 2*cos(2*pi*(t+15)/50)    # signal
w  = rnorm(500)               # noise
par(mfrow=c(3,1))
tsplot(cs,    col=4, main=expression(2*cos(2*pi*(t+15)/50)))
tsplot(cs+w,  col=4, main=expression(2*cos(2*pi*(t+15)/50+N(0,1))))
tsplot(cs+5*w,col=4, main=expression(2*cos(2*pi*(t+15)/50)+N(0,5^2)))
```

◊

Problems

1.1. (a) Generate $n = 100$ observations from the autoregression

$$x_t = -.9x_{t-2} + w_t$$

with $\sigma_w = 1$, using the method described in Example 1.9. Next, apply the moving average filter

$$v_t = (x_t + x_{t-1} + x_{t-2} + x_{t-3})/4$$

to x_t, the data you generated. Now plot x_t as a line and superimpose v_t as a dashed line.

(b) Repeat (a) but with

$$x_t = 2\cos(2\pi t/4) + w_t,$$

where $w_t \sim \text{iid N}(0, 1)$.

(c) Repeat (a) but where x_t is the log of the Johnson & Johnson data discussed in Example 1.1.

(d) What is seasonal adjustment (you can do an internet search)?

(e) State your conclusions (in other words, what did you learn from this exercise).

1.2. There are a number of seismic recordings from earthquakes and from mining explosions in `astsa`. All of the data are in the dataframe `eqexp`, but two specific recording are in `EQ5` and `EXP6`, the fifth earthquake and the sixth explosion, respectively. The data represent two phases or arrivals along the surface, denoted by P $(t = 1, \dots, 1024)$ and S $(t = 1025, \dots, 2048)$, at a seismic recording station. The recording instruments are in Scandinavia and monitor a Russian nuclear testing site. The general problem of interest is in distinguishing between these waveforms in order to maintain a comprehensive nuclear test ban treaty.

To compare the earthquake and explosion signals,

(a) Plot the two series separately in a multifigure plot with two rows and one column.

(b) Plot the two series on the same graph using different colors or different line types.

(c) In what way are the earthquake and explosion series different?

1.3. In this problem, we explore the difference between random walk and moving average models.

(a) Generate and (multifigure) plot *nine* series that are random walks (see Example 1.10) of length $n = 500$ without drift ($\delta = 0$) and $\sigma_w = 1$.

(b) Generate and (multifigure) plot *nine* series of length $n = 500$ that are moving averages of the form (1.1) discussed in Example 1.8.

(c) Comment on the differences between the results of part (a) and part (b).

1.4. The data in `gdp` are the seasonally adjusted quarterly U.S. GDP from 1947-I to 2018-III. The growth rate is shown in Figure 1.4.

Doubt

(a) Plot the data and compare it to one of the models discussed in Section 1.3.

(b) Reproduce Figure 1.4 using your colors and plot characters (`pch`) of your own choice. Then, comment on the difference between the two methods of calculating growth rate.

(c) Which of the models discussed in Section 1.3 best describe the behavior of the growth in U.S. GDP?