STAT 626: Outline of Lecture 1

1. **A Quick Review of the Syllabus**

2. **Time Series Data:**
   Anything measured regularly over time is a time series.

3. **Examples of Time Series Data**

4. **Goals of Time Series Analysis**

   Forecasting

   Description or dynamic of the data

SYLLABUS: Summer, 2016
STATISTICS 626: Methods in Time Series Analysis

IMPORTANT: Please note that the course meets during May 31- July 27. There will be one additional meeting during each of the first three weeks to account for the missed week of May 23.

INSTRUCTOR: Mohsen Pourahmadi
TIME & PLACE: MTW 2:15-3:30, Blocker 457
OFFICE HOURS: TuW 3:40-4:30 or by Appointment
OFFICE: 439 Blocker
PH. & FAX NOs: 979-845-3164, 979-845-3144
E-MAIL: pourahm@stat.tamu.edu

Teaching Assistant: Raanju Ragavendar (Blocker 406E, raanju@stat.tamu.edu)
Office Hours: MTu 1-2 pm.

TEXT: Time Series Analysis and Its Applications with R Examples, Shumway, R. and Stoffer, D., 2011, Third ed. Springer. **An electronic version is available through TAMU library free of charge.**

PREREQUISITE: STAT 601 or 642 and a working knowledge of basic math such as complex numbers 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. 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 introducing these concepts 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 (local, distance, stat majors, non-stat majors,...) of about three-five who are interested in similar/complementary application areas, develop a project plan and identify the relevant research papers and dataset to study, analyze and present at various 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 look at the data examples in Chap. 1 of the text ASAP.

GRADE POLICY:

1. **Exams**: There will be two midterms and no final. The midterms will constitute 25% and 35% of the grade, respectively.

2. **Homework**: Will be assigned regularly and posted on eCampus, it will contribute 10% to your grade. 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.**

3. **Missed assignments**: Each homework must be turned in by 2:15pm 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.

4. **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 presentations in class by local students (**there must be a local student in each group**). The project starts by choosing a suitable time series dataset, the first written report could be a page long describing the data and application area, and the first oral presentation will be about five minutes long describing the data, etc. Subsequent project presentations may take 15 and 30 minutes, respectively. The project is worth 30% of your grade. The project reports should be organized and typed following the format of a research article in statistics or your area of applications, the 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 presentation in class will contribute greatly to the grade for this part of the course.

5. 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 89 percent will be a B, etc.

6. Attendance and classroom participation are encouraged and will be rewarded, they are integral parts of the learning process .

A Tentative Course Outline: The dates and topics 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, Appendix R
	Stationary TS, Linear Processes	1.5
2	Estimation of autocorrelation	1.6
	Bivariate TS and cross-correlation	1.6
	A Review of Regression: $y = X\beta + e$.	2.1-2.2
3	Exploratory Data Analysis	2.3-2.4
	Periodogram, 5-min Project Presentation	4.2
	TS Models II: AR(1), MA	3.2
4	TS Models II: ARMA, $\phi(B)x_t = \theta(B)w_t$.	3.2
	ARMA : Invertibility& Causality (Roots $\|z\| > 1$,$\pi-$, $\psi-$) weights)	
	Difference Equations	3.3
	ARMA ACF and PACF	3.4
5	**EXAM I, June 22**	
	(Monday July 4th- Holiday) ARMA Forecasting	3.5
6	Estimation, Yule-Walker Eqs., MLE	3.6
	Building ARIMA Models	3.6-3.8
	SARIMA Models, (PPR, 15-min.)	3.9
7	Spectral Analysis, Filter Theorem	4.3-4.5
	FARIMA, Long-memory	5.2
8	Unit-Root Testing	5.3
	GARCH Models	5.4
	Multivariate ARMAX Models	5.8
	Multivariate ARMAX Models	5.8
9	Spurious Reg. and Cointegration	
	EXAM II, July 20	
	Final PPR. (Written report due.)	
10	Final PPR. (Written report due.)	
	Final PPR. (Written report due.)	

7. 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.

8. 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 person and turn it in as your own, even if you should have the permission of that person. Plagiarism is one of the worst academic sins, for the plagiarist destroys the trust among colleagues without which research cannot be safely communicated. If you have any questions regarding plagiarism, please consult the latest issue of the Texas A&M University Student Rules, under the section "Scholastic Dishonesty."

9. STATEMENT ON DISABILITIES: The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation for their disabilities. If you believe you have a disability requiring an accommodation, please contact the Office of Disability Services in Room B118 of Cain Hall. The phone number is 845-1637.

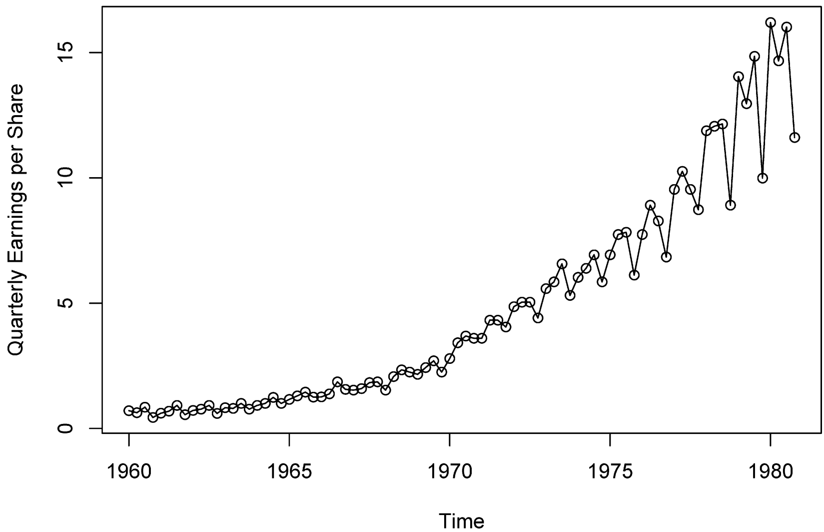

Fig. 1.1. Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

from different subject areas. The following cases illustrate some of the common kinds of experimental time series data as well as some of the statistical questions that might be asked about such data.

Example 1.1 Johnson & Johnson Quarterly Earnings

Figure 1.1 shows quarterly earnings per share for the U.S. company Johnson & Johnson, ~~furnished by Professor Paul Griffin (personal communication) of the Graduate School of Management, University of California, Davis~~. 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 gradually increasing underlying trend and the rather regular variation superimposed on the trend that seems to repeat over quarters. Methods for analyzing data such as these are explored in Chapter 2 (see Problem 2.1) using regression techniques and in Chapter 6, §6.5, using structural equation modeling.

To plot the data using the R statistical package, type the following:[1]

```
load("tsa3.rda")     # SEE THE FOOTNOTE
plot(jj, type="o", ylab="Quarterly Earnings per Share")
```

Example 1.2 Global Warming

Consider the global temperature series record shown in Figure 1.2. The data are the global mean land–ocean temperature index from 1880 to 2009, with

[1] We assume that `tsa3.rda` has been downloaded to a convenient directory. See Appendix R for further details.

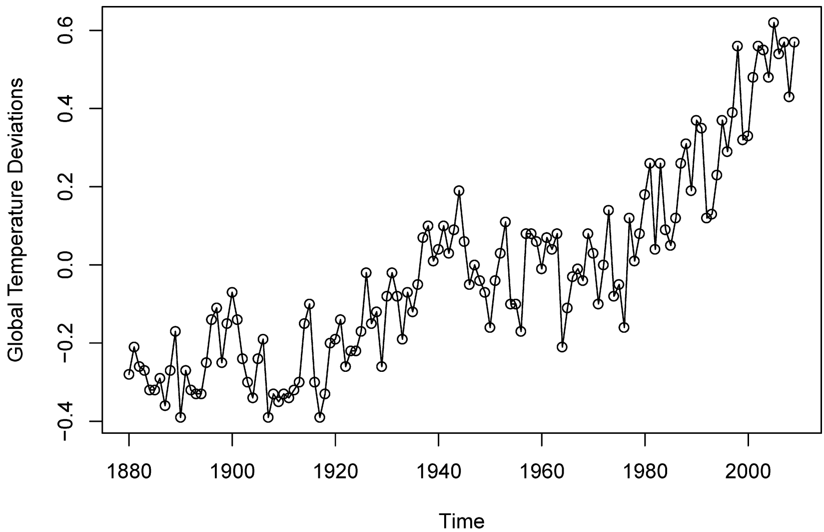

Fig. 1.2. Yearly average global temperature deviations (1880–2009) in degrees centigrade.

the base period 1951-1980. In particular, the data are deviations, measured in degrees centigrade, from the 1951-1980 average, and are an update of Hansen et al. (2006). We note an apparent upward trend in the series during the latter part of the twentieth century that has been used as an argument for the global warming hypothesis. Note also the leveling off at about 1935 and then another rather sharp upward trend at about 1970. The question of interest for global warming proponents and opponents is whether the overall trend is natural or whether it is caused by some human-induced interface. Problem 2.8 examines 634 years of glacial sediment data that might be taken as a long-term temperature proxy. Such percentage changes in temperature do not seem to be unusual over a time period of 100 years. Again, the question of trend is of more interest than particular periodicities.

The R code for this example is similar to the code in Example 1.1:

```
plot(gtemp, type="o", ylab="Global Temperature Deviations")
```

~~Example 1.3 Speech Data~~

~~More involved questions develop in applications to the physical sciences. Figure 1.3 shows a small .1 second (1000 point) sample of recorded speech for the phrase $aaa \cdots hhh$, and we note the repetitive nature of the signal and the rather regular periodicities. One current problem of great interest is computer recognition of speech, which would require converting this particular signal into the recorded phrase $aaa \cdots hhh$. Spectral analysis can be used in this context to produce a signature of this phrase that can be compared with signatures of various library syllables to look for a match~~.

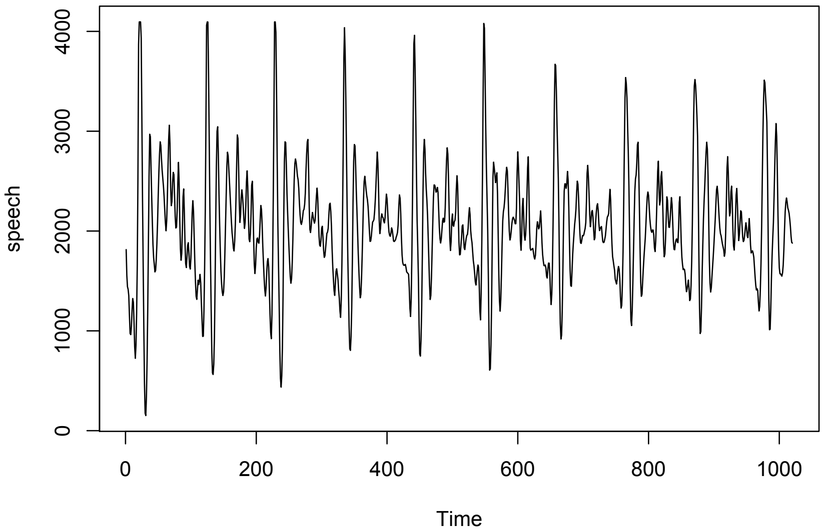

Fig. 1.3. Speech recording of the syllable $aaa \cdots hhh$ sampled at 10,000 points per second with $n = 1020$ points.

~~One can immediately notice the rather regular repetition of small wavelets. The separation between the packets is known as the pitch period and represents the response of the vocal tract filter to a periodic sequence of pulses stimulated by the opening and closing of the glottis.~~

In R, you can reproduce Figure 1.3 as follows:

```
plot(speech)
```

Example 1.4 New York Stock Exchange

As an example of financial time series data, Figure 1.4 shows the daily returns (or percent change) of the New York Stock Exchange (NYSE) from February 2, 1984 to December 31, 1991. It is easy to spot the crash of October 19, 1987 in the figure. The data shown in Figure 1.4 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 changes over time. In fact, the data show volatility 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 such as ARCH and GARCH models (Engle, 1982; Bollerslev, 1986) and stochastic volatility models (Harvey, Ruiz and Shephard, 1994) have been developed to handle these problems. We will discuss these models and the analysis of financial data in Chapters 5 and 6. The R code for this example is similar to the previous examples:

```
plot(nyse, ylab="NYSE Returns")
```

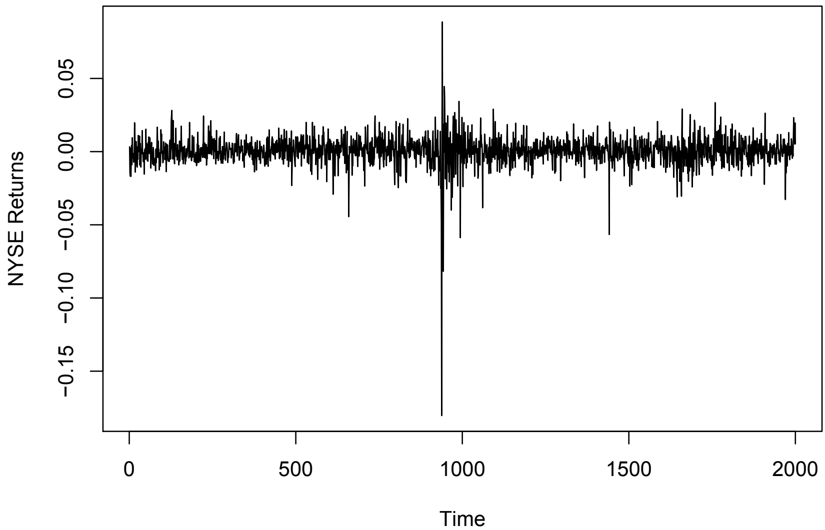

Fig. 1.4. Returns of the NYSE. The data are daily value weighted market returns from February 2, 1984 to December 31, 1991 (2000 trading days). The crash of October 19, 1987 occurs at $t = 938$.

Example 1.5 El Niño and Fish Population

We may also be interested in analyzing several time series at once. Figure 1.5 shows monthly values of an environmental series called the Southern Oscillation Index (SOI) and associated Recruitment (number of new fish) ~~furnished by Dr. Roy Mendelssohn of the Pacific Environmental Fisheries Group (personal communication).~~ Both series are for a period of 453 months ranging over the years 1950–1987. The 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 El Niño effect, which has been blamed, in particular, for the 1997 floods in the midwestern portions of the United States. Both series in Figure 1.5 tend to exhibit repetitive behavior, with regularly repeating cycles that are easily visible. This periodic behavior is of interest because underlying processes of interest may be regular and the rate or frequency of oscillation characterizing the behavior of the underlying series would help to identify them. One can also remark that the cycles of the SOI are repeating at a faster rate than those of the Recruitment series. The Recruitment series also shows several kinds of oscillations, a faster frequency that seems to repeat about every 12 months and a slower frequency that seems to repeat about every 50 months. The study of the kinds of cycles and their strengths is the subject of Chapter 4. The two series also tend to be somewhat related; it is easy to imagine that somehow the fish population is dependent on the SOI. Perhaps even a lagged relation exists, with the SOI signaling changes in the fish population. This possibility

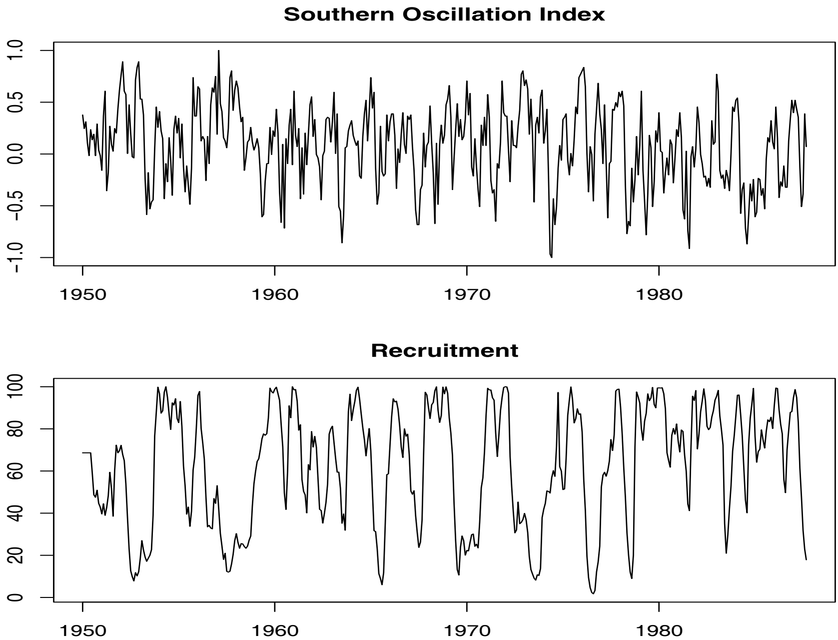

Fig. 1.5. Monthly SOI and Recruitment (estimated new fish), 1950-1987.

suggests trying some version of regression analysis as a procedure for relating the two series. Transfer function modeling, as considered in Chapter 5, can be applied in this case to obtain a model relating Recruitment to its own past and the past values of the SOI.

The following R code will reproduce Figure 1.5:

```
par(mfrow = c(2,1))  # set up the graphics
plot(soi, ylab="", xlab="", main="Southern Oscillation Index")
plot(rec, ylab="", xlab="", main="Recruitment")
```

~~**Example 1.6 fMRI Imaging**~~

~~A fundamental problem in classical statistics occurs when we are given a collection of independent series or vectors of series, generated under varying experimental conditions or treatment configurations. Such a set of series is shown in Figure 1.6, where we observe data collected from various locations in the brain via functional magnetic resonance imaging (fMRI). In this example, five subjects were given periodic brushing on the hand. The stimulus was applied for 32 seconds and then stopped for 32 seconds; thus, the signal period is 64 seconds. The sampling rate was one observation every 2 seconds for 256 seconds ($n = 128$). For this example, we averaged the results over subjects (these were evoked responses, and all subjects were in phase). The~~

1.3 Time Series Statistical 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. In order to provide a statistical setting for describing the character of data that seemingly fluctuate in a random fashion over time, we assume a time series can be defined as a collection of random variables indexed according to the order they are obtained in time. For example, we may consider a time series as a sequence of random variables, $x_1, x_2, x_3, \ldots$, where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, 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. The observed values of a stochastic process are referred to as a realization of the stochastic process. Because it will be clear from the context of our discussions, we use the term time series whether we are referring generically to the process or to a particular realization and make no notational distinction between the two concepts.

It is conventional to display a sample time series graphically by plotting the values of the random variables on the vertical axis, or ordinate, with the time scale as the abscissa. It is usually convenient to connect the values at adjacent time periods to reconstruct visually some original hypothetical continuous time series that might have produced these values as a discrete sample. Many of [illegible] the previous section, for example, could have been [illegible]s point in time and are conceptually more properly treated as continuous time series. The approximation of these series by discrete time parameter series sampled at equally spaced points in time is simply an acknowledgment that sampled data will, for the most part, be discrete because of restrictions inherent in the method of collection. Furthermore, the analysis techniques are then feasible using computers, which are limited to digital computations. Theoretical developments also rest on the idea that a continuous parameter time series should be specified in terms of finite-dimensional distribution functions defined over a finite number of points in time. This is not to say that the selection of the sampling interval or rate is not an extremely important consideration. The appearance of data can be changed completely by adopting an insufficient sampling rate. We have all seen wagon wheels in movies appear to be turning backwards because of the insufficient number of frames sampled by the camera. This phenomenon leads to a distortion called aliasing (see §4.2).

Always plot the data.

The fundamental visual characteristic distinguishing the different series shown in Examples 1.1–1.7 is their differing degrees of smoothness. One possible explanation for this smoothness is that it is being induced by the supposition 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 model expresses a fundamental way in which we might think about generating realistic-looking time series. To begin to develop an approach to using collections of random variables to model time series, consider Example 1.8.

Example 1.8 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 and indicates that all possible periodic oscillations are present with equal strength.

We will, at times, also require the noise to be independent and identically distributed (iid) random variables with mean 0 and variance σ_w^2. We shall distinguish this case by saying white independent noise, or by writing $w_t \sim$ iid$(0, \sigma_w^2)$. A particularly useful white noise series is Gaussian white noise, wherein the w_t are independent normal random variables, with mean 0 and variance σ_w^2; or more succinctly, $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$. Figure 1.8 shows in the upper panel a collection of 500 such random variables, with $\sigma_w^2 = 1$, plotted in the order in which they were drawn. The resulting series bears a slight resemblance to the explosion in Figure 1.7 but is not smooth enough to serve as a plausible model for any of the other experimental series. The plot tends to show visually a mixture of many different kinds of oscillations in

WN: The ideal regression residuals.

WN: The building blocks of TS.

vior of all tim ed in terms of the white noise model, classical statis ice. Two ways of introducing serial correlation and more smoothness into time series models are given in Examples 1.9 and 1.10.

Example 1.9 Moving Averages

We might replace the white noise series w_t by a moving average that smooths the series. For example, consider replacing w_t in Example 1.8 by an average of its current value and its immediate neighbors in the past and future. That is, let

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

which leads to the series shown in the lower panel of Figure 1.8. Inspecting the series shows a smoother version of the first series, reflecting the fact that the slower oscillations are more apparent and some of the faster oscillations are taken out. We begin to notice a similarity to the SOI in Figure 1.5, or perhaps, to some of the fMRI series in Figure 1.6.

To reproduce Figure 1.8 in R use the following commands. 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`.

~~**2.29** (Random cosine wave extended further) Suppose that~~

$$Y_t = \sum_{j=1}^{m} R_j \cos[2\pi(f_j t + \Phi_j)] \qquad \text{for } t = 0, \pm1, \pm2, \ldots$$

~~where $0 < f_1 < f_2 < \cdots < f_m < ½$ are m fixed frequencies, and $R_1, \Phi_1, R_2, \Phi_2, \ldots, R_m, \Phi_m$ are uncorrelated random variables with each Φ_j uniformly distributed on the interval (0,1).~~

~~**(a)** Show that $E(Y_t) = 0$ for all t.~~

~~**(b)** Show that the process is stationary with~~ $\gamma_k = \frac{1}{2}\sum_{j=1}^{m} E(R_j^2)\cos(2\pi f_j k)$ ~~.~~
~~Hint: Do Exercise 2.28 first.~~

~~**2.30** (Mathematical statistics required) Suppose that~~

$$Y_t = R\cos[2\pi(ft + \Phi)] \qquad \text{for } t = 0, \pm1, \pm2, \ldots$$

~~where R and Φ are independent random variables and f is a fixed frequency. The phase Φ is assumed to be uniformly distributed on (0,1), and the amplitude R has a Rayleigh distribution with pdf~~ $f(r) = re^{-r^2/2}$ ~~for $r > 0$. Show that for each time point t, Y_t has a normal distribution. (Hint: Let~~ $Y = R\cos[2\pi(ft + \Phi)]$ ~~and $X =$~~ $R\sin[2\pi(ft + \Phi)]$~~. Now find the joint distribution of X and Y. It can also be shown that all of the finite dimensional distributions are multivariate normal and hence the process is strictly stationary.)~~

Appendix A: Expectation, Variance, Covariance, and Correlation

In this appendix, we define expectation for continuous random variables. However, all of the properties described hold for all types of random variables, discrete, continuous, or otherwise. Let X have probability density function $f(x)$ and let the pair (X,Y) have joint probability density function $f(x,y)$.

The **expected value** of X is defined as $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

(If $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$; otherwise $E(X)$ is undefined.) $E(X)$ is also called the **expectation** of X or the **mean** of X and is often denoted μ or μ_X.

Properties of Expectation

If $h(x)$ is a function such that $\int_{-\infty}^{\infty} |h(x)|f(x)dx < \infty$, it may be shown that

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx$$

Similarly, if $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |h(x,y)|f(x,y)dxdy < \infty$, it may be shown that

$$E[h(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y)f(x, y)dxdy \tag{2.A.1}$$

As a corollary to Equation (2.A.1), we easily obtain the important result

$$E(aX + bY + c) = aE(X) + bE(Y) + c \tag{2.A.2}$$

We also have

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x, y)dxdy \tag{2.A.3}$$

The **variance** of a random variable X is defined as

$$Var(X) = E\{[X - E(X)]^2\} \tag{2.A.4}$$

(provided $E(X^2)$ exists). The variance of X is often denoted by σ^2 or σ_X^2.

Properties of Variance

$$Var(X) \geq 0 \tag{2.A.5}$$

$$Var(a + bX) = b^2 Var(X) \tag{2.A.6}$$

If X and Y are independent, then

$$Var(X + Y) = Var(X) + Var(Y) \tag{2.A.7}$$

In general, it may be shown that

$$Var(X) = E(X^2) - [E(X)]^2 \tag{2.A.8}$$

The positive square root of the variance of X is called the **standard deviation** of X and is often denoted by σ or σ_X. The random variable $(X - \mu_X)/\sigma_X$ is called the **standardized version** of X. The mean and standard deviation of a standardized variable are always zero and one, respectively.

The **covariance** of X and Y is defined as $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.

Properties of Covariance

$$Cov(a + bX, c + dY) = bdCov(X, Y) \tag{2.A.9}$$

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) \tag{2.A.10}$$

$$Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z) \tag{2.A.11}$$

$$Cov(X, X) = Var(X) \tag{2.A.12}$$

$$Cov(X, Y) = Cov(Y, X) \tag{2.A.13}$$

If X and Y are independent,

$$Cov(X, Y) = 0 \tag{2.A.14}$$

The **correlation coefficient** of X and Y, denoted by $Corr(X, Y)$ or ρ, is defined as

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Alternatively, if X^* is a standardized X and Y^* is a standardized Y, then $\rho = E(X^*Y^*)$.

Properties of Correlation

$$-1 \leq Corr(X, Y) \leq 1 \tag{2.A.15}$$

$$Corr(a + bX, c + dY) = sign(bd)Corr(X, Y)$$

$$\text{where } sign(bd) = \begin{cases} 1 \text{ if } bd > 0 \\ 0 \text{ if } bd = 0 \\ -1 \text{ if } bd < 0 \end{cases} \tag{2.A.16}$$

$Corr(X, Y) = \pm 1$ if and only if there are constants a and b such that $Pr(Y = a + bX) = 1$.