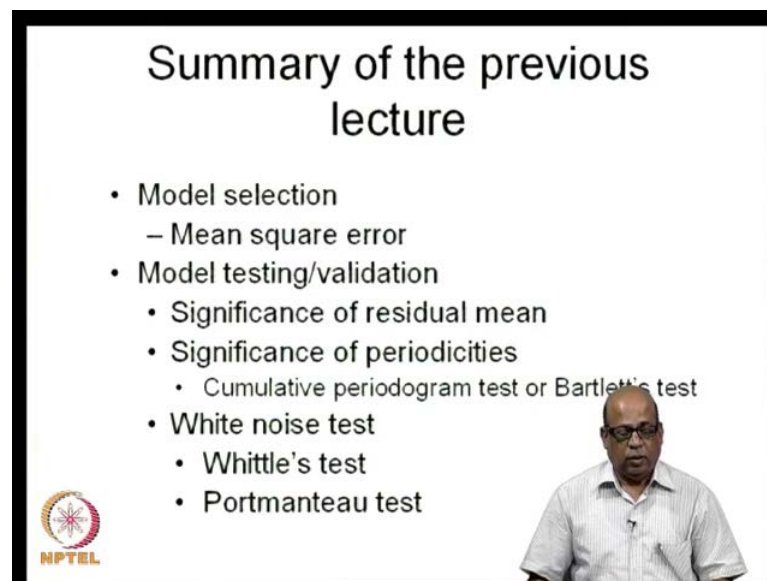


Stochastic Hydrology
Prof. P.P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore



Lecture No. # 18
Case Studies – I

(Refer Slide Time: 00:25)



Summary of the previous lecture

- Model selection
 - Mean square error
- Model testing/validation
 - Significance of residual mean
 - Significance of periodicities
 - Cumulative periodogram test or Bartlett's test
 - White noise test
 - Whittle's test
 - Portmanteau test

Good morning and welcome to this the lecture number 18, of the course, “stochastic hydrology”. In the last lecture, we went through the ARIMA model selection, based on the means square error criterion. In the earlier lecture, we had discussed the model selection based on the maximum likelihood criterion. Essentially, what we do is that we select a set of candidate models and then estimate the parameters, using the calibration data, the first part of the data. We use so the parameter estimation methods, typically, the Markov’s algorithm or algorithms that are available in matlab, and then estimate all the parameters of the candidate models.

Then corresponding to each of the candidate models for the validation data, we compute the likelihood values in the maximum likelihood case. In the mean square error case, we calculate the mean square error when the model is applied for the validation period. We choose that particular model in the case of forecasting that particular model which resorts

in the minimum mean square error. In the case of long term synthetic generation of the data, we choose that particular model, which results in the maximum likelihood.

Then in the last lecture, we also take we also discussed the procedure for model testing and validation. Essentially, if you recall, we do all this test on the residuals or the residuals that arise out of application of the model, on the validation data. You have built the model and apply that model for the validation period, because you have the data for the validation period, you obtain the residuals as differences between the actual data, actual observed data and the model simulated data that gives you the residuals.

So, you get a residuals sequence for the length of the validation period. That residual series you examine for the major assumptions that you we have made in building the model, namely that the residuals have a zero mean. Therefore, we test for the significance of the residual mean. If the residual mean is significant then it fails the test; if the residual mean is not significant that means that it can be approximated to be 0 and therefore, it passes the test.

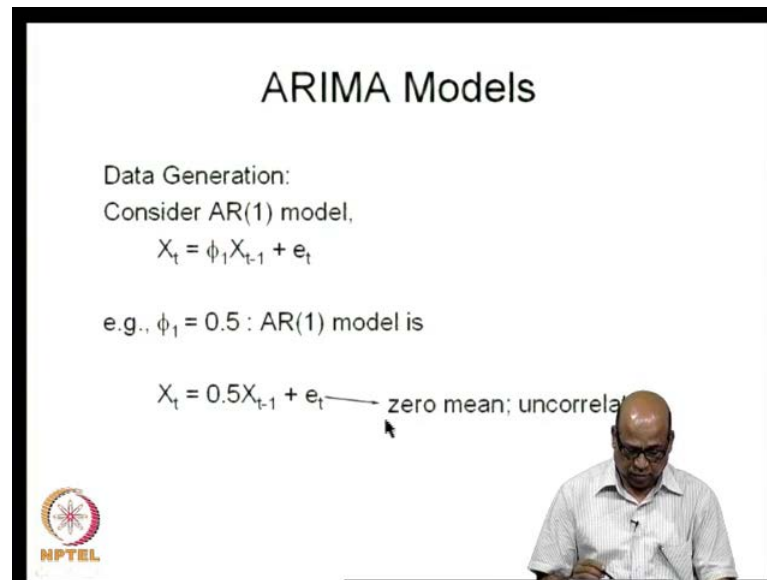
Similarly, we examined two tests for periodicities, one by defining a particular statistic where it examines the periodicities one at a time. So, you form the statistic and then compare it with the associated F value to examine whether that particular periodicity is significant or not. So like this one by one you test the periodicities.

We also discussed the cumulate periodogram test, which is also called as the Bartlett's test for significance of periodicities where we formulate the cumulative periodogram and then draw the confidence bands, 95 percent or 99 percent confidence bands, around the cumulative periodogram. If the cumulative periodogram resulting from the residual series lies within this band, significant band, then the series passes the test. In the sense, that the series does not have any significant periodicity present in the in the series.

The last series of test was on examining whether the series is in fact uncorrelated or it forms a white noise, for which we examine two tests we discuss two tests, one is Whittle's test and the Portmanteau test. Both of which define a particular statistic and this statistic, if you recall is dependent on the covariance matrix. What we will do in today's lecture is that we will see how a model is actually applied to several case studies, and how we carry out this test and so on. In the last lecture, I could not discuss any numerical examples, so today we will see what we mean by one time step forecasting?

How do we apply these models in the onetime step forecasting? How we select the models for long term synthetic generation of the data? Then we will also discuss primarily in today's lecture, I will discuss certain case studies of actual observed data how we apply that observed data for model building.


(Refer Slide Time: 05:35)



ARIMA Models

Data Generation:
Consider AR(1) model,
$$X_t = \phi_1 X_{t-1} + e_t$$

e.g., $\phi_1 = 0.5$: AR(1) model is
$$X_t = 0.5X_{t-1} + e_t$$
 — zero mean; uncorrela



The slide features a presenter in the bottom right corner, a mouse cursor pointing to the text 'zero mean; uncorrela', and the NPTEL logo in the bottom left corner.

Now, for the data generation, remember the data generation, we are talking about long sequences of data generated based on the models that we have built on the historical data. So, let us say that we have an AR 1 model that is auto regressive model of order one. We write this as X_t is equal to $\phi_1 X_{t-1}$ plus e_t .

So, let us take one simple example where I fix up ϕ_1 as 0.5. We will see how we apply this model. So, I will write this as X_t is equal to $0.5 X_{t-1}$ plus e_t . Now, e_t is a series, which has a zero mean and the series is uncorrelated. So, to apply this we choose the e_t such that they have zero mean and they are all uncorrelated. For example, if you take the standard normal deviates they all has zero mean and coming from a purely random sequence they are uncorrelated.

(Refer Slide Time: 07:01)

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

$$X_1 = 3.0 \text{ (assumed)}$$
$$X_2 = 0.5 \cdot 3.0 + 0.335$$
$$= 1.835$$
$$X_3 = 0.5 \cdot 1.835 + 1.226$$
$$\approx 2.14$$

And so on...

In the bottom left corner, there is a logo for NPTEL (National Programme on Technology Enhanced Learning). In the bottom right corner, a man with glasses and a light-colored shirt is visible, likely the presenter.

We choose the e_t sequence, following these two. Let say, how do we apply then. To start the generation process as we did in the **Thomas firing model case**, we start the generation process by assuming an initial value. Let say X_1 is equal to 3.0. Typically, this is assumed to be the mean of the series itself. Then X_2 , I write as $0.5 X_{t-1}$ which is X_1 in this case plus the e_t term.

Now, these e_t terms I may be choosing from the standard normal deviates; so 0.335. So, I get X_2 as 1.835. Then I will use this 1.835 in the generation of the next value, ϕ remains the same, X_2 now when I am generating X_3 , and then I put another e_t value from the same sequence of random numbers. I get 2.14 and this 2.14 I use to generate X_4 and so on. So, this is how we use the AR 1 model to generate numbers, to generate the particular variable values. X_1 we assumed and X_2 X_3 etcetera these are generated.


(Refer Slide Time: 08:22)


ARIMA Models

Consider ARMA (1, 1) model,
$$X_t = \phi_1 X_{t-1} + \theta_1 e_{t-1} + e_t$$

e.g., $\phi_1 = 0.5$, $\theta_1 = 0.4$: ARMA(1, 1) model is written as

$$X_t = 0.5X_{t-1} + 0.4e_{t-1} + e_t$$






Let us take an ARIMA models that is auto regressive moving average model of order (1,1), which means it has one AR parameter and one MA parameter. So, X_t is equal to $\phi_1 X_{t-1}$; this is AR parameter plus $\theta_1 e_{t-1}$ plus e_t . This is the ARIMA 1 1 model. Let say ϕ_1 is equal to 0.5 and θ_1 is equal to 0.4. This is now written as X_t is equal to $0.5 X_{t-1}$ plus $0.4 e_{t-1}$ plus e_t . when we are doing the model testing, let say this is our model that we have built, when we are doing the model testing we use this e_{t-1} as the e_t that is used in the previous term, that is for $t-1$. I will explain what I mean by that.


(Refer Slide Time: 09:20)

ARIMA Models

Say $X_1 = 3.0$

$$X_2 = 0.5 \cdot 3.0 + 0.4 \cdot 0 + 0.667 = 2.167$$
$$X_3 = 0.5 \cdot 2.167 + 0.4 \cdot 0.667 + 1.04 = 2.39$$
$$X_4 = 0.5 \cdot 2.39 + 0.4 \cdot 1.04 + 2.156 = 3.767$$





Let say we start with the same value X_1 is equal to 3.0. so, I get X_2 is equal to ϕ_1 into X_{t-1} , which is ϕ_1 into X_1 plus θ_1 into e_{t-1} to begin with we assume there is no e_{t-1} here. Therefore, it is zero plus e_t . Now, this is an e_t term that I am using, choosing from a set of numbers which have zero mean and they are all uncorrelated. So, this is 0.667, I get 2.167.

When I go to X_3 to generate X_3 , I write ϕ_1 into X_{t-1} which is X_2 which is this 2.167 plus θ_1 0.4 into e_{t-1} . So, e_{t-1} is the random term that we have used in generating the previous values, so that becomes 0.667 plus e_t . Now, e_t is a new number drawn from the same sequence of random numbers. Like this it keeps on going, so you generate several values in the series like X_1, X_2, X_3, X_4 and so on. So, this is how we do the long term synthetic generation of the data using the ARIMA models. Remember here, I have talked about AR and MA, I is the order of differencing which we do before we build the ARIMA model.

(Refer Slide Time: 10:58)

ARIMA Models


Data Forecasting:
Consider AR(1) model,
$$X_t = \phi_1 X_{t-1} + e_t$$

Expected value is considered.

$$E[X_t] = \phi_1 E[X_{t-1}] + E[e_t]$$

$$\hat{X}_t = \phi_1 X_{t-1}$$

Expected value zero



Now, let say we want to use the ARIMA model that we have chosen for one time step ahead forecasting. That is we chose the model from amongst set of candidate models by computing the mean square error of those models and then finally, we chose this particular model.

Now, as mentioned in the last lecture, what do we mean by forecasting? One time step ahead forecasting indicates that we are in a particular time period and then we want to

forecast what is likely to be the value of that particular variable for the next time period. Let say at the end of the June month, we know the data that has actually occurred for the June month. Now, we want to forecast what is likely if you are talking about the stream flows, what the likely stream flow is during the next time period which is July, if we are talking about the monthly time series. So that is a problem here.

Now, the when we are talking about the forecast, we are just looking at the expected value of that particular variable during the next time period from arising from this particular model. So, if we take the expected value of X_t ϕ_1 e expected value of X_{t-1} plus expected value of e_t , now expected value of e_t is the mean of e_t which is 0 because that is how we have used in the model. And therefore, I can write the forecast, which is typically written as X_t cap for the variable X_t is equal to $\phi_1 X_{t-1}$.

(Refer Slide Time: 12:50)

ARIMA Models

Consider ARMA(1, 1) model,


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

$$E[X_t] = \phi_1 X_{t-1} + \theta_1 e_{t-1} + 0$$

Error in forecast in the previous period

e.g., $\phi_1 = 0.5$, $\theta_1 = 0.4$: Forecast model is written as

$$X_t = 0.5X_{t-1} + 0.4e_{t-1}$$



Now, if we want to do the same thing for ARIMA 1 1 model, X_t is equal to $\phi_1 X_{t-1}$ plus $\theta_1 e_{t-1}$ plus e_t . Remember, this e_{t-1} is coming from the model application for the time period $t-1$, which is a known value; e_{t-1} is a known value, which was responsible for generating the previous value.

So, expected value of X_t is equal to $\phi_1 X_{t-1}$ plus $\theta_1 e_{t-1}$ plus 0. This sequence has a zero mean. Therefore, I write it as 0. So, this is an error in forecast in the previous time period. So, you had applied this in the previous time period compared it with the actual data for the validation period and then you obtain this e_t

minus 1. Therefore, that is a known value and when you take the expected value of a constant if you recall it will be the constant itself. Let say for example, phi 1 is equal to 0.5 and theta 1 is equal to 0.4 as we considered earlier. We write this as X_t is equal to $0.5 X_{t-1} + 0.4 e_{t-1}$.

(Refer Slide Time: 14:08)

ARIMA Models

Say $X_1 = 3.0$

Initial error assumed to be zero

$$\hat{X}_2 = 0.5 \times 3.0 + 0.4 \times 0$$

$$= 1.5$$

Forecasted

$X_2 = 2.8$ ← *Observed Value*

Error $e_2 = 2.8 - 1.5 = 1.3$

X_2 at $t=2$ and X_3 at $t=3$

$$\hat{X}_3 = 0.5 \times 2.8 + 0.4 \times 1.3$$

$$= 1.92$$

Actual value to be used

For the forecasting when we are applying it for the validation period, let us say X_1 is equal to 3.0 and I want to forecast the X_2 value. I write it as X_2 cap using that value model, which is phi 1 which is 0.5 into X_1 as 3.0 plus theta as 1.4 and e_{t-1} which is 0. So that will be equal to 1.5. Now, look at what we do for the second value. X_2 is 2.8, this is an observed value, this is not the forecasted value, this is observed or it is the data and this is the forecasted value. so, we compute the error e_2 as observed minus the forecasted, which is the 1.3 and this is what goes into generating the next value, so in forecasting the next value.

So X_3 cap, I write again as phi 1 X_2 plus theta 1 e_2 and e_2 is 1.3 here. Now, when we are doing X_2 , remember you take the observed value so you take the actual observed value for forecasting the next one. So, essentially as I mentioned in the last lecture essentially what we are doing is in this particular case, we are standing in time period t is equal to 2, and then forecasting for t is equal to 3. Now, this is X_t that is X_3 cap and X_2 is known, so we obtain the error based on the forecast. The time period to use that error in generating the next value for X_3 , next forecast for X_3 that is 1.3 this is what we are

using this 2.8 is the observed value like this we keep going. Now, X_3 again, we compute the error based on the observed value of X_3 and then generate X_4 and so on.

(Refer Slide Time: 16:36)

ARIMA Models

$X_3 = 1.8$
 Error $e_3 = 1.8 - 1.92 = -0.12$

$\hat{X}_4 = 0.5 \times 1.8 + 0.4 \times (-0.12)$
 $= 0.852$

and so on.

So, X_3 is 1.8. This is again the observed value. X_3 cap is the forecasted value, so we get the error as minus 0.12, this minus 0.12 is what we use in the forecasting of the next value and so on. So, this is how we obtain the forecast based on the ARIMA models. What we will do is we know how to generate the values and we know how to forecast the values; one time step ahead forecasting. I keep on mentioning that one time step ahead forecasting models but, you can also have two time step ahead forecasting models two time step ahead forecasting models and so on.

The difference there would be in one time step ahead forecasting models what did we do we forecasted X_t based on X_{t-1} , which means standing at the end of the time period $t-1$, we are forecasting X_t . Let say we want to use two values, two previous values, typically in hydrology, we do not deal with such situation, but any many of the other areas of engineering, typically they use several of these time lag values for forecasting.

Of course, when I say that in monthly time series forecasting and so on, we may consider previous lags; X_{t-3} X_{t-4} and so on. When I discuss the case studies I will make that more clear. So, in two time step ahead forecasting what we do is standing at X_t ; that is we want to forecast X_{t+2} based not only on X_{t+1} , but also on X_t .

So, we want to forecast in two time step ahead forecasting. I am sorry I will explain that correctly. One time step ahead forecasting we did from t to $t + 1$ let us say.

So, we obtain up to t , it is known and you are forecasting here, for one time step ahead. So, this known data we use in the two time step ahead forecasting. Up to this point, the whole history is known. We use all of that to forecast not only $t + 1$, but also $t + 2$. So, this is two times step ahead forecasting. Let us say that you are talking about forecasting of a flow increase in a particular river. So, standing in particular time period, let us say one hour you are standing a particular hour.

Then you would like to say that the next hour, the level will be so much, the hour next to that the level will be so much. In the case of flows standing in June month, you may want to say what the forecast is for July and what the forecast is for August, based on these models. Now, when we are applying it for two times step ahead forecast, the assumption is that you have the observed values only up to this time period t .

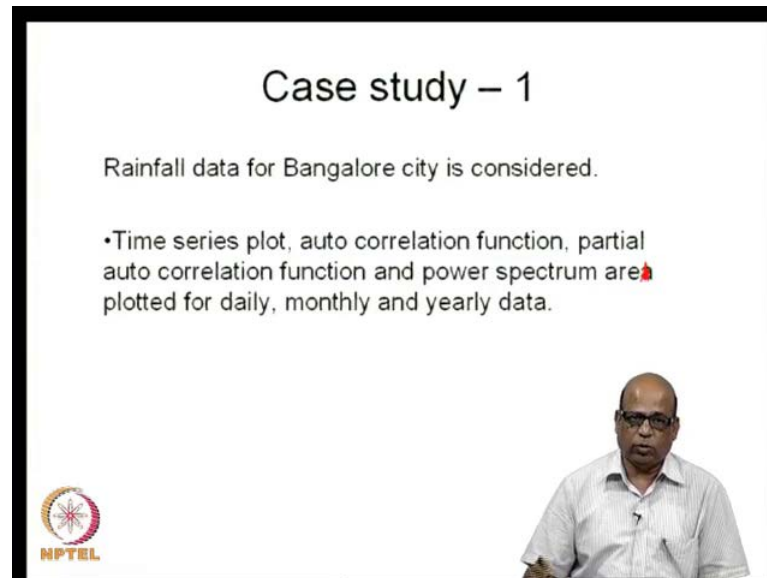
So, using this forecast you have forecasted X_{t+1} . Now, when we are forecasting for X_{t+2} , you use the forecasted value for X_{t+1} , namely the X_{t+1} cap that you use, because you do not have when you are doing two time step ahead forecasting. Both these two time step ahead actual values will be available only at the end of this two time steps. So, when you are doing two times ahead forecasting in obtaining your errors here, you will use the forecasted values, because you are doing two time step ahead, the values will be known only when you reach the end of time period $t + 2$.

Anyway, this we will discuss further when we look at some applications. Now, what we will do is that the mean time series analysis procedure that we have studied so far, namely the correlogram; correlogram gives us certain information, the partial auto correlation function gives us certain information, the power spectrum gives us certain information. All of these available information from the historical data, we synthesize and build the models. Once we build the models, we validate the models, based on the residual series. Now, these several steps, we will apply to several case studies, several different case studies.

In certain case studies, I may only discuss correlogram and spectral analysis and so on, in certain case studies we will go all the way up to building the ARMA model, and see how they can be applied for actual data generation as well as forecasting. Now, you must

remember, I have kept on telling in the course that the daily rainfall data, behaves much differently from let say a much normalized, much smoothen process, like a monthly stream flow or seasonal stream flow and so on.

(Refer Slide Time: 22:24)



The slide is titled "Case study - 1" and contains the following text:

Rainfall data for Bangalore city is considered.

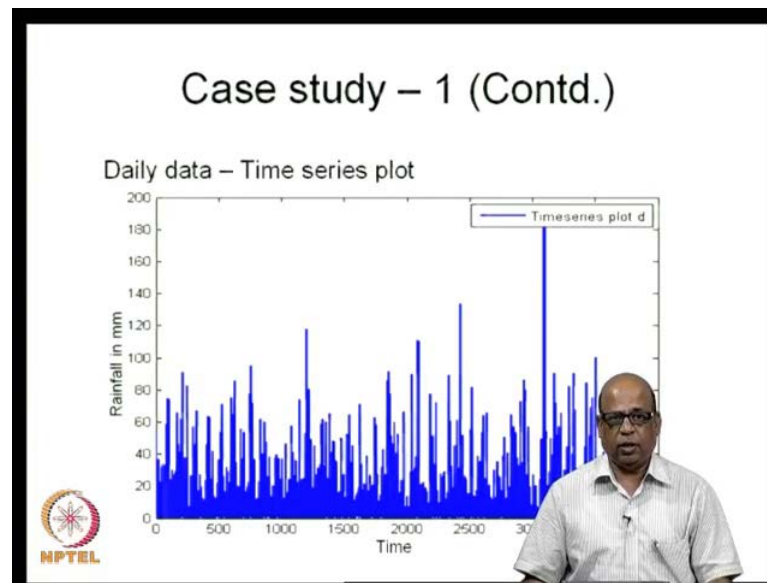
- Time series plot, auto correlation function, partial auto correlation function and power spectrum are plotted for daily, monthly and yearly data.

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

Let us look at first the daily rainfall data and we consider this daily rainfall data for Bangalore city, for certain duration and we plot for this daily data, we plot the time series, auto correlation function, serial partial auto correlation function and power spectrum.

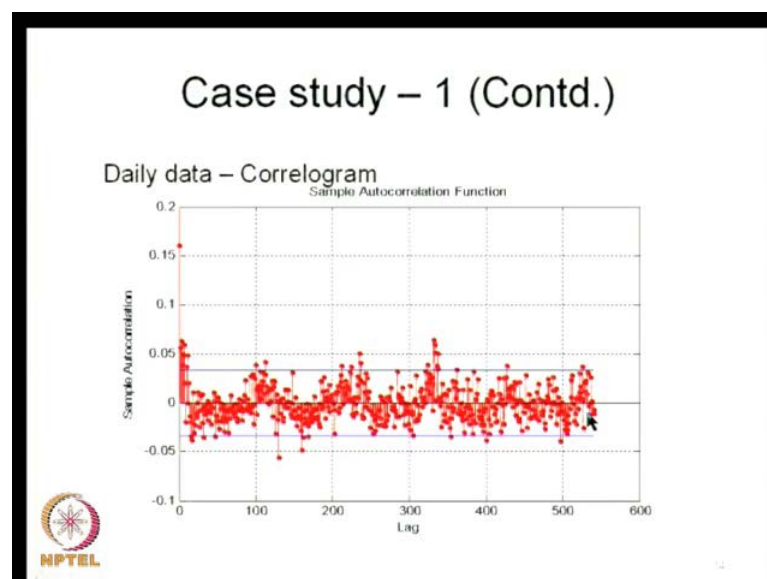
Then what we will do is we will add the daily data to obtain the monthly data and do the same analysis and see how they appear differently. They appear different from each other. Then add all the monthly data and obtain the yearly data and look at how the yearly data looks compared to the daily data. What we may expect in an urban area? The rainfall, daily rainfall, is in some sense it indicates a random process. I mean, the number the rainfall data may be uncorrelated and they you may not get any significant periodicities if you are looking at short duration rainfall such as daily rainfall. In fact these short duration rainfalls are what cause the urban flooding. Therefore, it is important to understand how the rainfall process itself is behaving.

(Refer Slide Time: 23:49)



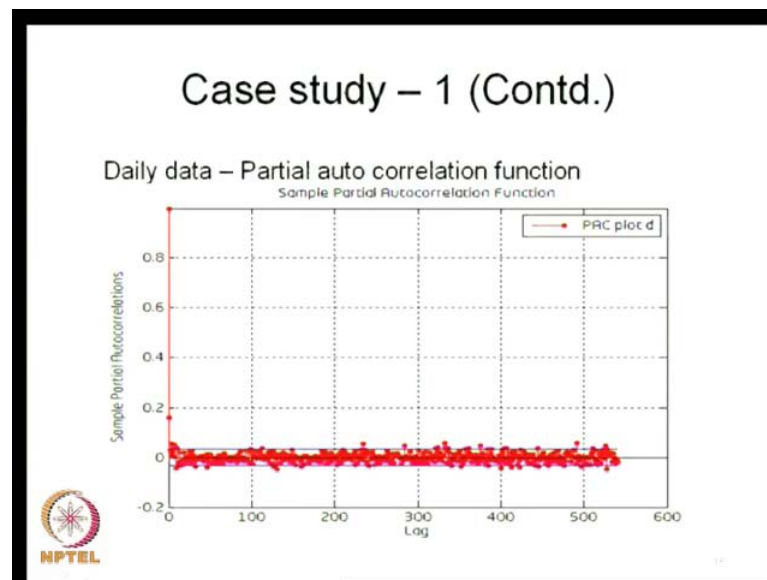
So, let us look at the time series. Certain length of the data, I have taken and then I have plotted the time series. This is how the time series looks like. There are large numbers of zeros here and then from the time series you cannot make out much in the daily case. If you go to monthly or yearly cases, we may perhaps suspect that there is some periodicity, but when we are talking about rainfall and you are plotting for short time duration, such as day, it is very difficult to understand whether there are any significant period which is present in the data and so on.

(Refer Slide Time: 24:27)



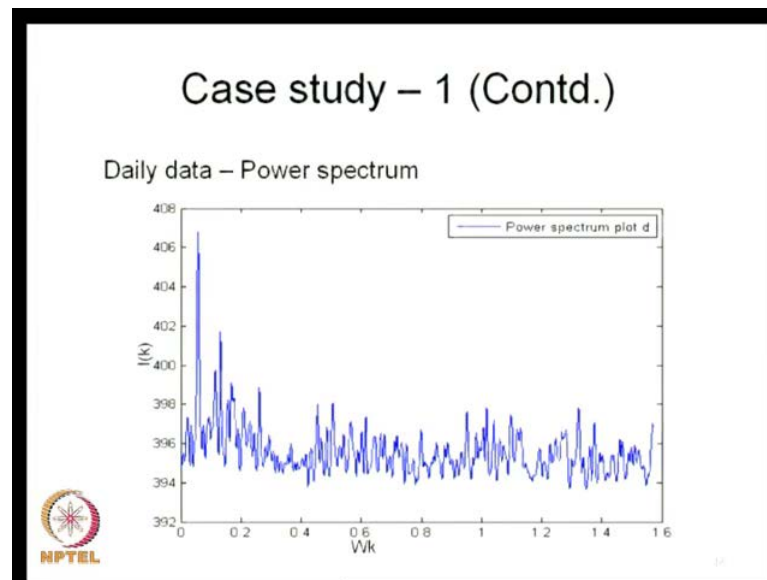
So, we will plot the correlogram. Now, when I plot the correlogram, as you can see these are the significant bands here that is 95 percent significant bands. Most of the correlations, with respect to most of the lags that is correlations with respect to most of the lags, are within the significant bands, indicating that they are all insignificant. There are hardly few initial lag correlation significant and there may be some of these, which are significant at perhaps regular intervals. We will have to check this with the spectral analysis. So, this indicates that the daily rainfall data is likely to be uncorrelated. Although, there are certain significant periodicities, for example, we are not getting a peak like this 0.15, 0.2, and so on, except for lag 0, which is lag one perhaps near to 0, which is close to 0.15. This value is not exactly for lag 0, lag 0 correlation is always one.

(Refer Slide Time: 25:42)



Now, we will see what happens when you look at the partial auto correlation. Again, partial auto correlation, this is a significant band here. Remember, I mentioned during discussion on auto correlation that being a auto correlation its significance band will also be same as the significance bands for the correlation, which is in this case we take it as plus minus 2 by root n; n being the data n being the length of the data I am sorry n being the number of data points. So, for most of the partial auto correlations are insignificant except for the very first few, may be first one, second one and so on. Then we will look at the power spectrum.

(Refer Slide Time: 26:34)

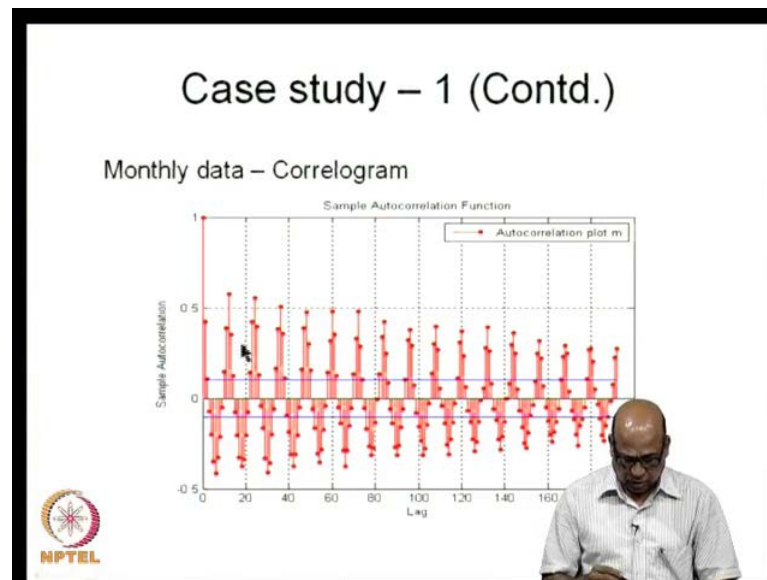


Power spectrum, if you had regular periodicities, what you would have seen that most of these spectral values will all be in certain band and suddenly you will see a large spike. But, that you are not seeing it. In the particular case, it indicates that there is an important contribution to variance somewhere around this point or may be around point just less than 0.1 and so on.

But, by enlarge, it is spread. There is no significant value around, which there is a contribution to variance. So, the daily data, typically behave like this, daily rainfall data typically behave like this where you are unable to say much on what kind of regularity exist in the data or what kind of deterministic component you can extract form the data and so on. Obviously, because this is all typically you find uncorrelated values of rainfall. When you are talking about smaller durations of daily rainfall or 6 hourly rainfall etcetera, which are important in the flooding situation, it is extremely difficult to get a good deterministic component. They behave mostly like a purely random process.

Let us now add up all the daily values and generate a monthly time series. In this case, we have about 32 years of data and this is the monthly time series. Now, the monthly time series appears to tell us that you may suspect certain periodicities. They seems to be certain regularities here, in terms of fluctuations and this one becomes more clear, when this information that we are suspecting the time series to hold, becomes much more prominent when we do the correlation or the correlogram analysis.

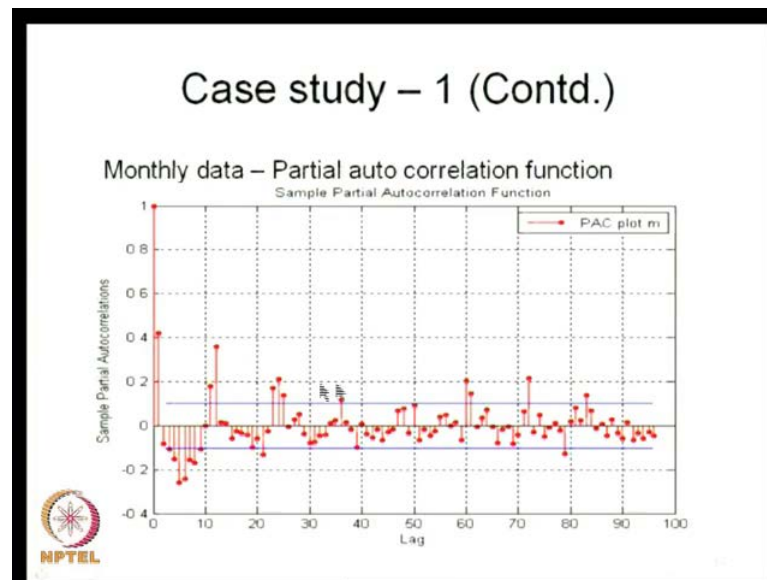
(Refer Slide Time: 28:48)



So, in the correlogram you see a sinusoidal variation. Remember, between this and this, notice that between this and this, there are 20 values, so it is going like this. There is a sinusoidal variation like this, it keeps going on like this and then there is a slow decay as you progress into time. We are going about 195 or something, some lag, there is a slow decay of the correlation. These bounds are again the significance bounds, 95 percent significance bounds and most of the correlations are significant as you can see.

Now, this typically indicates that there is a periodicity present in the data. For example, this may corresponds to 6, this may corresponds to 12 and so on. We will what exactly it means when we look at the spectral analysis the correlogram or the monthly data definitely indicates a presence of periodicity. The correlogram is dying down in a sinusoidal form. Compare this with what we did for the daily. So, this was the daily data correlogram where most of correlations were insignificant, whereas, when you come to the monthly data, most of the correlations are significant and then there is also a sinusoidal variation, indicating that there is a periodicity present in the data.

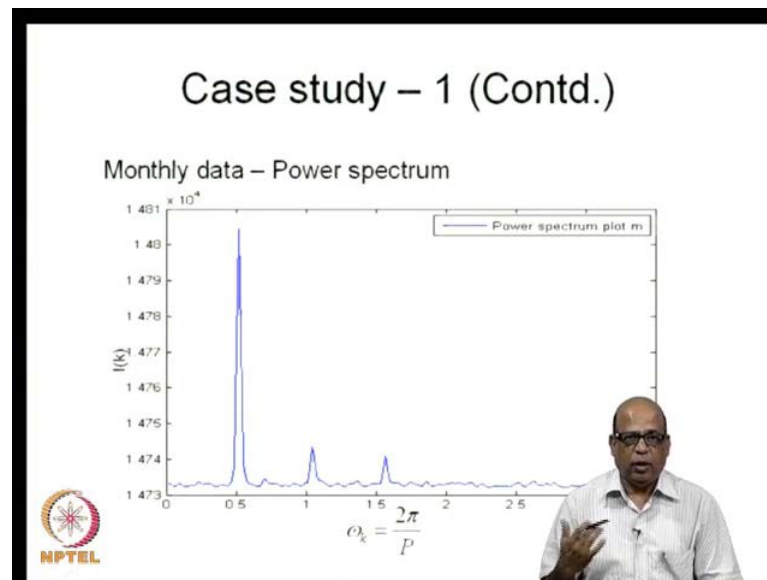
(Refer Slide Time: 30:26)



Now, we will see how the partial auto correlations show up? The partial auto correlations, there are many significant partial auto correlations as you can see, the corresponding to one perhaps, there is a significant partial auto correlations. These are significant; these are significant and so on. As you go further, however most of the partial auto correlations become insignificant. This is for the monthly data. Now, what does this tell in terms of our model building that we had discussed earlier that if your auto correlogram or the correlogram indicates that that is a sinusoidal or exponential decay, and there are certain partial auto correlations, which are significant, this indicates a MA model, MA process, in which you can use these significant partial auto correlations to build the MA model.

However, we are not adapting that procedure of identification as I have mentioned. We simply build candidate models based on this information and then look at look at the particular models to be chosen. **I am sorry**, I just want to correct that when you have the auto correlation decreasing, decaying, slowly with respect to time, and there are certain partial auto correlations significant, it indicates an AR process not the MA process. It indicates an AR process and then the order of the AR, auto regression, we choose based on the number of partial auto correlation which is significant.

(Refer Slide Time: 32:14)



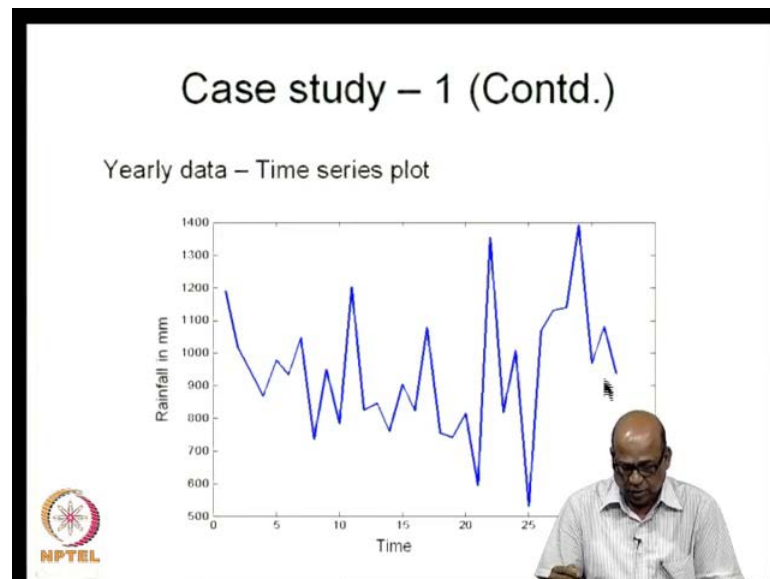
Now, we look at for the monthly data, we look at the power spectrum. see this indicated the correlations, correlogram indicated that there is a certain periodicities and we also suspected that there may be a periodicity corresponding to 6 months; there may be periodicity corresponding 12 months and so on. So, this information is strengthened, this indication is strengthened by the power spectrum. When I plot the power spectrum, I get a periodicity corresponding to somewhere around 0.52 or something, then 1.23 and so on.

So, this corresponds to, if we use ω_k is equal to 2π by p and associated ω_k value if we put, this corresponds to a periodicity of 12 months, this 6 months, 4 months and so on. So, in the monthly data, for Bangalore city, we observed that there is a periodicity of 12 months, which is in monsoon situation it is quite common, where there is a regularity in the rainfall pattern. But, these are not so intuitively clear that why there should be a 4 month periodicity, why there should be a 6 month periodicity and so on. As I mentioned earlier, while the spectral analysis brings out the fact that there are certain periodicities present in the data whether these periodicities are in fact significant or not, we need to test separately.

So, this information only indicates that there is a 12 month periodicity, 6 month periodicity and 4 month periodicity here. How do we obtain that is simply by p is equal to 2π by ω_k and the ω_k is associated with that particular spike. You pick up the ω_k

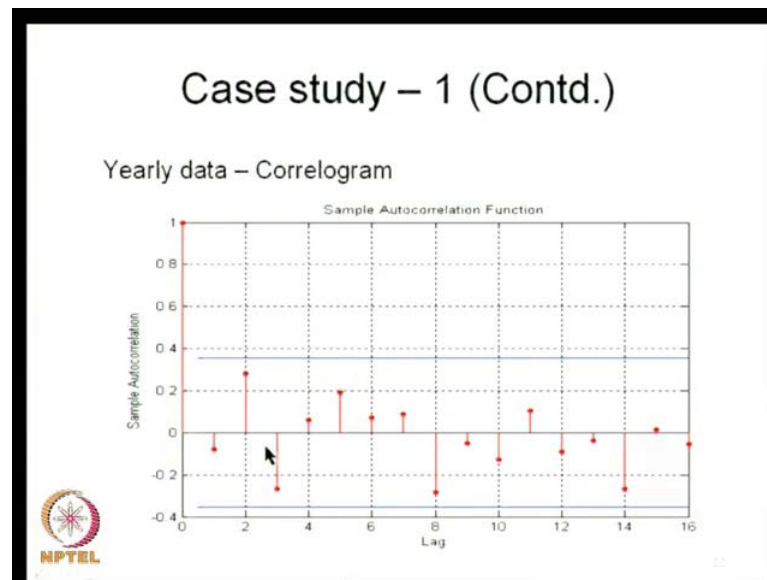
associated with this and then get the periodicity. Now, we will go one step ahead. Now, we started with the daily data. We did all this analysis; we could not get much of the information out of that. Then we added all the daily data obtain the monthly data and did the analysis for the monthly data, the monthly data correlogram showed up periodicities present in the data and that was fortified by the power spectrum. Now, we will add all the monthly data for the 12 periods in the year, 12 months in the years, and then formulate a yearly time series and plot the time series.

(Refer Slide Time: 34:47)



So, yearly time series if you draw, this is between 0 to about 1400 millimeters; annual rainfall, in Bangalore city. If you plot that this looks slightly haphazard like this, so the time series for last about 32 years or something, we have taken here and this is how it varies.

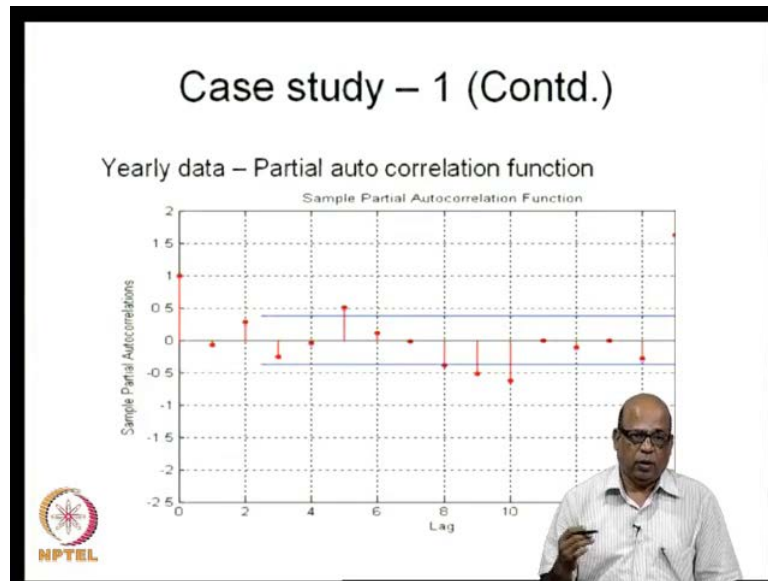
(Refer Slide Time: 35:15)



Let us see what happens to the correlogram of this; correlogram of the annual data. Now, the significance bands are here now. Why the significance bands are different from monthly to daily to annual? It is because of the number of values that you are using. This is $2 \sqrt{n}$ and n in the case of monthly time period is let say it is you have annual data and one year consist of 12 months and therefore, the number of months will be much larger than the number of years, and therefore, the number of values that you use to compute this for the annual data, will be much smaller than the number of values that you used for monthly data.

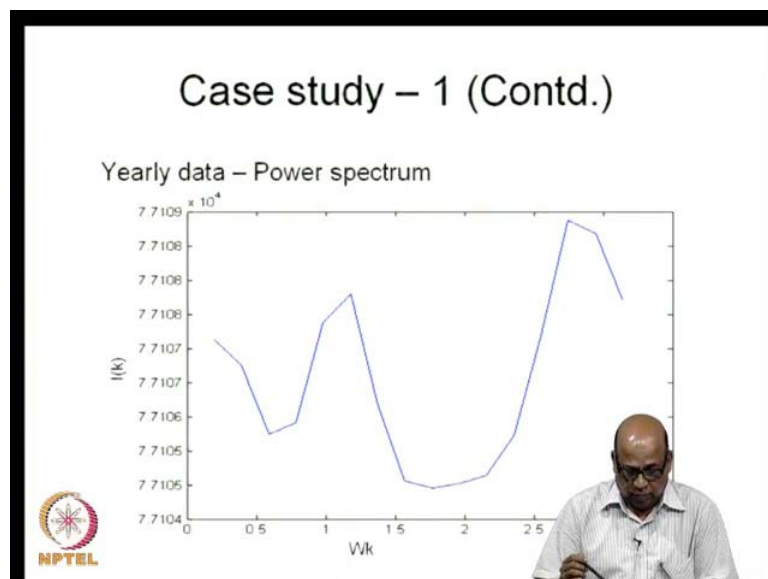
And this is $2 \sqrt{n}$ and therefore, as n decreases, this will be larger and larger. So, you are getting larger and larger insignificance bands, which mean that if the correlation, lie within these bands. Then they are all insignificant. So, as you can see, all the correlation for the annual rainfall data, they are all insignificant; statistically insignificant. We leave out the lag 0 correlations because that is always one, so all other remaining correlations are all insignificant.

(Refer Slide Time: 36:52)



Let us look at what the partial auto correlations say? Partial auto correlations also bring up the same conclusion that all the partial auto correlations, most of the auto, except this and this here somewhere around lag 10 and lag 9, they are significant, and then lag 5 is significant, zero you leave out and this is what the partial auto correlation function brings up.

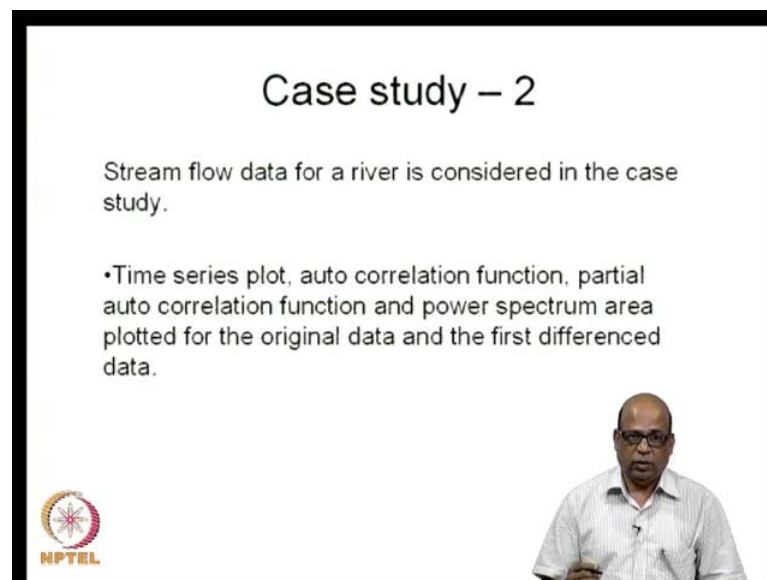
(Refer Slide Time: 37:24)



So, as we progress from daily to monthly to annual, the behavior of the process changes and this is important in hydrology, because what you are applying the models for; is it daily forecasting that you are looking for or is it monthly time forecasting and so on.

So, depending on that, the information contained in the time series is much different. The power spectrum; look at the power spectrum, now compare this with what we had; this is for the annual data, compare this with what we had for the monthly data. Monthly data showed definite periodicities present in the data and the daily data showed that power spectrum more or less behave like it is for a white noise. Now, here there are some lower frequencies and some higher frequencies that are dominant, so you cannot say much about the time series from the power spectrum data alone.

(Refer Slide Time: 38:32)



The slide is titled "Case study - 2" and contains the following text:

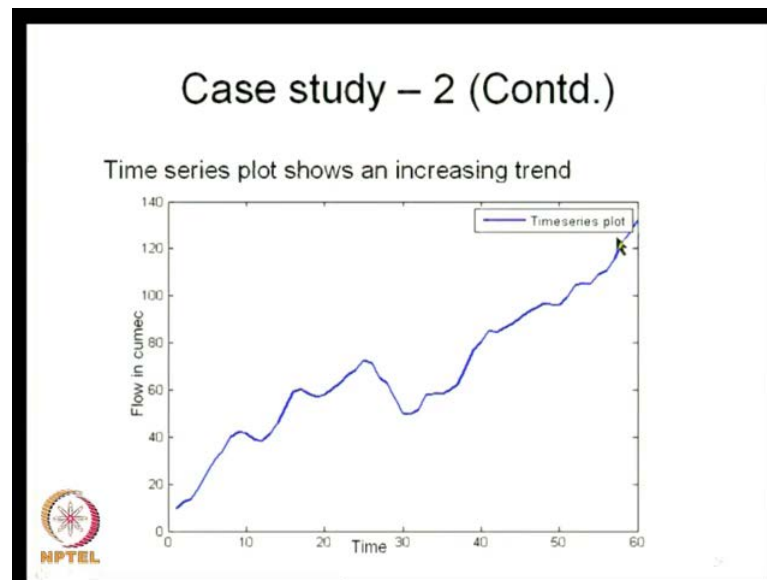
Stream flow data for a river is considered in the case study.

- Time series plot, auto correlation function, partial auto correlation function and power spectrum area plotted for the original data and the first differenced data.

In the bottom right corner of the slide, there is a small video inset showing a man with glasses and a white shirt. In the bottom left corner, there is the NPTEL logo.

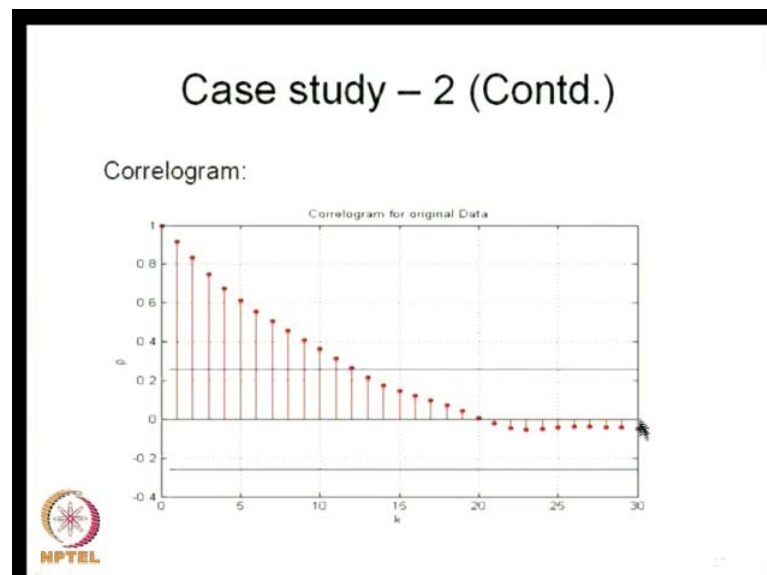
Now, we will go to a much smoothen process of stream flow data and we will consider the annual stream flow data. There may be certain trends in the annual stream flow data. Let us say, the stream flow data may be increasing or decreasing slowly and so on. This information on the trend, existing trend, is important for water resource managers, to use for planning for the next 10 years 15 years and so on. But, apart from the trend itself, we would like to build models for synthetic generation of data as well as for forecasting of data.

(Refer Slide Time: 39:18)



So, let us look at annual stream flow data. So, this is for 60 years at a particular location. There is a certain trend on an average, these are all increasing, so there is an increasing trend over the last about 60 years. These are in cumec, so this is a discharge value with respect to time in years.

(Refer Slide Time: 39:42)

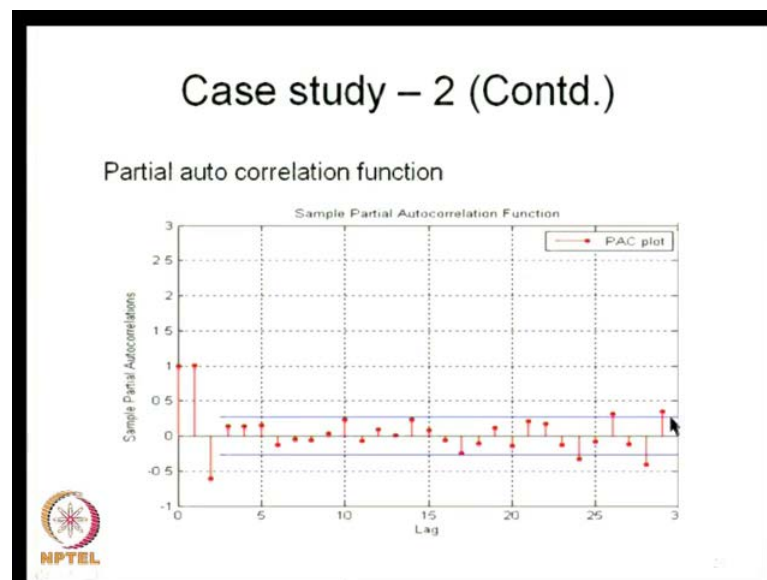


Now, we will plot the correlogram for this. I have gone up to about 30, which is actually 50 percent of the data available, but typically we go up to about 15 percent of the data available. So, if you plot the correlogram, the correlogram indicates an exponential decay

here. As you are progressing with respect to time, there is an exponential decay on the correlogram. this is the stream flow data.

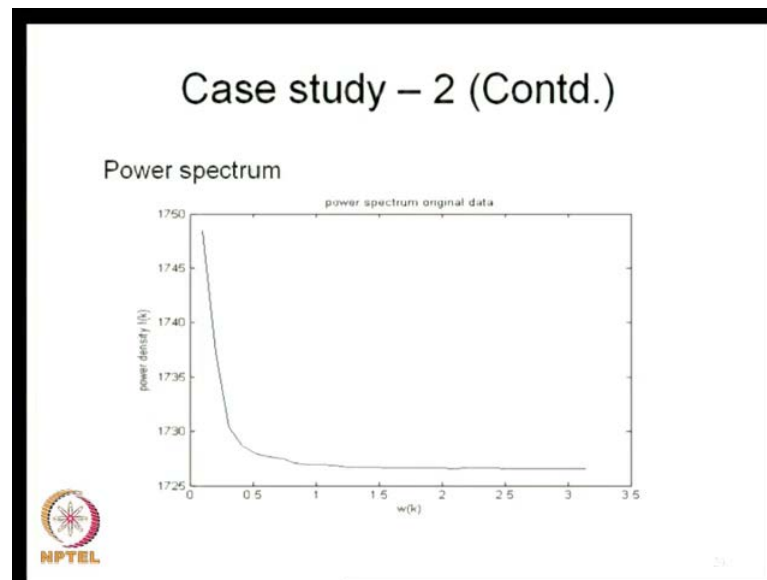
Then we look at the partial auto correlation. So, we plotted the time series, we plotted the correlogram; correlogram indicates that there is an exponential decay in the correlogram. Now, we look at what is a partial auto correlation, what does a partial auto correlation say? If you see that there is decay in the correlogram and there are certain partial auto correlations significant, you can suspect it to be an AR process auto regressive process.

(Refer Slide Time: 40:39)



So, let us look at the partial auto correlation. At lag 2, the partial auto correlation is one, nearly one here and then at lag three also it is significant, and then there are no significant partial auto correlations. This is just coming up out. These two are just above the significance bands, but typically you can say that one and two, lag one and two are significant, so you may suspect this to be an AR 2 process.

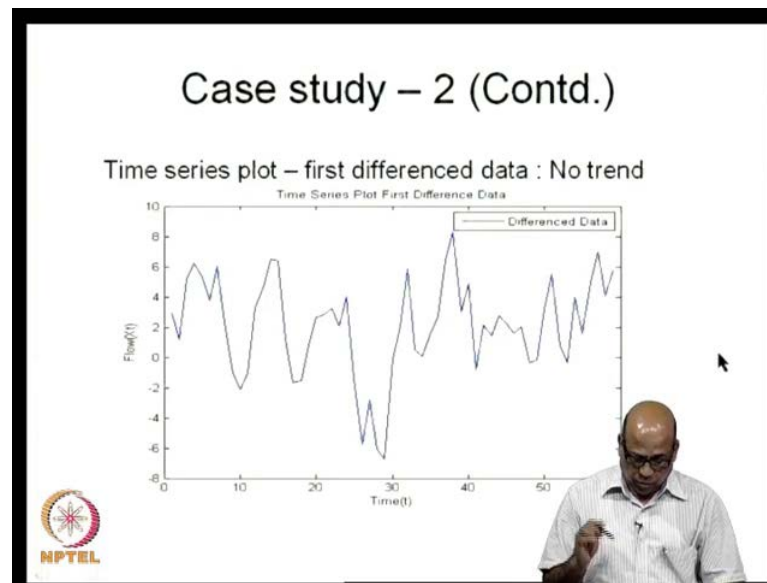
(Refer Slide Time: 41:19)



Now, we look at the power spectrum, the power spectrum indicates for this particular case study that the lower frequency dominant. When I discussed examples of theoretical AR models, typically this is what we brought out that the power spectrum for one AR two model; you may have lower frequencies dominating, for some other AR 2 model you may have middle level frequencies dominating and so on. But, typically it is a correlogram and the partial auto correlation function, these will indicate or this will give good indications for the auto regressive process.

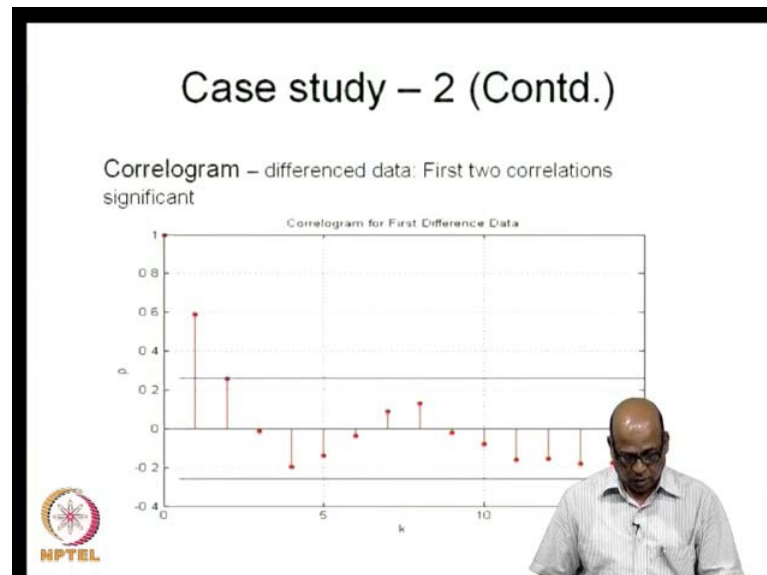
In this particular case, we can say that this follows closely the AR two processes. Now, what we do is because we saw a trend here, the time series had a trend. We want to remove the trend. So, what is the procedure? In the ARMA model procedure that we adopt to remove the deterministic components, we do the differencing first. So, X_t minus X_{t-1} that is a first differencing. So, let us see on this if we do the differencing, we get the time series for the first difference data that is first order difference data.

(Refer Slide Time: 42:43)



As you can see here what was appearing like this in the original time series, now when we do the differencing, it appears something like this, which means that the trend that or so apparent in the time series, has been removed now. We will do the correlation, correlogram analysis on this time series.

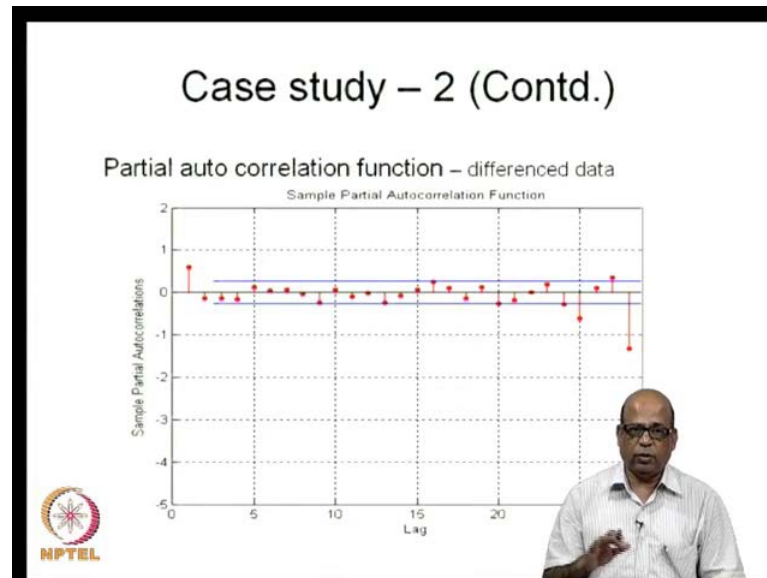
(Refer Slide Time: 43:15)



Now, when we do that, the correlogram for the first difference data, you can see that there is only one correlation. The lag one correlation that is significant and unlike what you saw in the earlier case where there is several correlations significant. (Refer Slide

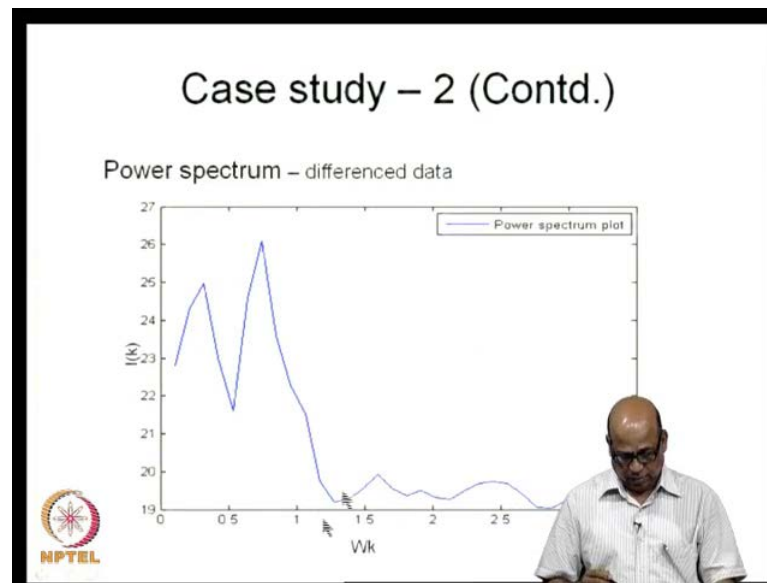
Time: 43:31) These were as you can see almost up to lag of about 12, they are all significant, and although they are decreasing with lag they are all significant, whereas for the difference data you see only lag one correlation coming up as significant.

(Refer Slide Time: 43:48)



The partial auto correlations again indicate that most of the partial auto correlations are insignificant. There is one at a very high lag that you can ignore when you are building a model and there is one at the beginning of the series, which is may be around lag two or something, it is significant. So, by differencing essentially you change the nature of the information contained in the data. One you remove the trend, because you remove the trend the correlogram behaved in different fashion, the partial auto correlation function behaved in a different fashion.

(Refer Slide Time: 44:21)

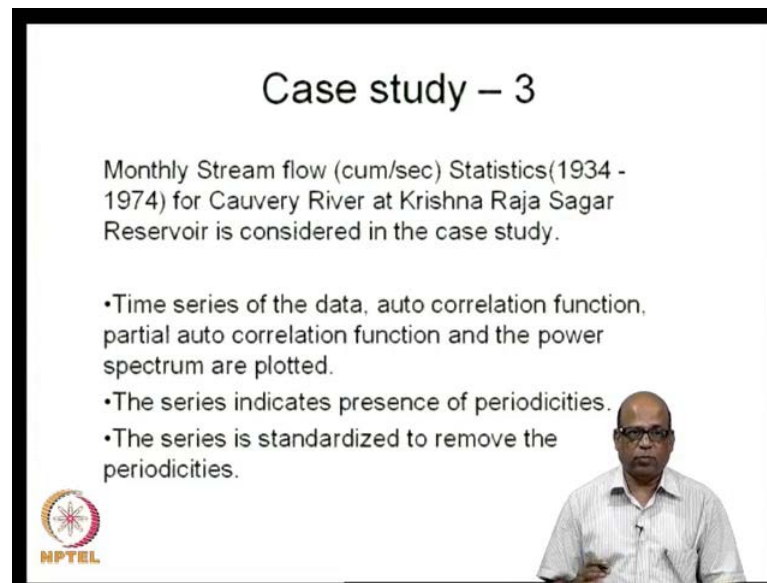


And now look at the power spectrum. The power spectrum for the difference data; it indicates unlike your power spectrum of your original data, which had significant contribution to various variance in the lower frequency. This is slightly more distributed now. So, you have frequency, you have contribution to variance almost up to about 1.2 or something, W_k up to about 1.2.

Now, what we will do is we will look at some case studies on ARMA models. So, both the case studies that I just discussed were with respect to only getting the information contained from the time series plot, from the correlogram plot, partial auto correlation function plot and the power spectrum plot. So, we just visually look at, we prepare all these plots and then visually look at this and then try to get the best information that is possible from these plots.

Now, we will do the actual model building exercise where starting with the observed data, we select certain candidate models, and then obtain the parameters, obtain the residual series that is for the calibration period we obtain the parameters. Then for the validation period, we obtain the residual sequence and then we do the test for the residual series to examine or to ascertain, whether the model that we choose either based on the maximum likelihood criterion or the minimum mean square error criterion, in fact passes the test all the thetas that I indicated earlier.

(Refer Slide Time: 46:26)



The slide is titled "Case study - 3" and contains the following text:

Monthly Stream flow (cum/sec) Statistics(1934 - 1974) for Cauvery River at Krishna Raja Sagar Reservoir is considered in the case study.

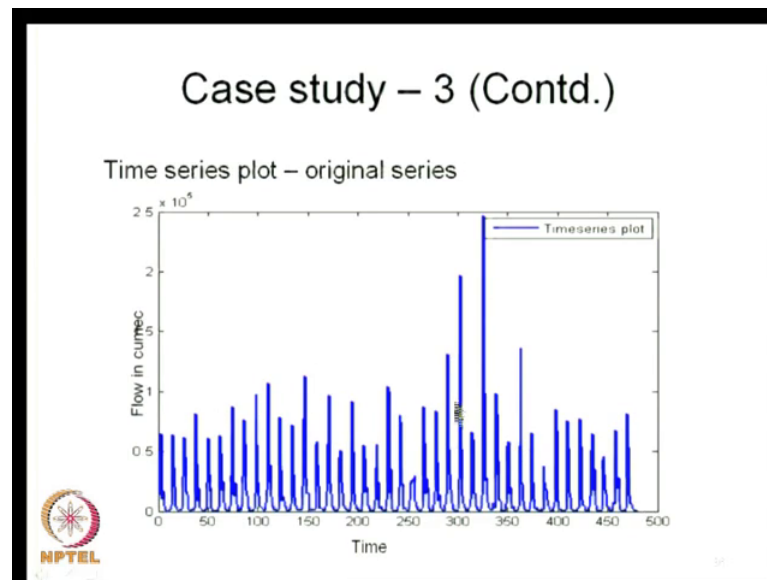
- Time series of the data, auto correlation function, partial auto correlation function and the power spectrum are plotted.
- The series indicates presence of periodicities.
- The series is standardized to remove the periodicities.

The slide also features the NPTEL logo in the bottom left corner and a small video inset of a speaker in the bottom right corner.

We will take the monthly stream flow data for Cauvery River at Krishna Raja Sagar, in Mysore. So, this is monthly stream flow data, this is cubic meters per second. The data is available for 1934 to 1974; we will plot the time series. We will plot the auto correlation function, partial auto correlation function and the power spectrum. Then we see that if there is periodicities present in the data, because we are talking about monthly stream flow data, as I have been saying repeatedly that the monthly stream flow data are almost certain to show up monthly, periodicity. I am sorry, periodicities corresponding to 12 months, especially in the monsoon regions like ours.

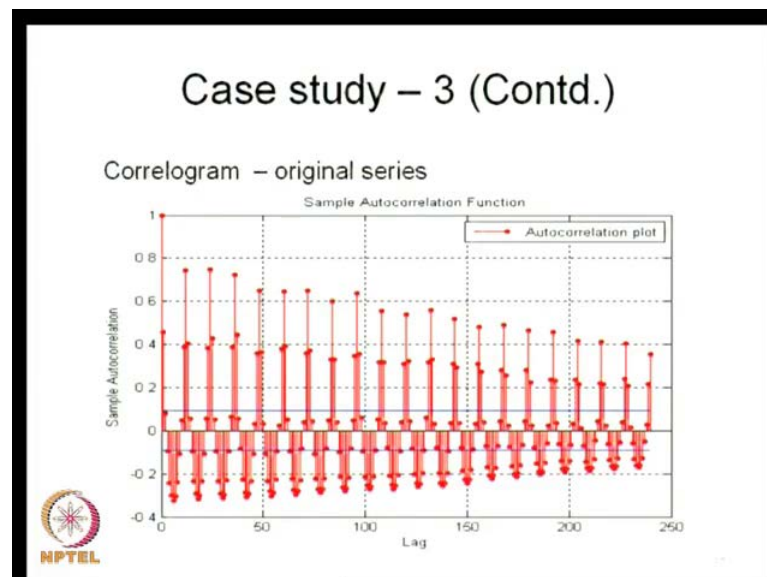
So, the time series of the data, we plot all of this. Then once the periodicity is indicated, let us see how we can remove the periodicities. So, in this particular case, we simply standardize the series, obtain another series by standardizing and then look at whether the periodicities are present or not.

(Refer Slide Time: 47:32)



So, that is a exercise we do now. So, this is the original time series. There are 480 values here that is about 24 years of data, so you have the monthly stream flow data. As you can see, there is a regular pattern here. So, you get suspect that they seems to be some periodicities present here.

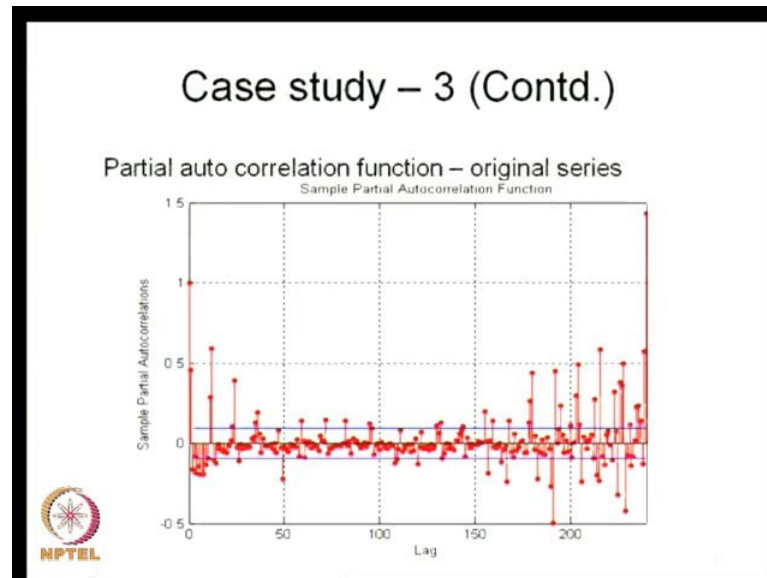
(Refer Slide Time: 47:56)



Then we plot the correlogram as was shown for the previous example, the correlogram shows a sinusoidal oscillation like this and large numbers of correlations are all significant. We have gone up to about 230 or 240; we had 480 values so we gone up to

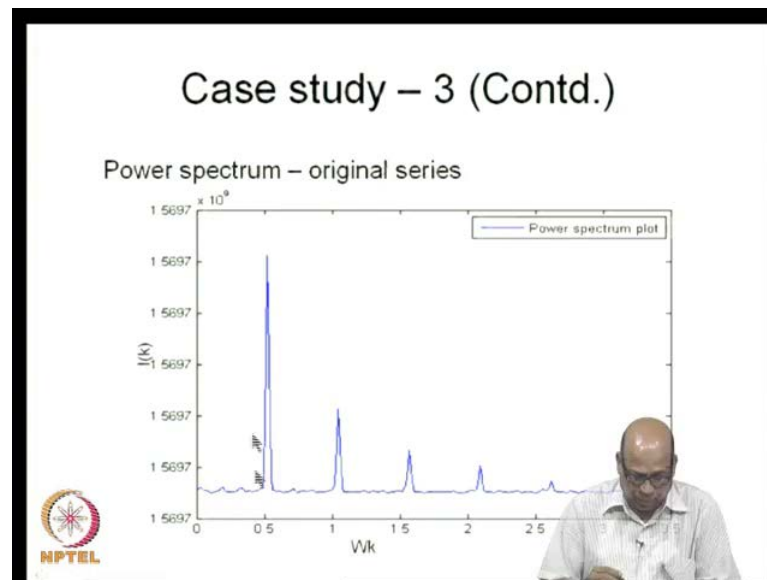
240 n by 2. There is a slow decay here but, the decay is rather slow, because we have gone up to 240 and then they are still significant. So, this shows that there is a significant periodicity indicated by the correlogram.

(Refer Slide Time: 48:37)



Now, let us look at what do the partial auto correlation say? The partial auto correlations also are significant at several lags, especially towards the end, but if you leave out these further points, there are significant partial auto correlations at the initial locations. Even these are significant, but there are somewhere around let say 8, 10 etcetera these are quite significant here.

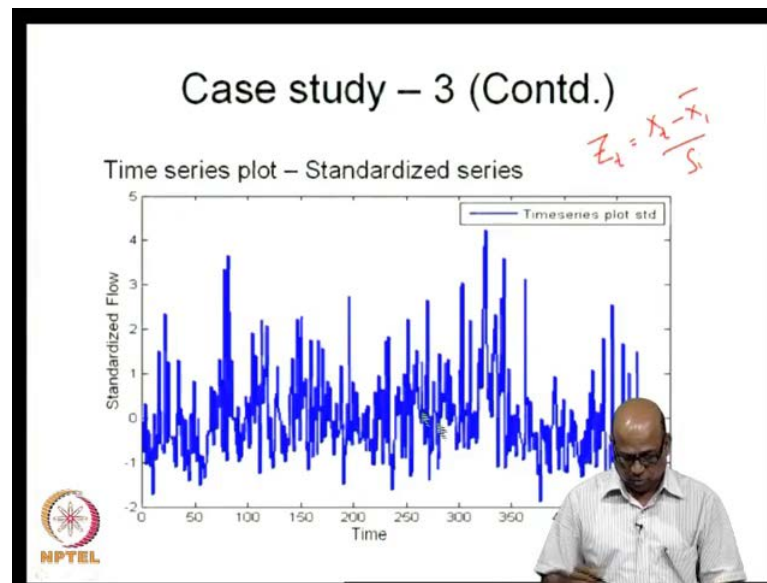
(Refer Slide Time: 49:07)



Then we do the power spectrum; the power spectrum much like the power spectrum that we considered in the stream flow case that I discussed in case two, these show up, I am sorry, in the case one, where we added up the monthly rainfall data, these show up significant periodicities and this periodicities corresponding to 12 months, this to 6 months, 4 months, 3 months and 2 months.

So, these periodicities in the monthly data are indicated by the power spectrum. Now, what we do is we want to remove the periodicities. We do not even examine whether these periodicities are significant or not. We have observed that there are periodicities present in the data, I just want to remove the periodicities. We could have tested with first order differencing, second order differencing, etcetera whether any effect is there by the differencing on periodicities. But what we do is we directly go to standardization. Let us see whether the standardization removes the periodicities.

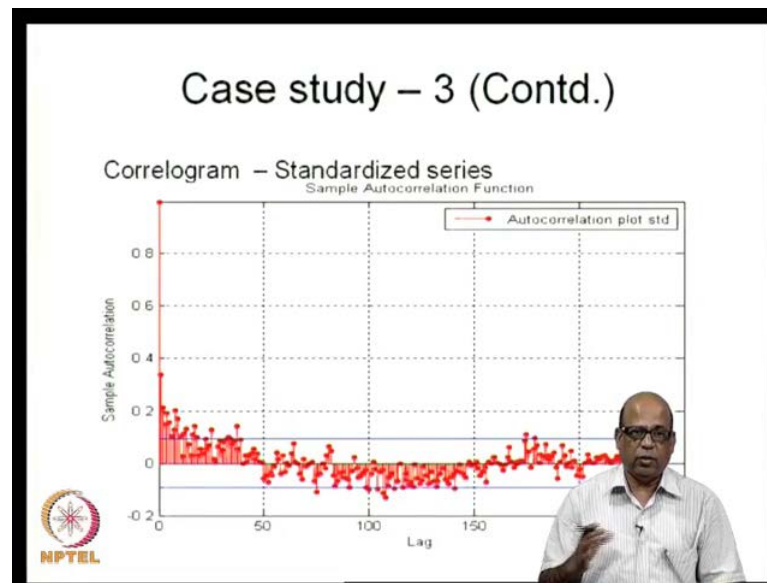
(Refer Slide Time: 50:16)



So, I do the standardization. How do I do the standardization? If you recall, we simply take; this is for the standardized data, so this is Z_t is equal to X_t minus \bar{X}_i by S_i . This is how I obtain the standardized stages. Now, \bar{X}_i is the mean of the month to which the time period t belongs and S_i is the associated standard deviation. So, that is how we standardized the data and then plot the standardized data.

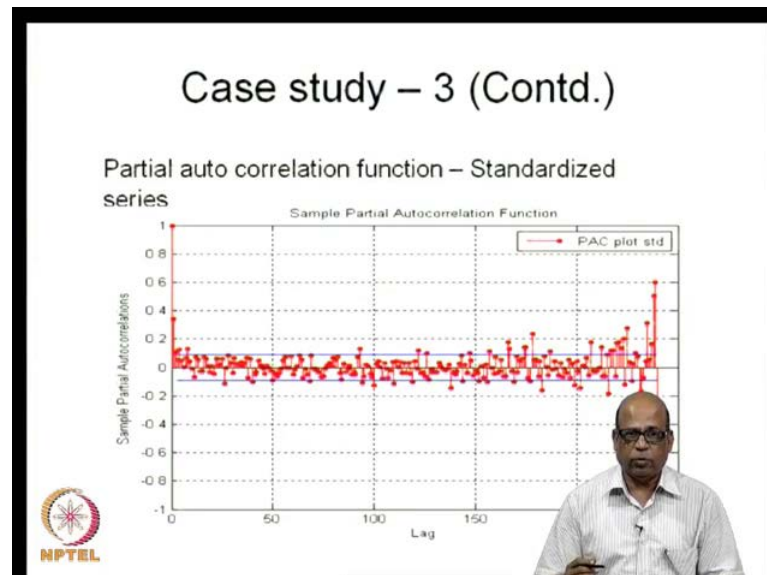
By the data alone, we would not be able to say much here, from the standardize data except that this looks much different from your original data. Original data did show up some regularity when that regularity is not seen in the standardized data but, we will plot the correlogram. That will give us much better information.

(Refer Slide Time: 51:10)



As we plot the correlogram as you can see here, the original data showed up a regular sinusoidal decay in the correlogram, whereas, that kind of behavior is completely absent in the correlogram in the standardized data. So by standardization you have removed much of the periodicities but, this we can ascertain by plotting the power spectrum.

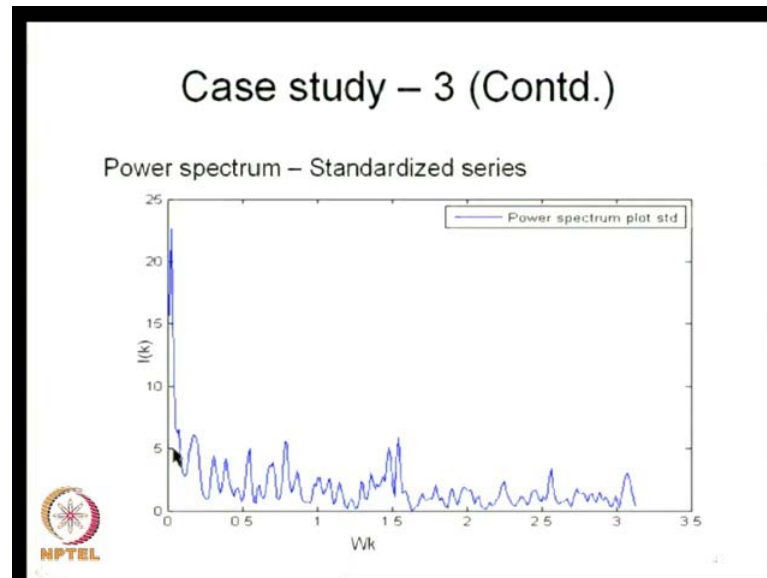
(Refer Slide Time: 51:38)



The partial auto correlation function again, most of the partial auto correlations, become insignificant and there are certain partial auto correlation, which are significant towards the end of this. However, we are not yet at the stage of model building, so we will not

consider this information at present. But, let us look at power spectrum, because our aim was to examine whether by standardization, we are in fact able to move the periodicities or not.

(Refer Slide Time: 52:05)




So, the power spectrum here shows that most of the periodicities are all removed, because you are seeing that there are no significant spikes here, which indicates that there is a major contribution to variance at that particular point. So, the power spectrum more or less shows that the variance is uniformly distributed or more or less uniformly distributed across all the frequencies.

(Refer Slide Time: 52:42)

Case study – 3 (Contd.)

- Standardized series is considered for fitting the ARMA models
- Total length of the data set $N = 480$
- Half the data set (240 values) is used to construct the model and other half is used for validation.
- Both contiguous and non-contiguous models are studied
- Non-contiguous models consider the most significant AR and MA terms leaving out the intermediate terms

11

So, what we now do is that we have removed the periodicities but, there are certain correlations present in the data. There are certain significant correlations present in the data. We use now the standardized series and use the fact that it does not have any significant periodicities present in the data. We will build a model, ARMA model on this data and do the entire test that are prescribed for the data.

I repeat here again, the procedure that we are adopting in this particular course is not just the identification, parameter estimation or the calibration and validation, as has been classically described by Box-Jerkin type of models. We adopt more or less the procedures prescribed by Kashyap and Rao, 1976, the text book, which I have been referring in the last lecture. We adopt that procedure in which we formulate the candidate models.

In formulating the candidate models, we use the information provided by the correlogram, information provided by the spectral density and the partial auto correlations, just to see which of the parameter, how many of the AR parameters and how many of the MA parameters to be included. So, the partial auto correlations, auto correlogram, auto correlations and the spectral density, together give us some indication on how many of AR terms, how many of MA terms, which of the terms to be included in the non-contiguous models and so on.

So, this is information that we get from the earlier analysis that we have done. Based on this, we form certain number of candidate model, we do not pin point to a particular model. We form number of candidate models and then associated with each of the candidate model; we obtain the likelihood values and the mean square values. So, this discussion we will continue in the next class. So, what we discussed today is essentially how to use the ARMA type of models for one time step ahead forecasting and for long term data generation.

Then we examined three case studies. In the first case study, we started with daily rainfall data, and then we aggregate it to monthly rainfall data and then considered also the annual rainfall data. Then we saw the different information contained in the information, contained in these different series like the daily series, monthly series and the annual series and then we considered also a monthly stream flow data.

Before that we considered the annual stream flow data, which showed a trend and by first order differencing, we showed how the trend has been removed from that particular data. Remember, these are all actual case studies, where we are using actual observed data. The first one was for Bangalore rainfall data, the second case study is from a U S site, the data has been downloaded from U S site but, the analysis has been done by one student here and then the third case study was for the Cauvery river data.

We have considered the actual observed data and then made these analysis to see what kind of information is obtain from all of these analysis techniques that we have discussed. Now, in the next class, we go one step ahead and then look at the same case study, the third case study, and see how we build the model, both for long term synthetic data generation as well as for forecasting. Typically, we will also discuss how the residual series are obtained, how we do the test statistical test, how we decide whether there is any significant periodicity present in the residual series or not and so on.

So, **thank you** very much for your attention, we will continue our discussion next time.