

# DEPARTMENT OF MATHEMATICS AND STATISTICS

Memorial University of Newfoundland  
CANADA A1C 5S7

St. John's, Newfoundland  
ph. (709) 737-8075 fax (709) 737-3010

---

Alwell Julius Oyet, PhD

email: aoyet@math.mun.ca

---

## TIME SERIES ANALYSIS - STAT 3540 LECTURE NOTES

### 1. INTRODUCTION

The purpose of this course is to introduce students to basic methods for analyzing a time series and for forecasting. By definition, a time series is any set or sequence of observations  $y_t$ , each one being recorded at a specific time  $t$ . The time parameter  $t$  could be hours, days, weeks, quarters, months, years, etc. For example, the time parameter in the time series, daily temperature in St. John's from 1999 to 2003, is days. Thus, the values of  $t$  are  $t = 1, 2, \dots, n$ , where  $n = 1826 =$  total number of days between January 1, 1999 and December 31, 2003, inclusive, and the observed values of the temperature will be denoted by  $y_1, y_2, \dots, y_n$ . In this example, the city of St. John's may be interested in forecasting the daily temperature for 2004, for the purpose of planning, based on the observed 5-year temperature data. Then, 2004 becomes the future time with time parameter denoted by  $t + l$ , where  $l$ , called the lead time (days), takes the values  $l = 1, 2, \dots, 366$  (2004 is a leap year). Let us denote the forecasted values by  $\hat{y}_t(l)$ . Then, the forecast for January 1, 2004, represented by  $\hat{y}_n(l) = \hat{y}_{1826}(1)$ , will have lead time  $l = 1$  and time parameter  $t + 1 = n + 1 = 1827$ . Since the city may base its plans on the forecast, it is expected that the forecasts  $\hat{y}_t(l)$  should be as close as possible to the actual values  $y_{t+l}$ . It is common to use the magnitude of the mean squared error (MSE) of the deviations  $\hat{y}_t(l) - y_{t+l}$  as a measure of accuracy of forecasts. We will aim at obtaining forecasts with MSE of deviations that is as small as possible for each lead time  $l$ .

Forecasting future events have become of great interest to governments, businesses and industries because of the significant role they play in decision-making processes. Such forecasts provide a basis for (a) business and economic planning (b) inventory and production planning, and (c) control and optimization of industrial processes. Some examples of predictions that are of interest are Fall 2004 - Fall 2007 MUN student enrollment for budgetary planning, the unemployment rate, inflation rate, volume of monthly sales, volume of weekly exports and imports, 2004 revenue for the province of Newfoundland and Labrador, and so on. A time series, such as electric signals or voltage, which can be recorded continuously in time, is said to be continuous. Other series, such as daily temperature,

hourly wind speed, daily stock prices, yearly earnings of a company, interest rates, yields and volume of sales which are taken only at specific time intervals is said to be discrete. Our focus will be on discrete time series. We will study methods which use data from past time points to develop models we can use for predicting future values of the time series.

### 1.1. Fundamental Concepts

The first step in the analysis of a time series is to plot the series and examine the plot in order to identify any obvious patterns that may be useful in its analysis. For example, consider the four time series plotted in Figure 1.

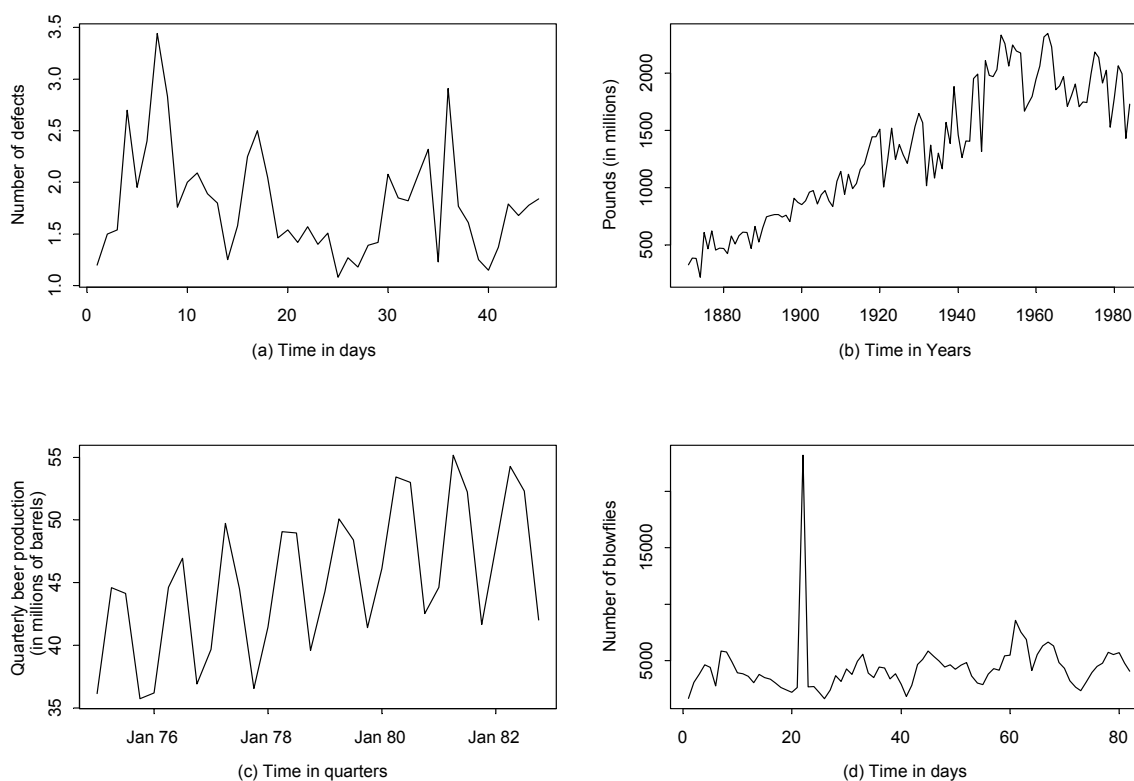


Figure 1: Plot of (a) Daily average number of truck manufacturing defects, (b) Yearly U.S. tobacco production, (c) Quarterly U.S. beer production (d) Contaminated blowfly data.

**Figure 1(a):** The daily average number of defects per truck  $y_t$ ,  $t = 1, \dots, 45$  found at the end of the assembly line of a truck manufacturing plant, shown in Figure 1(a) appears to vary about a fixed level. Time series that exhibit this phenomenon are said to be stationary in the mean and are special

cases of stationary time series.

1 :	1.20	1.50	1.54	2.70	1.95	2.40	3.44	2.83	1.76	2.00	2.09	1.89	1.80	1.25	1.58
16 :	2.25	2.50	2.05	1.46	1.54	1.42	1.57	1.40	1.51	1.08	1.27	1.18	1.39	1.42	2.08
31 :	1.85	1.82	2.07	2.32	1.23	2.91	1.77	1.61	1.25	1.15	1.37	1.79	1.68	1.78	1.84

We note that these  $y_t$ 's are the observed values of a more general random variable  $Y_t =$  the daily average number of defects per truck at the manufacturing plant. Any family of random variables  $\{Y_t, t \in I\}$  indexed by time, where  $I$  is an indexing set, and defined on the sample space of an experiment is called a stochastic process. In our discussion, we will assume that  $I$  is the set of all integers denoted by  $Z$ . It follows from the above discussion that a time series is a sample or a set of realizations from a stochastic process or sequentially observed values of a stochastic process  $Y_t$ . Since a time series is not a random sample but observed sequentially, the observed values are not independent. Observations that are, say  $k$  ( $k = 1, 2, 3, \dots$ ), distances apart will be correlated. The larger the value of  $k$ , or the further apart the observations are from each other, the weaker the correlation. For instance, the correlation between observation 1 and observation 3 will be stronger than the correlation between observations 1 and 7. This concept of correlation between observations  $k$  distances apart (which we shall call autocorrelation) is important because if we can determine the structure of the relationship between observations that are  $k$  distances apart, one can use that relationship to predict what will happen  $k$  distances from the last observation under the assumption that the relationship will continue into the future.

Now, consider a set of  $n$  random variables  $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$  from the stochastic process  $\{Y_t, t \in Z\}$ . Let  $\{y_{t_1}, y_{t_2}, \dots, y_{t_n}\}$  be observed values of  $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$  and let  $F(y_{t_1}, \dots, y_{t_n})$  be the  $n$ -dimensional joint probability distribution function of  $\{Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}\}$ . A stochastic process  $\{Y_t, t \in Z\}$  is said to be  $n$ th order stationary in distribution if

$$F(y_{t_1}, \dots, y_{t_n}) = F(y_{t_1+k}, \dots, y_{t_n+k}) \tag{1.1}$$

for any integers  $(t_1, \dots, t_n)$  and  $k$ . If (1.1) is true for any  $n = 1, 2, \dots$ , the process is said to be strictly or strongly or completely stationary. Note that if (1.1) is true for a given value of  $n$ , it is also true for any  $m < n$ .

Recall that the mean function of  $Y_t$  is

$$\mu_t = E(Y_t),$$

the variance function is

$$\sigma_t^2 = E(Y_t - \mu_t)^2,$$

the covariance function between  $Y_{t_1}$  and  $Y_{t_2}$  is

$$c(t_1, t_2) = E(Y_{t_1} - \mu_{t_1})(Y_{t_2} - \mu_{t_2}),$$

and the correlation function between  $Y_{t_1}$  and  $Y_{t_2}$  is

$$r(t_1, t_2) = \frac{c(t_1, t_2)}{\sqrt{\sigma_{t_1}^2} \sqrt{\sigma_{t_2}^2}}.$$

If  $Y_t$  is a strictly stationary process, it means that the distribution function of  $Y_t$  is the same for all  $t$ . Thus,  $\mu_t$  will be the same for all  $t$ , hence  $\mu_t$  is a constant. That is,  $\mu_t = \mu$  (is not a function of  $t$ ) provided  $E(|Y_t|) < \infty$ , for a strictly stationary process. In the same way, if  $E(Y_t^2) < \infty$ , then  $\sigma_t^2 = \sigma$  for all  $t$  and hence is also not a function of time  $t$ . Similarly, it can be shown that, if  $Y_t$  is strictly stationary, then for any integers  $t_1, t_2$  and  $k$ , we have

$$c(t_1, t_2) = c(t_1 + k, t_2 + k) = c_k \quad \text{and} \quad r(t_1, t_2) = r(t_1 + k, t_2 + k) = r_k.$$

Thus, the covariance and correlation between  $Y_t$  and  $Y_{t+k}$  depends only on the time difference  $k$ . The problem with the definition of a strictly stationary process is that it is very difficult or impossible to obtain the joint distribution function of a process  $Y_t$  from the observed series  $y_t$ . Hence, it is very difficult to verify strict stationarity. Consequently, we often use a weaker definition of stationarity which depends only on the moments of the process, such as mean, variance, covariance and correlation. These, we can easily compute from an observed series.

A process is said to be  $n$ th order weakly stationary if all its joint moments up to order  $n$  exist and does not depend on time  $t$ . By this definition a second order weakly stationary process (sometimes called covariance stationary or stationary in the wide sense) have (i) constant mean, (ii) constant variance, (iii) covariance and correlation functions which depend on time difference only.

**Figure 1(b):** The yearly U.S. tobacco production from 1871 to 1984  $y_t$  shown in Figure 1(b) does not vary about a fixed level. Rather, the general direction of the series  $y_t$  is that it appear to be increasing as  $t$  increases and hence exhibits an overall upward trend. In addition, the fluctuations of this tobacco series increases as the level of the series increases, indicating that the variance  $\sigma_t^2$  of the series depends on time. Any time series that exhibits these patterns are said to be nonstationary in mean and variance and are examples of nonstationary time series.

1871 :	327	385	382	217	609	466	621	455	472	469	426	579	509
1884 :	580	611	609	469	661	525	648	747	757	767	767	745	760
1897 :	703	909	870	852	886	960	976	857	939	973	886	836	1054
1910 :	1142	941	1117	992	1037	1157	1207	1326	1445	1444	1509	1005	1254
1923 :	1518	1245	1376	1289	1211	1373	1533	1648	1565	1018	1372	1085	1302
1936 :	1163	1569	1386	1881	1460	1262	1408	1406	1951	1991	1315	2107	1980
1949 :	1969	2030	2332	2256	2059	2244	2193	2176	1668	1736	1796	1944	2061
1962 :	2315	2344	2228	1855	1887	1968	1710	1804	1906	1705	1749	1742	1990
1975 :	2182	2137	1914	2025	1527	1786	2064	1994	1429	1728			

**Figure 1(c):** In this figure we observe a regular pattern that is repetitive in nature. Within each year, the level of production increases from the first quarter to a peak then declines gradually to the fourth quarter. Such repetitive patterns that are regular are due to seasonal variations. That is, the different seasons within each year affect the level of beer production. Now, since these seasons are the same each year we expect the pattern to repeat itself thereby leading to what is normally called a seasonal time series.

	1Q	2Q	3Q	4Q
1975 :	36.14	44.60	44.15	35.72
1976 :	36.19	44.63	46.95	36.90
1977 :	39.66	49.72	44.49	36.54
1978 :	41.44	49.07	48.98	39.59
1979 :	44.29	50.09	48.42	41.39
1980 :	46.11	53.44	53.00	42.52
1981 :	44.61	55.18	52.24	41.66
1982 :	47.84	54.27	52.31	42.03

Seasonal time series are also nonstationary time series. We also note that the level of the series seem to be increasing as we move from 1975 to 1982. Thus, the beer series is also nonstationary in the mean. The importance of stationarity in modelling a time series is that the nondependence of the variation of the series and the level of the series on time, makes the series more stable, easier to model and lead to a more reliable forecast. Thus, in analyzing a time series we will examine the series for signs of nonstationarity, attempt to identify the components of the series that are responsible for the nonstationarity of the series, extract these components or transform the series into a stationary series, and then attempt to model the stationary component.

### Components of a Time Series

The examples in Figures 1, have been used to illustrate certain patterns that are commonly found in time series. There are four main components of a time series. These are: Trend, Cycle, Seasonal Variations and Irregular Fluctuations. These components usually do not occur alone in a time series.

They can occur in any combination or all can occur together.

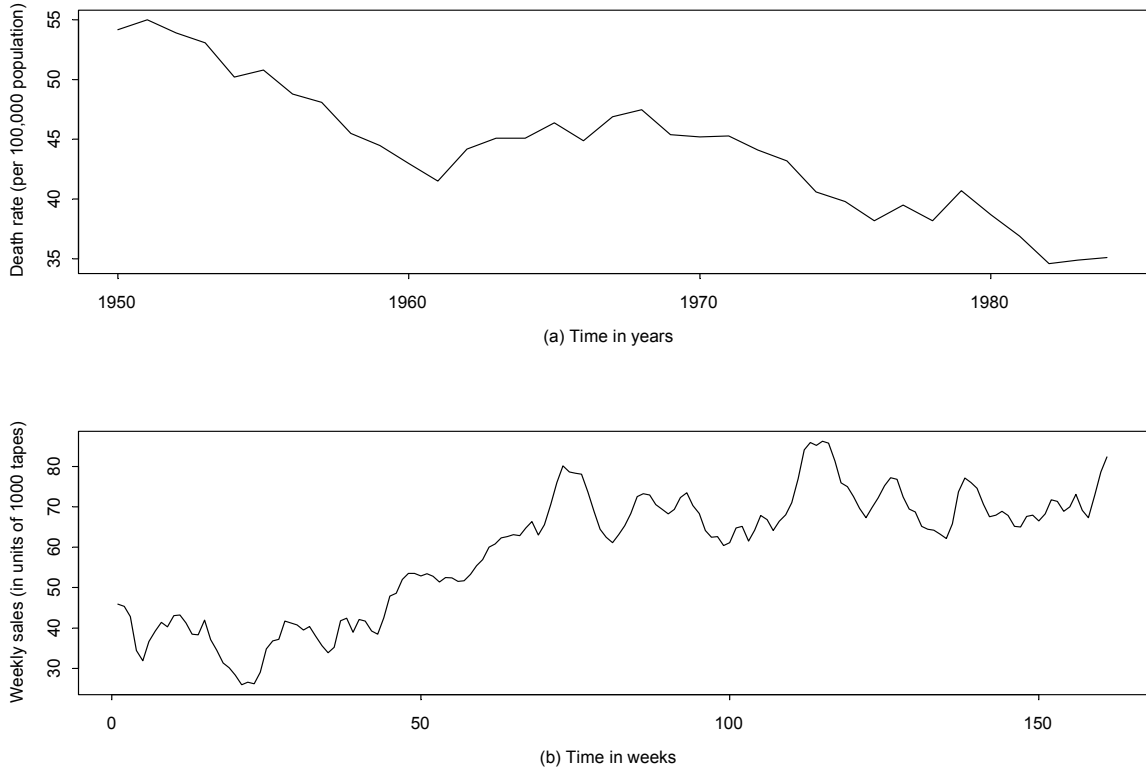


Figure 2: Plot of (a) Yearly Pennsylvania accidental death rate between 1950 and 1984, (b) Weekly sales of Super Tech videocassette Tape.

**Trend:** By trend we mean the general upward or downward or curvilinear direction in which a time series appear to be moving over a period of time. Thus, trend reflects the long-term growth or decline in the series. Figures 1(b) and 1(c) show two time series with increasing linear trend. Other examples of time series with trend components are shown in Figure 2. Figure 2(a) is a plot of the yearly accidental death rate for Pennsylvania between 1950 and 1984 showing a decreasing trend component.

1950 :	54.2	55.0	53.9	53.1	50.2	50.8	48.8	48.1	45.5	44.5	43.0	41.5
1962 :	44.2	45.1	45.1	46.4	44.9	46.9	47.5	45.4	45.2	45.3	44.1	43.2
1974 :	40.6	39.8	38.2	39.5	38.2	40.7	38.7	36.9	34.6	34.9	35.1	

Figure 2(b) which is the weekly sales of Super Tech Videocassette Tape (in units of 1000 tapes) shows an initial increasing trend which appear to start changing directions or level off from the 74th week. These examples show that the trend component can be either linear or nonlinear. Some factors

that can be represented by the trend component of a time series are: market growth, increase in population, inflation or deflation (price changes), changes in consumer taste, etc. The weekly sales in Figure 2(b) are:

1 :	45.9	45.4	42.8	34.4	31.9	36.6	39.2	41.4	40.3	43.1	43.2	41.2	38.4
14 :	38.3	41.9	37.1	34.5	31.3	30.2	28.3	25.9	26.6	26.2	29.0	34.8	36.8
27 :	37.2	41.7	41.2	40.7	39.5	40.4	38.0	35.6	33.9	35.2	41.8	42.4	38.9
40 :	42.1	41.7	39.2	38.5	42.5	47.9	48.6	52.0	53.5	53.5	52.9	53.4	52.8
53 :	51.4	52.5	52.4	51.5	51.7	53.3	55.4	56.9	60.0	60.8	62.3	62.6	63.1
66 :	62.8	64.7	66.3	63.0	65.5	70.6	76.0	80.1	78.6	78.3	78.1	73.6	68.8
79 :	64.4	62.4	61.1	63.1	65.3	68.3	72.5	73.2	72.9	70.5	69.4	68.2	69.3
92 :	72.3	73.5	70.3	68.3	64.1	62.5	62.6	60.4	61.1	64.7	65.1	61.5	64.2
105 :	67.8	66.8	64.1	66.4	68.0	71.0	76.9	84.1	85.9	85.2	86.2	85.7	81.3
118 :	75.9	75.0	72.5	69.6	67.3	69.8	72.2	75.2	77.2	76.8	72.4	69.4	68.7
131 :	65.1	64.4	64.2	63.2	62.1	65.8	73.7	77.1	76.0	74.6	70.6	67.5	67.9
144 :	68.9	67.8	65.1	65.0	67.6	67.9	66.5	68.2	71.7	71.3	68.9	70.0	73.1
157 :	69.1	67.3	72.9	78.6	82.3								

**Seasonal Variation:** Seasonal variations refer to periodic patterns in a time series that are repeated from year to year. An example of a series with seasonal variation is shown in Figure 1(c). The periodic pattern in Figure 1(c) is repeated after every 4th quarter. Thus, we say that the period of the series is 4. Some factors that cause seasonal variation are: regular yearly weather patterns, regular yearly religious celebrations or customs or holidays. For instance the average monthly temperature in St. John's is clearly seasonal since it directly measures the various weather patterns in St. John's. Other examples of time series affected by seasonal variations are shown in Figure 3. The seasonal effects on the series in Figure 3 are apparent. The employment figures shown in Figure 3(a) increase dramatically in the summer months, with peaks occurring in June when schools are not in session and decrease in the fall months when school reopen.

	<i>Jan</i>	<i>Feb</i>	<i>Mar</i>	<i>Apr</i>	<i>May</i>	<i>Jun</i>	<i>Jul</i>	<i>Aug</i>	<i>Sep</i>	<i>Oct</i>	<i>Nov</i>	<i>Dec</i>
1971 :	707	655	638	574	552	980	926	680	597	637	660	704
1972 :	758	835	747	617	554	929	815	702	640	588	669	675
1973 :	610	651	605	592	527	898	839	614	594	576	672	651
1974 :	714	715	672	588	567	1057	949	683	771	708	824	835
1975 :	980	969	931	892	828	1350	1218	977	863	838	866	877
1976 :	1007	951	906	911	812	1172	1101	900	841	853	922	886
1977 :	896	936	902	765	735	1234	1052	868	798	751	820	725
1978 :	821	895	851	734	636	994	990	750	727	754	792	817
1979 :	856	886	833	733	675	1004	956	777	761	709	777	771
1980 :	840	847	774	720	848	1240	1168	936	853	910	953	874
1981 :	1026	1030	946	860	856	1190	1038	883	843	857	1016	1003

The number of rooms occupied at the Traveller's Rest Inc. in Figure 3(b) peaks in July or August and decreases in the fall. The plot also shows that the level of the seasonal fluctuations about the

trend is increasing as time increases.

	<i>Jan</i>	<i>Feb</i>	<i>Mar</i>	<i>Apr</i>	<i>May</i>	<i>Jun</i>	<i>Jul</i>	<i>Aug</i>	<i>Sep</i>	<i>Oct</i>	<i>Nov</i>	<i>Dec</i>
1977 :	501	488	504	578	545	632	728	725	585	542	480	530
1978 :	518	489	528	599	572	659	739	758	602	587	497	558
1979 :	555	523	532	623	598	683	774	780	609	604	531	592
1980 :	578	543	565	648	615	697	785	830	645	643	551	606
1981 :	585	553	576	665	656	720	826	838	652	661	584	644
1982 :	623	553	599	657	680	759	878	881	705	684	577	656
1983 :	645	593	617	686	679	773	906	934	713	710	600	676
1984 :	645	602	601	709	706	817	930	983	745	735	620	698
1985 :	665	626	649	740	729	824	937	994	781	759	643	728
1986 :	691	649	656	735	748	837	995	1040	809	793	692	763
1987 :	723	655	658	761	768	885	1067	1038	812	790	692	782
1988 :	758	709	715	788	794	893	1046	1075	812	822	714	802
1989 :	748	731	748	827	788	937	1076	1125	840	864	717	813
1990 :	811	732	745	844	833	935	1110	1124	868	860	762	877

The pattern in these series repeat itself every 12 months, and thus the seasonal period is 12.

**Cycle:** By cycle we mean recurring periodic patterns around trend lines with duration longer than a year. These fluctuations can have a period of anywhere from two to ten years measured from peak to peak or trough to trough. One very common cyclical fluctuation is the “business cycle.” Business cycles are caused by recurrent periods of economic or business expansion (boom) followed by a period of economic or business recession or contraction.

**Irregular Fluctuations:** These are fluctuations in a time series that follow no regular or recognizable pattern. Many irregular fluctuations in a time series are caused by “unusual” events that are difficult to predict such as earthquakes, accidents, hurricanes, wars, strikes, etc.

### **A General Approach to Time Series Modelling**

The general approach to analyzing a time series is as follows:

1. Plot the series and examine the main features of the graph, checking in particular if
  - (a) the series contains a trend component,
  - (b) fluctuations about the trend curve are increasing with time,
  - (c) the series contains a seasonal component,
  - (d) there are any apparent sharp changes in behaviour

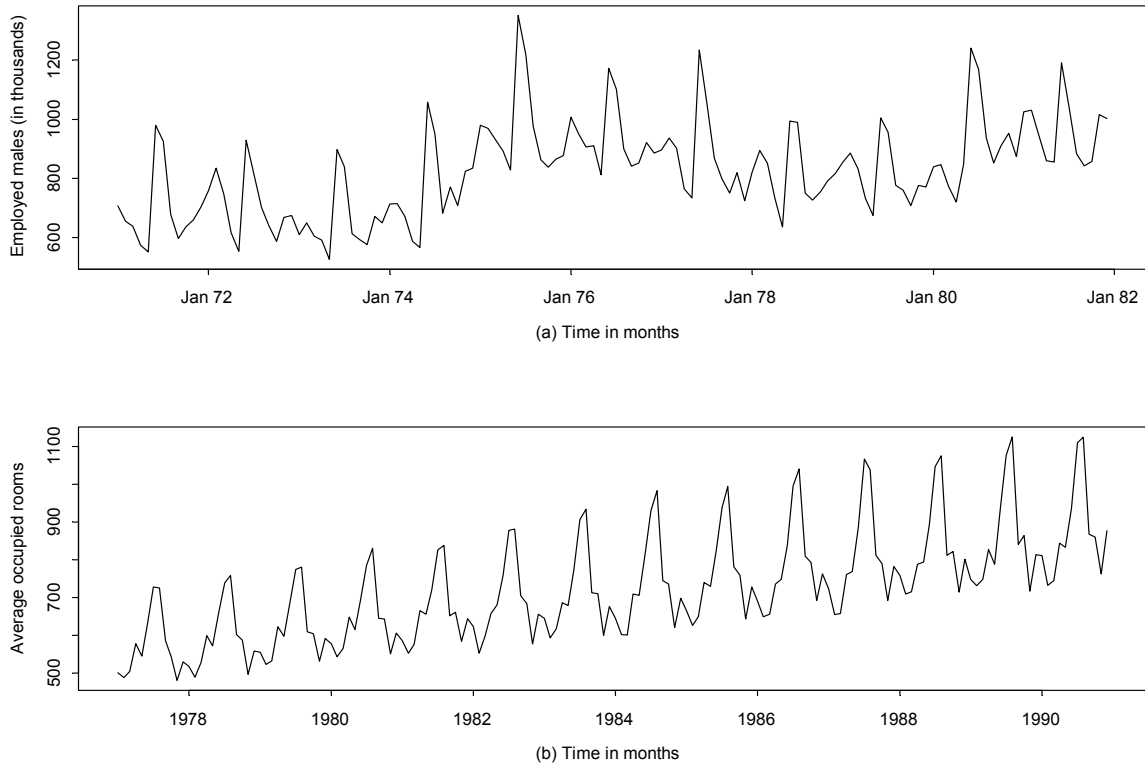


Figure 3: Plot of (a) Monthly employment figures for males between ages 16 and 19 from January 1971 to December 1981 (in thousands), (b) Monthly occupied hotel room averages for traveller's rest Inc. for 1977 - 1990.

(e) there are any outlying observations.

2. If the series is nonstationary by means of any of the components mentioned in (1), reduce the time series to a stationary series by extracting the components and/or by transformation.
3. Identify a suitable model for the stationary series using various sample statistics we shall discuss later.
4. Fit the identified model to the series and perform model diagnostics to check goodness of fit.
5. Combine the extracted components with the fitted model to obtain forecasts of the original series.