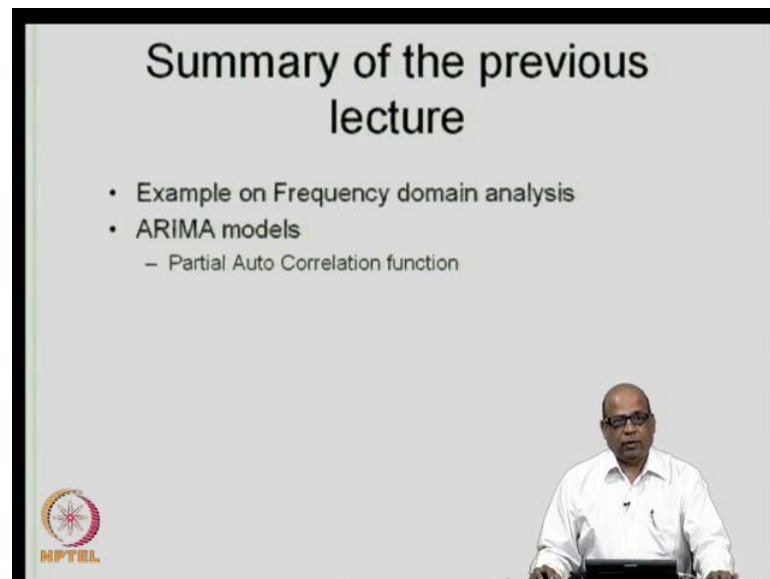


**Stochastic Hydrology**  
**Prof. P. P. Majumdar**  
**Department of Civil Engineering**  
**Indian Institute of Science, Bangalore**

**Lecture No. # 15**  
**ARIMA Models – II**

Good morning and welcome to this lecture number 15 of the course, Stochastic Hydrology.

(Refer Slide Time: 00:24)



The slide is titled "Summary of the previous lecture" and contains the following bullet points:

- Example on Frequency domain analysis
- ARIMA models
  - Partial Auto Correlation function

In the bottom right corner of the slide, there is a small inset video showing the professor, Prof. P. P. Majumdar, wearing glasses and a white shirt, sitting at a desk. In the bottom left corner of the slide, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) featuring a stylized sun and the text "NPTEL".

If you recall in the last lecture, we continued our earlier discussion on frequency domain analysis. We introduced the spectral density function and then in the last lecture, we solved an example starting with the monthly stream flow data. Then we plotted the correlogram for the monthly stream flow data. We examined from the time series plot and the correlogram that there are periodicities indicated in the data. So, the monthly data, time series data, when we express in the frequency domain and carry out the spectral density analysis, spectral analysis, the line spectrum as well as the power spectrum show prominent peaks in the spectral densities indicating that the periodicities

that were shown up by the correlogram as well as the time series plot are indeed present in the data.

So, the spectral analysis actually brings to the forth the periodicities present in the data and through the example, we could see that the periodicities in the monthly data that we considered were present at period of 12 months, 6 months 4 months and 3 months how do we identify this we look at the spikes provided in the line spectrum or the power spectrum the associated  $\omega$  value the omega value we convert that into the corresponding periodicity that is  $2\pi$  by  $\omega$ .

So, we identified that there are periodicities corresponding to twelve months 6 months 4 and 3 months for the monthly stream flow data that we considered then in the same example we had also could see how many of these periodicities are infect statistically significant. So, we introduced a statistical test by which you can examine the periodicities that you identified through the spectral analysis whether those periodicities are statistically significant or not in the same example what we then did is that we converted the stream flow data to a standardized series by deducting the mean of the associated month and by dividing by the standard deviation of the corresponding month and then carried out the same analysis of correlogram of plotting the correlogram and the spectral density function we saw that the periodicities that were shown in the original data were absent in the standardized data.

(Refer Slide Time: 04:54)

### ARIMA Models

Box Jenkins Time series models:

- For stationary time series
- If the time series is stationary, the correlogram dies down fairly quickly (e.g., within 4 or 5 lags, in most hydrologic applications)
- If the time series is non stationary, the decay is very slow

NPTEL

Then we also introduced to the arima models subsequent to that example we introduced to the arima model how to formulate the arima model if you recall we said arima model that is auto regressive integrated moving average model. So, you have the auto regressive terms and the moving average terms and the term I there indicates the order of differencing.

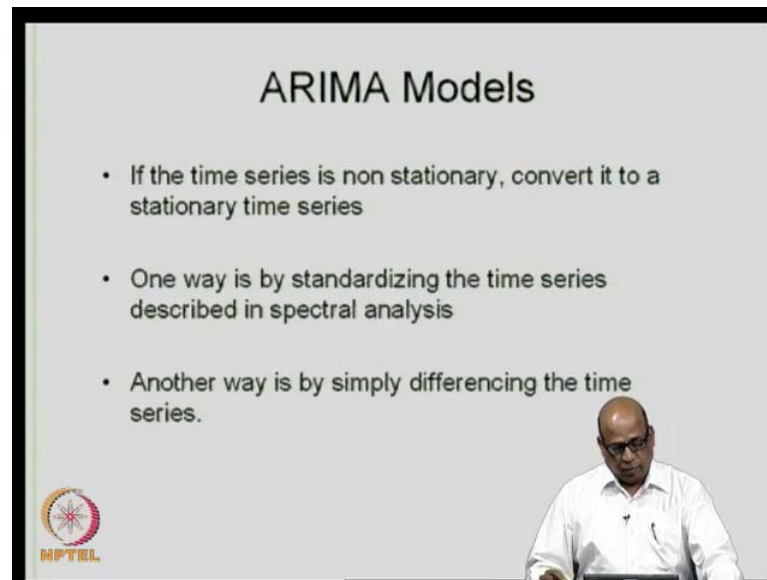
In discussing the arima models we also introduced the concept of partial auto correlation if you recall the partial auto correlation indicates the relationship between the dependant variable let say  $X_t$  on the independent variable  $X_{t-1}$  or  $X_{t-k}$  let us say you are looking at lag  $k$  when it is dependants that is  $X_t$ 's dependants on all the terms are partial out or removed out. So, this indicates the partial auto correlation the partial auto correlation in addition to the correlogram and spectral analysis gives some other information that is present in the time series.

The time series plot itself the correlogram the spectral density and the partial auto correlation function all of these together will tell us which kinds kind of models are useful for the particular data. So, in today's lecture we will continue with that discussion and see how we formulate the arima models auto regressive integrated moving average models and perhaps how we identify which type of models how many terms in the models are there needs to be included how many A R terms need to be included how many of MA terms need to be included in the model for the specific application in question.

In the last lecture I just mentioned that these are the type of models that I will be covering in the course are called as the box Jenkins time series models and they are valid that the coverage that we will do in this course is only meant for stationary time series we do not cover the non stationary time series and therefore, your original time series if it is non stationary you must first convert that into a stationary term series which means first you have to identify whether the series that you are considering is stationary or non stationary if you look at the correlogram if the series is non stationary then the correlogram decays rather slowly where as if the stationary in the case of stationary time series the decay is quite fast for example, here i have shown a correlogram of a stationary time series the decay is quite rapid whereas, for the case of non stationary time series the decay is rather slow indicating that the dependence does not die down quickly and therefore, it becomes a non stationary time series.

So, the first indication is of non stationary is if the correlogram does not die down fairly rapidly then you must suspect non stationary in the data then we must have the  $(())$  or the means to remove the non stationary and convert the time series that you have into a stationary time series only then apply the type of models that we will be discussing now.

(Refer Slide Time: 06:51)



**ARIMA Models**

- If the time series is non stationary, convert it to a stationary time series
- One way is by standardizing the time series described in spectral analysis
- Another way is by simply differencing the time series.

NPTEL

One way of removing the non stationary is by standardizing the time series in the example that we saw in the last lecture we saw that once you remove the once you standardize the series and plot the spectral density as well as the correlogram it indicates that there is no there is no correlation present in the data or the data becomes random and the series becomes stationary.

So, one way of doing removing the non stationary is simply standardizing the time series, but, in the arima models we also considered what is called as the differencing the differencing the time series first order differencing second order differencing etcetera which I will introduce presently this is a simple way of removing the periodicities and making the time series a stationary time series not only the periodicities if you have trends trends can also be removed by differencing.

(Refer Slide Time: 08:07)

The slide is titled "ARIMA Models". It contains the following text and graphics:

- Differencing:  
$$Y_t = X_t' = X_t - X_{t-1}$$
  
 $X_t'$  is First order differencing
- Below the equations, there are two sets of data and graphs:
  - On the left:  $\{X_t\} = 2, 4, 6, 8, 10, \dots$ . Below this is a graph with a vertical axis labeled  $X_t$  and a horizontal axis labeled  $t$ . A straight line starts from the origin and slopes upwards, representing an increasing trend.
  - On the right:  $\{Y_t\} = 2, 2, 2, \dots$ . Below this is a graph with a vertical axis labeled  $Y_t$  and a horizontal axis labeled  $t$ . A horizontal line is drawn at a constant value, representing a stationary time series.
- In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning).

So, in general we use a differencing to convert the non stationary time series into a stationary time series what I mean by differencing is that if you have a series  $X_t$  you take the first difference; that means,  $X_t - X_{t-1}$  just deduct the previous value and compute the new series  $Y_t$  constitute the new series  $Y_t$ . So, if you have  $X_t - X_{t-1}$ ; that means, the first order differencing where you are simply deducting the previous term this is called as the first order differencing what does it do for example, if you have a series like this 2 4 6 8 10 12 and. So, on if you plot  $X_t$  versus  $t$  you have an increasing trend in the data 2 4 6 8 and. So, on let say I do the first order differencing here then what I will get 4 minus 2 6 minus 4 8 minus 6 10 minus 8 etcetera. So, the new series will consist of two two two two and.

So, one. So, if you plot now  $Y_t$  versus  $t$  you have a horizontal line this is; obviously, stationary. So, what was originally non stationary just by first order differencing in this particular example we have converted into a stationary time series. So, in general the differencing has the effect of removing now some amount of non stationary in the data why I said some amount of non stationary is that in this particular case all the non stationary has been removed, but, in the actual data when we are considering let us say stream flow at a particular location and. So, on depending on the strength of the periodicity the strength of the trend etcetera that are present in the nature of trend the nature of periodicities that are present not all the non stationary may be removed just by the first order differencing may be you all have to go to the second order third order

etcetera even then it may not be possible to remove completely non stationary in the data in which case we try other methods for example, standardization and. So, on.

As we explain in the previous lecture the presence of non stationary in the data can be examined by various statistical test. So, once you do the differencing and constitute the new time series on the new time series newly constituted time series by differencing you again do the analysis of correlogram and then the spectral density spectral analysis and. So, on to examine whether this series that you constituted now is. In fact, stationary when when you achieve the satisfactory degree of stationary then you can use the type of models that we will be discussing in the course.


(Refer Slide Time: 10:57)


**ARIMA Models**

$$X_t' = X_t - X_{t-1}$$

$X_t''$  is Second order differencing

$$\begin{aligned} X_t'' &= X_t' - X_{t-1}' \\ &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\ &= X_t - 2X_{t-1} + X_{t-2} \end{aligned}$$





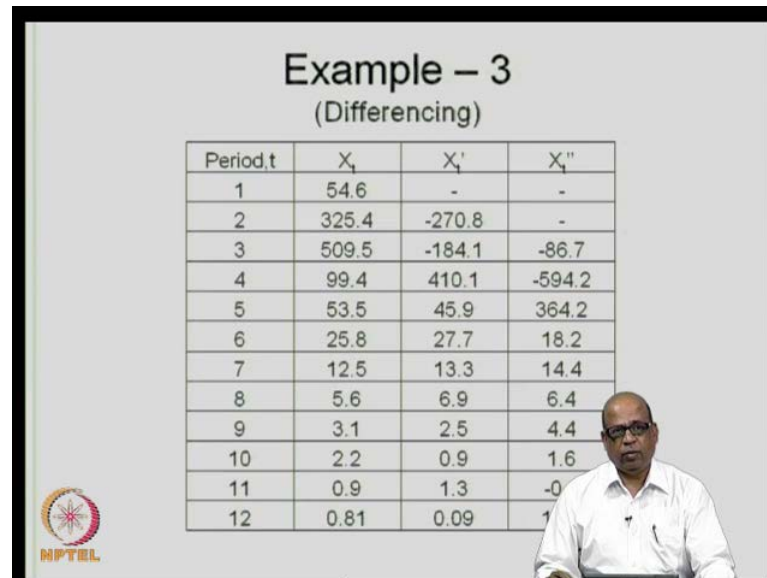
So, similar to the first degree of first order of differencing you also have second order of differencing where you take the first difference of the difference series itself the x dash t here is X t minus X t minus 1. So, X dash t is the first order difference series in the second order differencing you take the differencing on the already difference series. So, x double dash t which is the second order differencing is given by X dash t minus X dash t minus 1. So, if you write like that if you write it in a long form. So, what is X dash t x dash t is X t minus X t minus 1 this is the first order differencing on X t and minus x dash t minus 1 is the first order differencing on X t minus 1 which is X t minus 1 minus X t minus 2 you are taking the first order differencing on X t minus 1 which means this will be X t minus 2 X t minus 1 plus X t minus 2. So, x dash x double dash t which is the

second order differencing can be written as  $X_t - 2X_{t-1} + X_{t-2}$  like this you carry on the third order differencing 4th order differencing etcetera in general in hydrology most hydrological applications where do we have autoregressive moving average type of models we go typically up to the second order differencing not more than that.

(Refer Slide Time: 12:30)

**Example – 3**  
(Differencing)

Period, t	$X_t$	$X_t'$	$X_t''$
1	54.6	-	-
2	325.4	-270.8	-
3	509.5	-184.1	-86.7
4	99.4	410.1	-594.2
5	53.5	45.9	364.2
6	25.8	27.7	18.2
7	12.5	13.3	14.4
8	5.6	6.9	6.4
9	3.1	2.5	4.4
10	2.2	0.9	1.6
11	0.9	1.3	-0.4
12	0.81	0.09	1.21



Let's look at one example here let us say you have this time series this is the same time series that we have considered earlier for 12 month I have shown here  $X_t$  these are the observed values let say observed monthly stream flow values you take the first differencing. So,  $X_t - X_{t-1}$  you get minus 270.8 then  $X_t - X_{t-2}$  here that is  $X_2 - X_3$  you get minus 184.1  $X_3 - X_4$  you get 410.1 and. So, on like this it you get the first order differencing.

Then for the second order differencing you do  $X_{t-2} - X_{t-1}$  which means  $X_2 - X_3$ . So, you get minus 86.7 because these are both are negative then this minus this that is minus 594.2 and. So, on. So, you get the second order difference series. So, like this from the original series you can get the first order difference series second order difference series and. So, on then you can examine whether this series that you obtained by first order differencing is in fact stationary or this is stationary if this is non stationary still you go on to the higher order differencing series and examine whether this is stationary still not satisfied then you go to the next order



differencing which is the third order differencing in this case and then examine whether that is stationary. So, on; obviously, I have shown only twelve values here, but, this has to be done for a longer series typically we may have a monthly monthly data for 50 years 60 years etcetera on that series you have to do this examine test of stationary or lack of it.

(Refer Slide Time: 14:33)

**Example – 4**

Monthly Stream flow (in cumec) statistics(1979-2008) for a river is selected for the study. (Part data shown below)

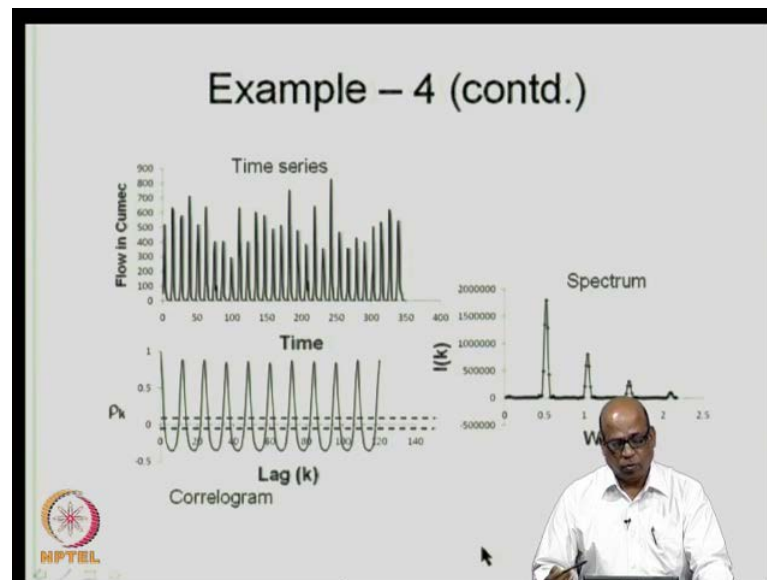
Year	Month	S.No.	Flow
1979	June	1	54.6
	July	2	325.4
	August	3	509.5
	September	4	99.4
	October	5	53.5
	November	6	25.8
	December	7	12.5
1980	January	8	5.6
	February	9	3.1
	March	10	2.2
	April	11	0.9
	May	12	0.81



Now, we will consider the same data that we considered in the example of spectral analysis. So, there are 300 and 48 values only 12 values I have shown here. So, this is between 1979 and 2008 this is monthly stream flow data. So, you have n is equal to 348 only part data is shown and the time series plot etcetera is shown earlier. So, this is the time series plot that you have for 300 and 48 values.



(Refer Slide Time: 14:58)

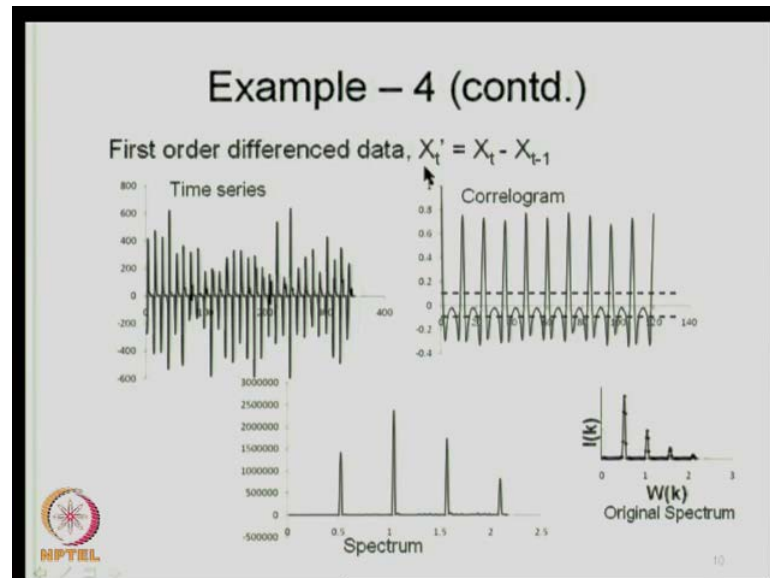


We have also shown earlier the correlogram for this as well as the spectral density for this the correlogram indicates that there is a significant periodicity present I again repeat that right from the time series plot you suspect that there is a periodicity because just if you look at the time series plot there is A Regular pattern the flows are increasing and then decreasing and. So, on in a fairly good regularity and that information is also seen in the correlogram which indicates that there is a periodicity present in the data we want to verify this and pin point exactly where the periodicity is present are present and therefore, we convert this into the frequency domain and carry out the spectral spectral analysis plot the line spectrum or the power spectrum which brings out the spikes and these spikes correspond to the periodicities and this periodicity is of 12 month periodicity and this is 6 months and this is 4 months and this is 3 months.

Which means that we have now seen that this is time series that we are considering is the not stationary time series because there are significant periodicities present in the data why do I say significant because we also examined for the periodicities that were identified in the spectral analysis corresponding to 12 months 6 months 4 months 3 months all of them were statistically significant recall that we formulated a statistic corresponding to these and then compared it with the f distribution with 2 degree of freedom and then concluded that all the periodicities that we had identified here are. In fact, statistically significant now the question is if you want to apply the time series models which are meant for stationary time series on this particular time series which

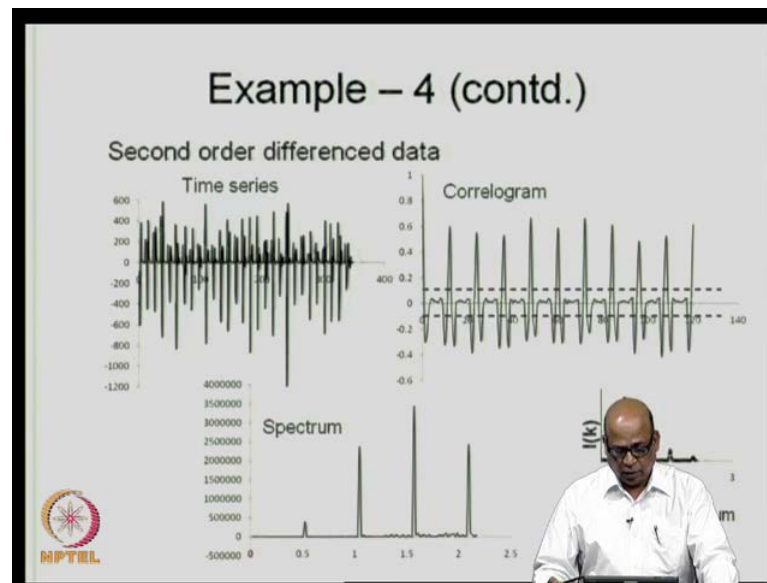
you know is. In fact, non stationary because of the presence of significant periodicities in the data then you have to convert that time series into non stationary into a stationary time series. So, let us see what happens if you do the first order differencing on the time series that we had consider. So, what is the first order differencing it is simply  $X_t - X_{t-1}$ .

(Refer Slide Time: 17:40)



So, let us look at what happens here for the difference first order difference data. So, in the original data which is shown here all I do is take  $X_t$  and deduct  $X_{t-1}$ . So, I formulate a new series  $X_t'$  as  $X_t - X_{t-1}$ . So, corresponding to the original series you have a new series now and that series I plot here. So, this is the time series of the first order difference data  $x_t'$  I plot this time series which is different from the original time series like this, but, still you see some kind of a pattern here that is the values are increasing periodically increasing with some regularity and then decreasing with some regularity and. So, on the corresponding correlogram appears like this again indicating that there are still periodicities present here and the line spectrum appears like this the original line spectrum was like this. So, the line spectrum is no different not much different in terms of its indication of the periodicities. So, the first order difference data still indicates that there is some non stationary present in the data.

(Refer Slide Time: 19:08)



We will go to the second order differencing now what do I do in the second order differencing I take  $X_t - X_{t-1}$  now when I do the same analysis on the second order difference data the time series appears like this which is different from what was there for the first order difference data the correlogram appears like this still indicating that there are significant periodicities that are present in the data and the spectral density this is original spectral density here and this is the spectral density for the difference data it again indicates that there may be some periodicity of course,, we need to test for the significance statistical significance of these periodicities, but, there is an indication that the the time series that you. So, formulated by taking the second order differences still is not devoid of periodicities.

And let us examine what happens in the third order differencing. So, because we are not satisfied with what we did in the second order differencing we go to the third order differencing what do I do in the third order differencing I take  $x_{t-2} - x_{t-1}$ . So, in the first order we take  $X_t - X_{t-1}$  in the second order we take  $x_t - x_{t-1}$  by and this we call it as  $x_{t-2}$  which is a second order differencing in the third order differencing I do the difference on the second order difference series which is  $x_{t-2} - x_{t-1}$  that is I will take the difference on this difference series itself when I do that I get the third order difference series you get the data like this the time series data like this and the correlogram is like this again it indicates these are the 95 percent significance lines recall

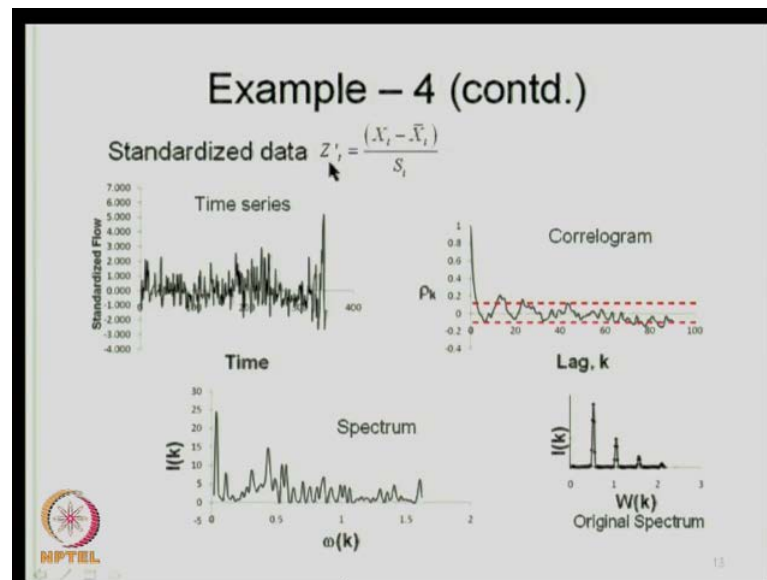
that the 95 percent significance lines you form by taking plus minus 1.96 by root n and for most practical purposes you take two by root n that is plus minus 2 by root n where n is the the number of sample data in which in this case it is 300 and 48. So, these are your significance bands. So, it indicates still that there are still periodicities present here and even the spectral density shows that there are periodicities present in the data.

Again for comparison this is original spectrum here and this is your new spectrum in the original spectrum you had a spike corresponding to point 5 3 or something indicating a periodicity of 12. 12 months, but, that is absent here where as the spike corresponding to 1.2 or around that is still present and these two have come up now corresponding to 2.2 the spike has become more prominent in the third order difference data.

As we saw in the example that we I discussed the spectral analysis the fact that these are prominent spikes does not necessarily mean that these are statistically significant periodicities we need to examine for the statistical significance of these it may. So, happen that these are statistically insignificant, but; however, you are seeing from the correlogram as well as from the spectral analysis that the periodicities are not entirely removed specially because the correlogram shows that there are significant correlation still present in the data and therefore, it indicates that the periodicity may still be present the data and therefore, now let us see what happens if we do the standardization on this data if you do the standardization you may perhaps end up with a completely random series devoid of any periodicities of course,, with the third order differencing you may have significant correlation which you may use in the models because you are talking about correlated data when you are looking at time series models.

In doing standardization essentially what we were doing is you are taking out the mean you are deducting the mean and dividing by the standard deviation.

(Refer Slide Time: 23:59)



So, if you look at the standardized data  $Z'_t$  is equal to  $X_t - \bar{X}_I$  by  $S_I$  recall that this  $I$  is the month corresponding to the time period  $t$ . So, in this case  $t$  goes from 1 to 300 and 48 and  $I$  goes from 1 to 12. So, corresponding to every  $t$  you have an association with the particular month in the year and you are taking out the particular mean of that month and dividing it by the standard deviation of that month when you standardize the series you will see that the time series looks more or less random like this and most of the correlations are all insignificant there is hardly any correlation that is outside the significance bar except for the first one and the line spectrum shows that the data is random.

So, there are no specific spikes here which are much different from the other spikes compared this to the original spectral density you see that this spectral density is indicates much more random data corresponding to in comparison with the original data.

So, standardization of it the data indicates that the by standardizing we have removed the periodicities present in the data now we can use the time series models on the standardized data or you can also use on your third order difference data if you are sure that by doing this you have removed this periodicities; that means, periodicities that are coming up now are statistically insignificant. So, we could have used you could use the arima type of models on the third order difference models where the difference the order of differencing is third order.

(Refer Slide Time: 26:04)

The slide is titled "ARIMA Models". It contains the following text and equations:

- Operator 'B':  
The effect of operator 'B' is to shift the argument to that one step behind.

$$BX_t = X_{t-1}$$
$$BX_{t-1} = X_{t-2}$$

AR (1) Model:

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$
$$X_t = \phi_1 BX_t + \varepsilon_t$$
$$X_t(1 - \phi_1 B) = \varepsilon_t$$

AR (1) component

The slide also features the NPTEL logo in the bottom left corner and a small inset image of a man in a white shirt in the bottom right corner.

Now, we introduce another important comments. So, what what did we do now we saw methods by which you can remove the trend or the periodicities present in the data by differencing first order differencing second order differencing and. So, on the presence of the periodicities let us say that even after differencing you still have certain periodicities present in the data **indicating** indicating that there are certain lag correlation at particular lags which are still quite significant now these can be addressed by arima models by introducing those a particular lag terms I will discuss this in greater detail later on where we are talking about contiguous and in contiguous herm models where specific terms can be included without including the previous terms anyway right now we will not worry too much about it, but, right now what we will do is we will introduce an a interesting operator called the operator b which is useful in writing down the arima models in more compact and relevant forms.

The effect of the operator b is to shift the argument to that one step behind simply shifted one step behind for example, when I use the operator b on X t it is simply equal to X t minus 1 when I use the operator b on X t minus 1 it simple X t minus 2. So, the operator b shifts the argument to one step behind that is all.

Now, this becomes a very handy tool in in expressing the various arima models in more compact and relevant forms say for example, you have the A R 1 model auto regressive model of order one it is written as X t is equal to phi 1 X t minus 1 plus epsilon t this is

our original AR(1) model. So, I will use the B operator now. So,  $X_t$  is equal to  $X_{t-1}$  plus  $\phi_1$  times  $X_{t-1}$  plus  $\epsilon_t$ . I will write it as  $B X_t$ . So,  $\phi_1$  into  $B X_t$  plus  $\epsilon_t$  that is  $X_t$  is equal to  $\phi_1 B X_t$  plus  $\epsilon_t$ . So, I will take all the  $X_t$  terms on one side. So, I will write it as  $X_t$  into  $1 - \phi_1 B$  is equal to  $\epsilon_t$ . So,  $1 - \phi_1 B$  becomes the term for AR(1) component. So, we are saying  $X_t$  into the AR component is equal to  $\epsilon_t$ .

(Refer Slide Time: 28:58)

**ARIMA Models**

AR (2) Model:  $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \epsilon_t$   
 $X_t = \phi_1 B X_t + \phi_2 B X_{t-1} + \epsilon_t$   
 $X_t = \phi_1 B X_t + \phi_2 B^2 X_t + \epsilon_t$   
 $X_t (1 - \phi_1 B - \phi_2 B^2) = \epsilon_t$   
AR (2) component

Generalized form for an AR(p) model is

$$X_t \left( 1 - \sum_{i=1}^p \phi_i B^i \right) = \epsilon_t$$

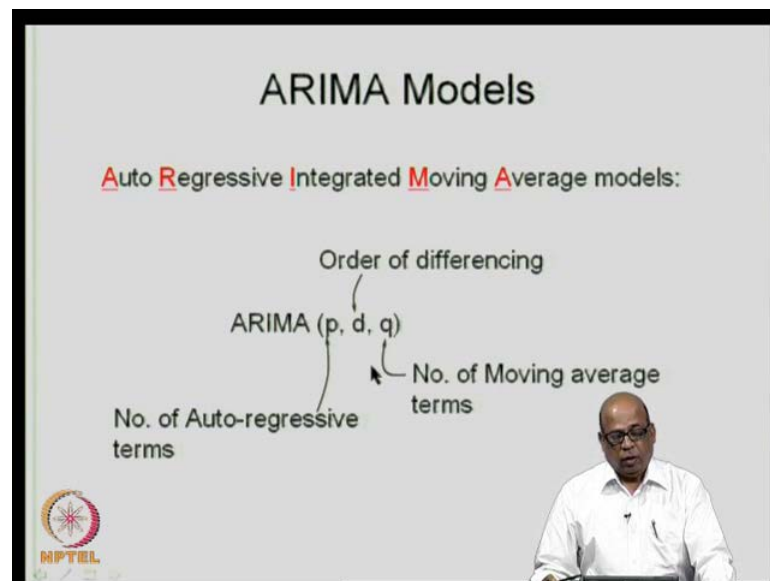
Let's say we want to write for AR(2) model what is AR(2) model it is  $X_t$  is equal to  $\phi_1 X_{t-1}$  plus  $\phi_2 X_{t-2}$  plus  $\epsilon_t$ . So, now, I will use the B operator wherever there is a  $X_{t-1}$  term I will put B into  $X_t$  and  $X_{t-2}$  term I will put it as  $B^2$  into  $X_t$ . So, I will write this as  $X_t$  is equal to  $\phi_1 B X_t$  plus  $\phi_2 B^2 X_t$  plus  $\epsilon_t$ . So,  $X_t$  minus  $\phi_1 B X_t$  minus  $\phi_2 B^2 X_t$  is  $\epsilon_t$ . So,  $X_t$  into  $1 - \phi_1 B - \phi_2 B^2$  is  $\epsilon_t$ . So, we write the AR models as  $X_t$  into some term in terms of B and in terms of the associated parameters and on the right side we keep the noise term  $\epsilon_t$  for the AR models. So, the this becomes AR(2) component compared this to AR(1) component this is  $1 - \phi_1 B$  and for the AR(2) component  $1 - \phi_1 B - \phi_2 B^2$ .

So, I will further expand this. So, this is  $\phi_1 B$  into  $X_t$  plus  $\phi_2 B^2$  this was  $B X_{t-1}$  plus  $B^2 X_{t-2}$ . So, I will write that as B into  $B X_t$  because  $X_{t-1}$  is  $B X_t$ . So, together I will write this as  $B^2 X_t$  that has to be  $X_{t-2}$ . So, this is  $X_t B^2 X_t$  plus  $\epsilon_t$  and therefore, I will write it as  $X_t$  into  $1 - \phi_1 B - \phi_2 B^2$  is equal to  $\epsilon_t$ . So, we write the AR models as  $X_t$  into some term in terms of B and in terms of the associated parameters and on the right side we keep the noise term  $\epsilon_t$  for the AR models. So, the this becomes AR(2) component compared this to AR(1) component this is  $1 - \phi_1 B$  and for the AR(2) component  $1 - \phi_1 B - \phi_2 B^2$ .

$\phi^2 b^2$ . So, in general you can say you can write it as  $X_t - \phi X_{t-1}$  into look at this for the second term what a what we have  $1 - \phi$  for the second term  $I$  is equal to  $1 - \phi + \phi^2 - \phi^3 + \dots$  to 2 of  $\phi^i b^i$  to the power  $I$ .

So, for a  $p$ th model we can write for the  $p$ th order AR model we write this as  $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \dots - \phi_p X_{t-p}$  into bracket  $1 - \sum_{i=1}^p \phi_i B^i$  is equal to  $1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$ . So, many terms of the AR model  $\phi_1 \phi_2$  and. So, on here  $b$  to the power  $I$ . So,  $b$  to the power  $1 + b$  to the power  $2$  negative is outside of the summation and right side is  $\epsilon_t$ . So, this is how we express the AR model in a more compact form using the  $B$  operator. So, any time you see the models expressed like this you immediately must relate it with AR models and the order of the AR model is given by how many terms you have in this. So, there are  $p$  terms here for an AR  $p$  model.

(Refer Slide Time: 32:05)



Now, recall that we said auto regressive integrated moving average models what we explained. So, far was AR models that is you do not have differencing you do not have MA terms MA terms is moving average terms. So, these are arima if you want to write it in the general form  $ARIMA(p, d, q)$ ; that means, there is no differencing involved there is order of differencing is 0 and there are no MA terms involved therefore, the order of MA terms is zero. So, AR  $p$  model is  $ARIMA(p, 0, 0)$  in general we write the arima model as  $ARIMA(p, d, q)$  it means that you have corresponding to auto regressive you have  $p$  number of terms the order of differencing is denoted by  $d$  and the number of moving average



terms is  $q$ . So,  $AR(p, d, q)$  indicates  $AR$  you have  $p$  number of terms of  $AR$  the order of differencing is  $d$  and the number of  $MA$  parameters is  $q$ . So, this is the general notation that we follow for any  $ARIMA$  model.

Given a time series  $X_t$  and the given order of differencing first you do the differencing on the time series of that particular order and then apply an  $ARMA$  model. What do I mean by  $ARMA$  model?  $ARMA$  model is with zero differencing; that means, you will have  $ARMA(p, q)$ . So, first you do the differencing and then write the associated  $ARMA$  model. We will discuss this through some examples later on subsequently in the lecture, but, we will continue our discussion on how to apply the  $B$  operator. So, let us say that you have an autoregressive moving average model; that means, you have done the differencing already there is no differencing you are simply writing the  $AR$  parameters and the  $MA$  parameters. So, you have  $p$  of  $AR$  parameters  $p$  number of  $AR$  parameters and  $q$  number of  $MA$  parameters. So, in general  $ARMA(p, q)$  model is written as  $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \epsilon_t$ .

(Refer Slide Time: 37:56)

**ARIMA Models**

First order differencing:

$$X_t - X_{t-1} = e_t$$

$$X_t - BX_t = e_t$$

$$X_t(1 - B) = e_t$$

Second order differencing:  $X'_t = X_t - X_{t-1}$

$$= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})$$

$$= X_t - 2X_{t-1} + X_{t-2}$$

$$= X_t - 2BX_{t-1} + X_{t-2}$$

$$= (1 - B)^2 X_t$$

So,  $\phi_p X_{t-p} + \dots$  there are  $q$  number of moving average terms. The moving average in the context of  $ARMA$  models is not the same as the moving average that we discussed earlier when we were talking about taking averages in the windows across which is moving across the time series. This was slightly different from the notation here.

we talk about the moving average in the context of the residuals that results from the model. So, this is you have  $q$  number of moving average terms  $\theta_1 e^{t-1}$  plus  $\theta_2 e^{t-2}$  plus etcetera plus  $\theta_q e^{t-q}$ . So, you have  $q$  number of moving average parameters  $\theta_1$   $\theta_2$  etcetera up to  $\theta_q$  and you have  $p$  number of auto regressive parameters  $\phi_1$   $\phi_2$  etcetera  $\phi_p$  plus the noise term or the residual term  $e_t$  now remember see here notice that  $e_{t-1}$   $e_{t-2}$  etcetera up to  $e_{t-q}$  is what you have written for  $X_t$ .

So, when you go to  $X_{t+1}$  the next term this  $e_t$  gets into the MA parameters here and then you will have a term of  $e_{t+1}$  when we do the numerical example it will be clear of how we account for the residuals in the moving average terms now these are. In fact, the residual terms that you have here are. In fact, important in examining whether the model that we have fit to a particular data passes all the tests or not.

So, we need to do the statistical test on the residual series after we fit the model on a particular time series let us say you have a stream flow every monthly stream flow for last fifty years data and then you have estimated the parameters and then fit the model and then you get the residual series on the residual series you need to do the test.

The assumptions that are involved in this model are that the series of residuals it has a zero mean and they are all uncorrelated in the numerical example we will show how to do the test on the series of residuals that you get. So, this how a general arma  $p$   $q$  model is written you have  $p$  parameters  $p$  terms of auto regressive terms and you have  $q$  terms of moving average terms plus the noise  $\epsilon_t$  or  $e_t$ .

Let us see how we express this using the difference using the  $b$  operator for which let us say you are talking about a term of first order differencing  $X_t - X_{t-1}$  is equal to  $e_t$  this is a your what is the model here you have only integration you have neither the AR terms nor the MA terms. So, you have only the integration which means in our arima  $p$   $d$   $q$  notation what does this mean  $p$  is 0  $q$  is 0 and  $d$  is 1. So, **only-** you are only doing the differencing of the series. So,  $X_t - X_{t-1}$  is equal to  $e_t$  if you write like that for clarity let me write it down. So, this is nothing, but, arima  $p$   $d$   $q$  or  $p$  is 0 here you do not have any AR terms  $d$  is 1 you are doing the first order differencing and  $q$  is 0 there are no MA parameters. So, this is arima zero one zero model how I had written  $X_t$  is equal to  $X_{t-1}$  plus  $e_t$  that is all. So, you are doing the differencing

here now let me express this in terms of the  $b$  operator. So,  $X_t - X_{t-1}$  is  $b$  into  $X_t$  will be equal to  $e_t$ . So, you are putting the operator  $b$  on  $X_t$  to get  $X_{t-1}$  that is  $X_t$  into  $1 - b$  is equal to  $b X_t$ . So, this is how the first order differencing looks let us see how the second order differencing looks let us say that I am not writing any arima model here I am simply looking at the second order differencing I want to express this using the operator  $b$ .

So, this is  $X_{t-2}$  which is a second order differencing is equal to  $X_{t-1} - X_{t-2}$  the second order differencing what is  $X_{t-1}$  it is the first order differencing therefore, I will write this as  $X_t - X_{t-1}$  and what is the  $X_{t-1}$  it is  $X_t - X_{t-2}$  that is the first order differencing on  $X_{t-1}$ . So, this will be  $X_t - X_{t-1} + X_{t-2}$  that is  $X_t$  minus what is  $X_{t-1}$  that is  $b$  into  $X_t$  plus what is  $X_{t-2}$   $X_{t-2}$  is  $b$  into  $X_{t-1}$  and  $X_{t-1}$  there is again  $b$  into  $X_t$  therefore, this will be  $b^2 X_t$ . So, if you take out  $X_t$  outside what you are left with  $1 - 2b + b^2$  which is  $(1 - b)^2$  the whole square into  $X_t$ .

So, the first order differencing was  $X_t$  into  $1 - b$  the second order differencing is  $(1 - b)^2$  into  $X_t$  that is  $X_t$  into  $(1 - b)^2$  here  $X_t$  into  $(1 - b)^2$  whole square if we take the third order differencing and do the same exercise again you will get  $X_t$  into  $(1 - b)^3$  fourth order  $X_t$  into  $(1 - b)^4$  and..

(Refer Slide Time: 41:23)

## ARIMA Models


In general  $d^{\text{th}}$  order difference is  $(1-B)^d X_t$

ARIMA (1, 1, 1)  $Y_t = X_t - X_{t-1}$

$$Y_t = \phi_1 Y_{t-1} + \theta_1 e_{t-1} + e_t$$

$$X_t - X_{t-1} = \phi_1 (X_{t-1} - X_{t-2}) + \theta_1 e_{t-1} + e_t$$

$$X_t - BX_t = \phi_1 (BX_{t-1} - B^2 X_{t-2}) + \theta_1 B e_{t-1} + e_t$$

$$X_t (1 - B - \phi_1 B + \phi_1 B^2) = e_t (1 + \theta_1 B)$$


So, on. So, in general the  $d$  eth order difference which we need for arima  $p$   $d$   $q$  models in general the  $d$  eth order differencing can be expressed as  $X_t$  into  $1 - B$  to the power  $d$ . So, this how we get the  $d$  eth order differencing in terms of the operator  $B$ .

Let us look at the arima one one one model; that means, the first order differencing is done on the series and then you apply the arma one one model. So, as I said any time you have the order there is a differencing order present that is any time when  $d$  is not equal to 0 first you carry out the differencing on the original time series. So, we first carry out first the differencing of this order. So, this is first order differencing. So,  $Y_t$  is equal to  $X_t$  minus  $X_{t-1}$  will constitute another series now which is a first order difference series and then apply the arma model of order one one on the series  $Y_t$  on the difference series  $Y_t$ .

So, I write  $Y_t$  is equal to  $X_t$  minus  $X_{t-1}$  and then write an arma model with 1 A R term and 1 MA term. So, on  $Y_t$  i write this as  $Y_t$  is equal to  $\phi_1 Y_{t-1}$  plus  $\theta_1 e_{t-1}$  plus  $e_t$ . So, you have 1 A R parameter and 1 MA parameter plus the noise term here and this you are writing it on the difference series  $Y_t$  equal to  $X_t$  minus  $X_{t-1}$  and this difference order is one here.

If you had an order two first you reconstitute the series  $Y_t$  by taking the second order difference and then you write the arma model on the difference series that you obtain.

So, first you carry out the differencing and then write the arma model on the difference series.

So, now we will use the  $b$  operator to express this model what is  $Y_t$  now I am just expanding  $Y_t$  is  $X_t - X_{t-1}$  is equal to  $\phi_1 X_{t-1} - X_{t-2}$  because  $Y_{t-1}$  is here plus  $\theta_1 \epsilon_{t-1}$  as it is plus the noise term  $\epsilon_t$ .

$X_t - X_{t-1}$  this I will write it as  $X_t - X_{t-1} = B X_t$  is equal to  $\phi_1 X_{t-1} - X_{t-2} = B X_t - X_{t-2}$  as we have done earlier plus  $\theta_1 \epsilon_t$  remember the  $b$  operator operates on any of the terms and it simply shifts a particular argument to one time step behind. So, when  $b$  operator operates on  $\epsilon_t$  you get  $\epsilon_{t-1}$ . So, I will write  $\epsilon_{t-1}$  as  $\theta_1 b \epsilon_t$  plus  $\epsilon_t$  as it is. So, collecting all the terms on of  $X_t$  on the left side you write  $X_t$  is equal to  $X_t$  into  $1 - b$  minus  $\phi_1 b$  plus  $\phi_1 b^2$  is equal to  $\epsilon_t$  into  $1 + \theta_1 b$  that is how you express an arima one one one model.

So, given any arima  $p d q$  model you should be able to use the  $b$  operator and express this in a more compact and more elegant form using the  $b$  operator.

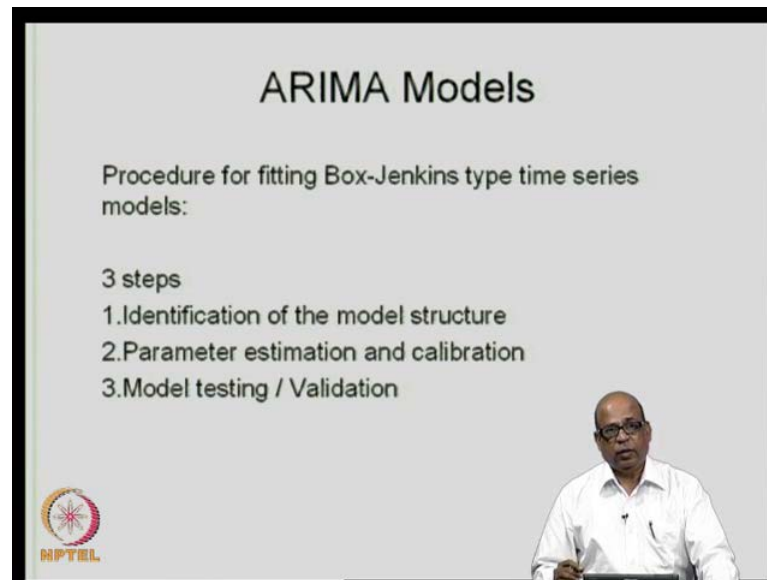
Now we will we now know that given the time series first you identify the order of differencing that you would like to do this the first order differencing second order differencing and. So, on and then identify the number of AR parameters and the numbers of MA parameters and therefore, you will have a particular model ready with you. So, you have this kind of model.

So, given any structure of arima model you now know how to express that arima model in in a compact form like this using the  $b$  operator now we will address the question of given the time series of a particular variable how do we identify which of these large number of models fit that particular time series or which among these large number of models can be **used-** to represent the particular time series.

What I mean by that is in the general form arima  $p d q$  you virtually have infinitely many models possible  $p$  can vary from one to let us say 20 25 and. So, on although there is you can keep on going depending on the data then similarly,  $q$  MA parameters there particularly keep on going 1 2 3 4 5 6 and. So, on. So, and the order of differencing first

order differencing second order differencing etcetera. So, virtually you can theoretically form infinitely many number of models using this general structure. So, for a given time series which among these infinitely many possible models are. In fact, feasible are can be used for the given time series is our primary question that we need to answer.

(Refer Slide Time: 47:14)




**ARIMA Models**

Procedure for fitting Box-Jenkins type time series models:

3 steps

1. Identification of the model structure
2. Parameter estimation and calibration
3. Model testing / Validation

 HPTEL

Now, in the box jerkin's type of analysis we follow three primary steps three major steps namely identification of the models structure that is  $p$   $d$   $q$  how many terms of  $p$  how many terms of  $d$  that is which order of differencing and how many MA terms that we would like to include. So, first you identify the model's structures.

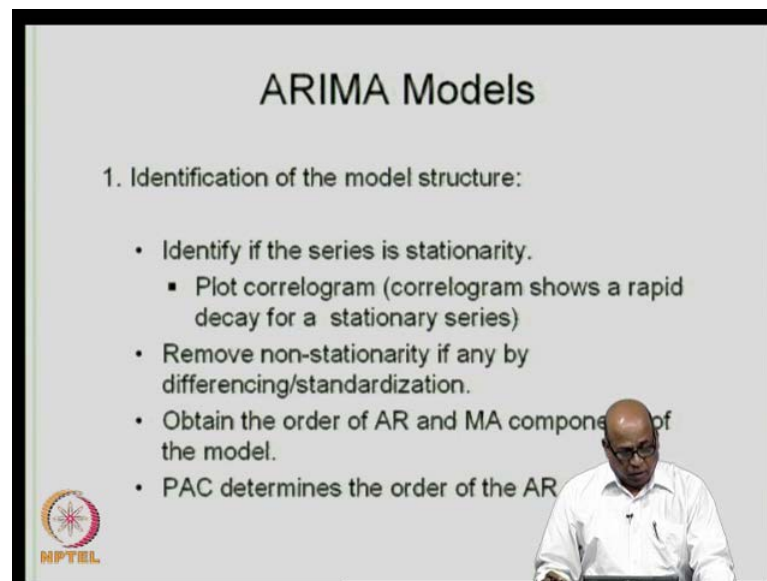
Then once you identify the model's structure let us say that you have identified that there is one A R term one order of differencing and one order of MA term then you are left with the problem of estimating the parameters you have a parameter  $\phi_1$  here you have a parameter  $\theta_1$  here in this particular case. So, in general for A R arma  $p$   $q$  model for example, you will have  $p$  parameters to estimate for the A R terms and  $q$  parameters to estimate for the MA terms.

So, once you write the model's structure you come with the second question of parameter estimation. So, you need to estimate the parameter and then calibrate the model what do I mean by **calibration** calibration that for the data you will have the residuals that arise from the data by applying this particular model that you have identified and those residuals must satisfy the assumptions that we have made for the

residuals namely that they should have a zero mean and they should all be uncorrelated they should constitute series with uncorrelated values then we do the model testing on the remaining data and the validation.

So, these are the primary three steps one is identification model identification another is parameter estimation and calibration and the third one is model testing or validation with the remaining part of the data. So, typically what we do is if you have a series of let say fifty years of monthly stream flow data how many values you have 50 into 2600 values typically you build the model on the first half of the data take  $n$  by 2 values and fit the data do the calibration etcetera and do the validation on the remaining part depending on the length of the data this can be either  $n$  by two that is you build the model on  $n$  by 2 or if the data is fairly small then you do not have fairly numbers of  $n$  by 2 and therefore, you may have to go with 3  $n$  by 4 or 75 percent of the data you use to build the model and the remaining part to test and **validate** validate the model.



(Refer Slide Time: 50:40)



## ARIMA Models

1. Identification of the model structure:

- Identify if the series is stationary.
  - Plot correlogram (correlogram shows a rapid decay for a stationary series)
- Remove non-stationarity if any by differencing/standardization.
- Obtain the order of AR and MA components of the model.
- PAC determines the order of the AR

All this nuances of model building we will discuss we show the application and case studies. So, one where depending on the length of data you may have to sacrifice on the validation part and validation amount of data that is available for validation and. So, on. So, you do not have to be realistically bound by always  $n$  by 2 for model building and model testing and. So, on it it actually depends on the length of data that you have and as

I mentioned right at the first right in the first lecture in hydrology mostly you are constrained by the length of the data that is available to you.

So, identification of the model's structure which means we need to see how many of the auto regressive models or auto regressive terms to be included and how many moving average terms and what level of differencing that you need to do and. So, on.

So, as I mentioned the first step in identification of the model is examining if the series that you have is stationary or not. So, you plot the correlogram if the correlogram shows a very slow decay either you may have sinusoidal correlations or the correlogram indicating sinusoidal variation or it may have a slow decay either exponential decay or normal non-linear non-linear decay if you have a slow decay then it indicates that the series is not stationary once you identify that the series is non stationary then we need to make the series stationary for use of the arima models. So, we may adopt like I just demonstrated you may have adopt differencing that is first order differencing second order differencing etcetera to make sure that you remove non stationary then you obtain the order of A R and MA components for the model. So, once you difference it then you can obtain the order of A R and MA components.

The partial auto correlations that we discussed in the last lecture we introduced in the last lecture is a very handy tool for identifying the A R components if there are significant partial auto correlations present in the data let us say there are only 2 significant partial auto auto correlation or auto partial auto correlation present in the data it indicates an A R model of two may be may be appropriate for the particular time series provided it also satisfies certain conditions on the auto that is correlogram.

So, if you want to identify just the A R components you have to look at both the correlogram as well as the partial auto correlations together if you are auto corral or the correlogram shows a decay and the partial auto correlations show significance presence of one or two or may be more partial auto correlations which are significant then it indicates the A R terms to be of that order let us say you may have A R 2 or a A R 3 depending on how many partial auto correlations are significant or not.

Now, this discussion on identification of exactly how many A R terms and how many MA terms to be included has significance only when you want to you want to short list on a few models and then examine which among them are more appropriate for your



particular application and. So, on we will continue this discussion in the next lecture of how to identify how many of A R terms and how many of MA terms are appropriate for the particular situation.

So, to summarize then in this lecture we discussed how to formulate the arima models the general formulation of arima models and how to do the differencing the first order differencing second order differencing and what effect the differencing has on removal of periodicity. So, we tested with one numerical example that the time series that we considered had significant periodicities present then we did the first differencing second differencing third differencing etcetera whereby we saw that the series becomes the periodicities are removed as we keep doing the differencing.

We also saw that standardization removes periodicities on the same time series we did the standardization by standardization I mean  $Z_t$  which is a standardized series we expressed it as  $X_t - \bar{x}_i$  by  $x_i$  where  $x_i$  in the case of monthly time series  $x_i$  is the mean  $\bar{x}_i$  is the mean of the particular month's flow to which the time  $t$  belongs then we wrote the arima models using the  $b$  operator. So,  $B X_t$  is equal to  $X_{t-1}$ . So, the effect of the  $b$  operator on any argument is simply to shift the argument one time step behind. So, we we examined how to write a general arima model using the  $b$  operator in a more compact form then towards the end of the lecture we has we have just listed out the steps that are involved one is identification of the model by which I mean identification how many A R terms and how many MA terms and what is the level or what is the order of differencing that you need to do

. So, this gives the identification of the structure of the model then we do the parameter estimation and calibration in the second step and in the third step we do the validation and model testing. So, in the next lecture we will see details of each of these three steps how do we identify how do we estimate the parameters and how do we do the testing. So, we will meet in the next lecture **thank you** very much for you attention..