Chapter 7

Time Series Analysis

7.1 Introduction

Definition:

The discrete time series can be defined as a set of time-ordered observations of some process over a given period of time. If the observations are made at equally spaced time intervals $t_1, t_2, t_3, \ldots$, the time series can then be represented as \{X_{t_1}, X_{t_2}, X_{t_3}, \ldots\}. Also, the length of data set is called as the length of Time Series. The observation values may be real or complex numbers. Time series facilitate us to understand the underlying mechanism that generates the observed data and, in turn, to forecast future values of the series. Without loss of generality we can think of the generating mechanism as probabilistic and accordingly model time series as stochastic processes. Thus we can consider each observation to be a value of some random variable; the time series, a single realization of a stochastic process (i.e., a sequence of random variables).

Examples of Time Series Around Us

Daily closing stock prices, monthly unemployment figures, the annual precipitation
CHAPTER 7. TIME SERIES ANALYSIS

index, crime rates, and earthquake aftershock frequencies are all examples of time series which we encounter. Virtually any quantity recorded over time yields a time series. To "visualize" a time series we plot our observations as a function of the time. This is called a time plot.

7.2 Processes

7.2.1 Auto-Regressive Process

In these type of processes current element is serially dependent on the previous (time lagged) ones. It can be modeled as follows:

AR(p) Model.

\[ Z_t = \phi_1 * Z_{t-1} + \phi_2 * Z_{t-2} + \cdots + \phi_p * Z_{t-p} + \epsilon \] (7.1)

where \( \epsilon \) is noise, and \( \phi_1, \phi_2, \cdots, \phi_p \) are the autoregressive model parameters. Thus each observation is made up of a random error component (random shock) and a linear combination of prior observations.

Stationarity Requirement

For the series to be stable the parameters should be within a certain range; for example, if there is only one autoregressive parameter then it must fall within the interval of \(-1 < 1\). Otherwise, past effects would accumulate and the values of successive \( Z_t \)'s would move towards infinity, that is, the series would not be stationary. If there is more than one autoregressive parameter, similar (general) restrictions on the parameter values can be defined, as described in [62].
7.2.2 Moving Average Process

Another approach for modeling univariate time series models is the moving average (MA) model as given below:

**MA(q) Model.**

\[
Z_t = a_t - \theta_1 * a_{t-1} - \cdots - \theta_q * a_{t-q}
\]  

(7.2)

where \(Z_t\) is the time series, \(a_t\)'s are white noise, and \(\theta_1, \cdots, \theta_q\) are the parameters of the model. The value of \(q\) is called the order of the MA model. In these processes each element in the series is affected by the past error (or random shock) that cannot be accounted for by the autoregressive component. Thus each observation is made up of a random error component (random shock) and a linear combination of prior random shocks.

**Invertibility Requirement**

We observe that there is a "duality" between the moving average process and the autoregressive process, which is explained in [62], [63]. It is shown that the moving average equation 7.2 can be rewritten (inverted) into an autoregressive form (of infinite order). However, analogous to the stationarity condition described above, this can only be done if the moving average parameters follow certain conditions, that is, if the model is invertible. Otherwise, the series will not be stationary.

7.2.3 Auto-Regressive Moving Average Models (ARMA)

Autoregressive moving average model is the general model introduced by Box and Jenkins (1976) which includes autoregressive as well as moving average parameters (ref. [62], [64] and [65]). This is described using equation 7.3:

**ARMA(p,q) model:**

\[
Z_t - \phi_1 * Z_{t-1} - \phi_2 * Z_{t-2} - \cdots - \phi_p * Z_{t-p} = a_t - \theta_1 * a_{t-1} - \cdots - \theta_q * a_{t-q}
\]  

(7.3)
It can also include differencing parameter in the formulation of the model. Thus there are three types of parameters in the model i.e. the autoregressive parameters \( p \), the number of differencing passes \( d \), and moving average parameters \( q \). The notation used is ARIMA \((p, d, q)\). For example, a model described as \((1, 2, 2)\) means that it contains 1 autoregressive \((p)\) parameter and 2 moving average \((q)\) parameters which were computed for the series after it was differenced twice.

### 7.3 Points to Remember

- The underline assumption in the Box-Jenkins model is the stationarity of the time series.

- For the non-stationary time series Box and Jenkins suggested for differencing the given series for one or more times.

- In this process we arrive at an ARIMA model after differencing, in which "i" refers to integrated.

- New time series corresponding to zero mean can be obtained by subtracting the mean of series from each data point.

- Box-Jenkins models can be extended to include seasonal autoregressive and seasonal moving average terms.

### 7.4 Box-Jenkins Modeling Approach

Important steps in time series model building using Box-Jenkins approach are as follows:

1. Model Identification.
CHAPTER 7. TIME SERIES ANALYSIS


3. Model Validation.

These steps are depicted in the flow chart as given in Figure 7.1 below:

Figure 7.1: Flow chart showing the modeling steps

7.4.1 Box-Jenkins Model Identification

Stationarity and Seasonality- As explained earlier the underline assumption in Box-Jenkins approach is that the given time series of data is stationary. The stationarity can be
detected with the help of indexRun sequence plot Run Sequence Plot. For station-
arity it is important that Run Sequence Plot should show the constant location and
scale. Another method to detect the stationarity is with the help of Autocorrelation
plot. In general it is observed that the Autocorrelation plot decays slowly for the
non-stationary series [24].

**Differencing** - To achieve the stationarity in the given series, Box and Jenkins suggested
the differencing approach.

**Seasonality or Periodicity** - As a first step it is equally important to check if there is any
significant seasonality that needs to be modeled. Seasonality can be detected with the
help of an either autocorrelation plot, or a seasonal sub-series plot / a spectral plot.

**Seasonal Differencing** - If the seasonality exists, it is important to identify the order for
the seasonal autoregressive and seasonal moving average terms. For most of the
experiments repetition period is decided, a priory. For example, for monthly data we
would typically include either a seasonal AR 12 term or a seasonal MA 12 term. In
certain situations seasonal differencing may be applied to remove seasonality effect.
In general Box-Jenkins models, we do not explicitly remove seasonality before fitting
the model.

**Model Order** (p and q) - Having got stationary time series, the next step is to identify the
order (i.e., the p and q) of the autoregressive and moving average terms.

**Autocorrelation and Partial Autocorrelation Plots** - The appropriate number of autore-
gressive and moving average parameters needed can be determined from Autocorre-
lation Function (ACF) and the Partial Autocorrelation Function (PACF). The method-
ology is outlined below:

1. **One autoregressive** (p) parameter:
ACF - exponential decay; PACF - spike at lag 1, no correlation for other lags.

2. Two autoregressive (p) parameters:
   ACF - sinc wave shaped pattern or a set of exponential decays; PACF - spikes at lag 1 and lag 2, no correlation for other lags.

3. One moving average (q) parameter:
   ACF - spike at lag 1, no correlation for other lags; PACF - damps out exponentially

4. Two moving average (q) parameters:
   ACF - spikes at lag 1 and lag 2, no correlation for other lags; PACF - a sinc wave shaped pattern or a set of exponential decays.

5. One autoregressive (p) and one moving average (q) parameters:
   ACF - exponential decay starting at lag 1; PACF - exponential decay starting at lag 1.

7.5 Time Series Model for Inter-arrival Times

7.5.1 Step I - Data

For the traffic traces used in this study, inter-arrival times are computed for individual users. Timestamp information within the segregated users is used for this purpose. Sample data is as given below:

<table>
<thead>
<tr>
<th>Time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000000</td>
</tr>
<tr>
<td>36.488482</td>
</tr>
<tr>
<td>0.988127</td>
</tr>
<tr>
<td>1.549466</td>
</tr>
<tr>
<td>45.323660</td>
</tr>
<tr>
<td>Value 1</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>1.231711</td>
</tr>
</tbody>
</table>
7.5.2 Step II- Check for: Stationarity, Seasonality, Outliers "Run Sequence Plot"

In this step we check the stationarity and seasonality of the above data with the help of run sequence plot. This plot is shown in Figure 7.2. From this plot, we observe that,

- The data show strong and positive autocorrelation.
Figure 7.2: Run sequence plot of the raw data

- There does not seem to be a significant trend or any obvious seasonal pattern in the data.

- Next we examine the autocorrelation plot:

  - The autocorrelation plot in Figure 7.3 shows that the sample autocorrelations are very strong, positive and decay very slowly.

  - The autocorrelation plot indicates that the process is non-stationary and suggests an ARIMA model.
- Next step is to difference the data.

- Run sequence plot of given time series after differencing is shown in Figure 7.4.

- Again examine the sample autocorrelations and Partial Autocorrelation plots of the time series data after differencing it. These plots are shown in Figures 7.5 and 7.6 respectively. As discussed previously these plots suggest two auto-regressive parameters.

7.5.3 Model Estimation

In this section, we compute the model parameters using nonlinear least square estimation of an ARIMA model using back forecasts. The initial conditions are, parameter $p = 2$, \ldots
CHAPTER 7. TIME SERIES ANALYSIS

Figure 7.4: Run sequence plot of the differenced (at lag(1)) time series.

Figure 7.5: Autocorrelation plot of the differenced time series.
differencing $d = 1$ and parameter $q = 0$. Data-plot generated the following estimation output for the AR(2) model:

\[
\begin{align*}
\text{AR } 1 &= -0.74100 \\
\text{AR } 2 &= -0.31519
\end{align*}
\]

Thus the model for the differenced data, $Y_t$, is an AR(2) model:

\[
Y(t) = -0.741 \cdot Y(t - 1) - 0.316 \cdot Y(t - 2) \quad (7.4)
\]

It is often more convenient to express the model in terms of the original data, $X_t$, rather than the differenced data. From the definition of the difference, $Y_t = X_t - X_{t-1}$, we can make the appropriate substitutions into the above equation:

\[
X(t) - X(t - 1) = -0.741 \cdot (X(t - 1) - X(t - 2)) - 0.316 \cdot (X(t - 2) - X(t - 3)) \quad (7.5)
\]

to arrive at the model in terms of the original series:

\[
X(t) = 0.259 \cdot X(t - 1) + 0.425 \cdot X(t - 2) + 0.316 \cdot X(t - 3) \quad (7.6)
\]
CHAPTER 7. TIME SERIES ANALYSIS

7.5.4 Model Validation

Residuals:

A good model should not only provide sufficiently accurate forecasts, it should also be parsimonious and should produce statistically independent residuals that contain only noise and no systematic components (e.g., the correlogram of residuals should not reveal any serial dependencies).

A good test of the model is:

- To plot the residuals and inspect them for any systematic trends.
- To examine the auto-correlogram of residuals (there should be no serial dependency between residuals).

As with standard non-linear least square fitting, the primary tool for model diagnostic checking is residual analysis.

Analysis of residuals:

The major concern here is that the residuals are systematically distributed across the series (e.g., they could be negative in the first part of the series and approach zero in the second part) or that they contain some serial dependency which may suggest that the ARIMA model is inadequate. The analysis of ARIMA residuals constitutes an important test of the model. The estimation procedure assumes that the residual are not (auto-) correlated and that they are normally distributed. The Time Series module will automatically compute the residuals, and make them available for further analyses with all other methods available in this module; residuals can also be saved to a data file, together with the original series or its transformations.

4-Plot of Residuals from ARIMA(2,1,0) Model:

The 4-plot is a convenient graphical technique for model validation.
CHAPTER 7. *TIME SERIES ANALYSIS*

**Normal Probability Plot**

This type of graph is used to evaluate the normality of the distribution of a variable, that is, whether and to what extent the distribution of the variable follows the normal distribution. The selected variable will be plotted in a scatter-plot against the values "expected from the normal distribution." The standard normal probability plot is constructed as follows.

First, the deviations from the mean (residuals) are rank ordered. From these ranks the program computes $z$ values (i.e., standardized values of the normal distribution) based on the assumption that the data come from a normal distribution. These $z$ values are plotted on the Y-axis in the plot. If the observed residuals (plotted on the X-axis) are normally distributed, then all values should fall onto a straight line. If the residuals are not normally distributed, then they will deviate from the line. Outliers may also become evident in this plot. If there is a general lack of fit, and the data seem to form a clear pattern (e.g., an S shape) around the line, then the variable may have to be transformed in some way (e.g., a log transformation to "pull-in" the tail of the distribution, etc.).

The normal probability plot for Residuals from ARIMA(2,1,0) Model is shown in Figure 7.7.

**Half-Normal Probability Plot**

The half-normal probability plot is constructed in the same way as the standard normal probability plot, except that only the positive half of the normal curve is considered. Consequently, only positive normal values will be plotted on the Y-axis.

The half normal probability plot for Residuals from ARIMA(2,1,0) Model is shown in Figure 7.8.
Figure 7.7: Normal probability plot of residuals.

Figure 7.8: Half normal probability plot of residuals.
- **De-trended Probability Plot**

This plot is constructed in the same way as the standard normal probability plot, except that before the plot is generated, the linear trend is removed. This often "spreads out" the plot, thereby allowing the user to detect patterns of deviations more easily. The de-trended normal probability plot for Residuals from ARIMA(2,1,0) Model is shown in Figure 7.9.

![De-trended Normal Probability Plot](image)

Figure 7.9: De-trended probability plot of residuals.

- **Autocorrelation Plot of Residuals from ARIMA(2,1,0) Model**

In addition, the autocorrelation plot of the residuals from the ARIMA(2,1,0) model was generated. The autocorrelation plot (Figure 7.10) shows that for the first 15 lags, all sample autocorrelations except those at lags 2 fall inside the 95 percent confidence bounds indicating the residuals appear to be random.

We can make the following conclusions based on these plots:

1. The run sequence plot shows that the residuals do not violate the assump-
Figure 7.10: Autocorrelations plot of residuals.

Figure 7.11: Histogram of residuals.
CHAPTER 7. TIME SERIES ANALYSIS

It shows that most of the residuals are in the range (-1, 1).

2. The lag plot indicates that the residuals are not autocorrelated at lag 1.

3. The histogram and normal probability plot indicate that the normal distribution provides an adequate fit for this model.

7.6 Summary

For IITB traces we found that there were a total of 832 sources. Out of these only 41% i.e. around 346 sources contribute to 90% utilisation. No machine sources were encountered. Users are grouped into three categories as follows:

High : 50 - 100
Moderate : 101 - 200
Low: 250 - 325

Important observation about the time series of inter-arrival transaction times is that about 60% users from moderate and low group fit well into AR(2) model. We have to come up with some distribution for AR(2) model parameters. Time series analysis of the group wise simultaneous sessions resulted in an MA(2) model after second differencing. However, the time series analysis of the bit volume process did not yield any ARIMA model.

For *kiwitraces* we encountered around 15000 different sources. After eliminating three machine sources we found that 1190 sources contribute to 90% of the total utilisation. Users are grouped into three categories as follows:

High : 5 - 50
Moderate : 101 - 250
CHAPTER 7.  TIME SERIES ANALYSIS

Low: 400 - 600