

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

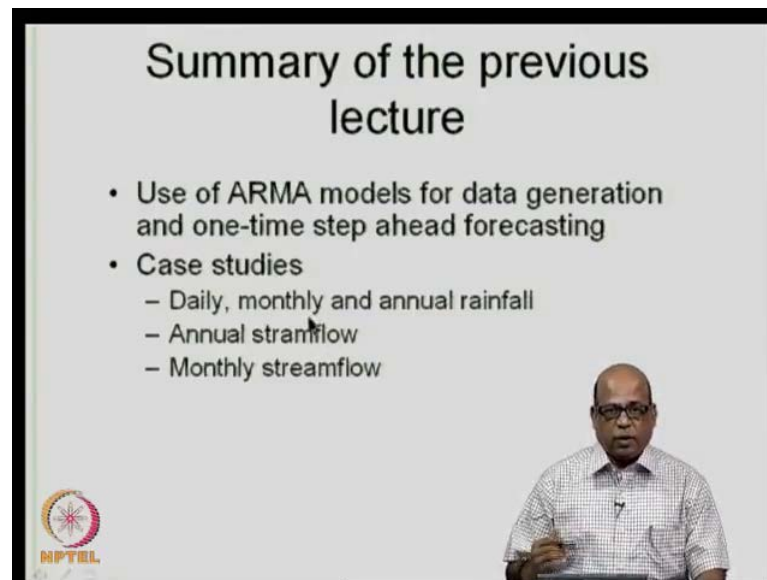
**Lecture No. # 19**  
**Case Studies – II**

Good morning, and welcome to this the lecture number 19 of the course on stochastic hydrology. In the last lecture, we have essentially seen how the ARMA type of models can be used for data generation and also for the data forecasting. By data generation, I mean you want to generate a sequence of data, which essentially reproduces the behavior of the historical time series. Now, this sequence typically is generated for a long length, typically of the order of 50 years 100 years and so on.

In fact, in certain cases where we would like to use this data for examining the performance of the system, long term performance of the performance of the system, we go as long as 200 years 300 years and so on, using these types of models. Then we also saw how we use the ARMA models for one time step ahead forecasting that is standing at the end of a particular month, let us say June month, we would like to forecast what would be the stream flow during the next month. So, this is called as one time step ahead forecasting.

The ARMA models can also be used for two times step ahead forecasting, three times step ahead forecasting and so on. Then in the previous lecture we also saw some case studies. We started with the case study of daily rainfall in an urban area; typically we took the Bangalore rainfall data on that. And then we saw from daily rainfall, if we aggregate it to monthly rainfall how the behavior changes. In the daily rainfall, we did not see any periodicities. In fact, the daily rainfall behaved much like a completely random process, in the sense that the values were all uncorrelated. We could not detect any significant correlation, autocorrelation and the spectrum showed that it behaves more or less like a white noise.

(Refer Slide Time: 02:28)



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

- Use of ARMA models for data generation and one-time step ahead forecasting
- Case studies
  - Daily, monthly and annual rainfall
  - Annual streamflow
  - Monthly streamflow

In the bottom right corner of the slide, there is a small video inset of a man with glasses and a light-colored shirt, who is the presenter. In the bottom left corner of the slide, there is a logo for NPTEL (National Programme on Technology Enhanced Learning).

But when we aggregated these to a monthly time series that is simply adding up all the daily rainfall values into a monthly rainfall value, we saw that the periodicities corresponding to 12 months came up, and periodicities corresponding to 6 months, and 4 months came up, which were not present in the daily data. Then we went on to aggregate this monthly data to annual rainfall. Again, in the annual rainfall, we did not detect any significant periodicities this may be because of the limitations of the data itself, because we do not have large number of yearly data.

If you had in fact large number of yearly data, let us say 100 years or 150 years of data and so on, and then you did the test for periodicities for annual values of rainfall. Perhaps, you may come out with some decadal periodicities and so on. But for the case study that we considered, we did not detect any significant periodicities in the annual rainfall. So, for the monthly rainfall, when we did the analysis, the monthly rainfall data showed up significant periodicities.

I'm sorry it showed up periodicities at 12 months, 6 months and 4 months. But, whether they are statistically significant or not, we need to examine using the test that I have discussed in the earlier lectures.

Then we consider another case study where we took the annual stream flow, which had a increasing trend. Then we examined how we can remove the trend from the data by differencing. So, we consider the first order differencing that means  $X_t - X_{t-1}$

1, so from the  $X_t$ , which is the annual stream flow, tiding corresponding to the time period of  $N$  year and  $X_{t-1}$  is the previous value. We constructed a new series  $Y_t$  is equal to  $X_t - X_{t-1}$ , and then also did check on the different series whether the trend has been removed. We in fact observed that the trend has been removed.

So, we did the same analysis then on the annual stream flow as we did for the rainfall data. We examined the correlogram, the specter density, and the partial autocorrelations and so on. Based, on this analysis that we said it may perhaps follow a purely AR model, AR 2 perhaps because the number of partial autocorrelations that were significant in that case study was two. Then we considered the monthly stream flow of the Cauvery River at KRS reservoir. We will continue this case study today, but essentially what we did in the last lecture is that we first plot the time series. The time series immediately shows that there may be perhaps some periodicities present. Then this fact is verified by the correlogram; correlogram again shows up that it is a sinusoidal correlogram and that indicates that there are periodicities present.

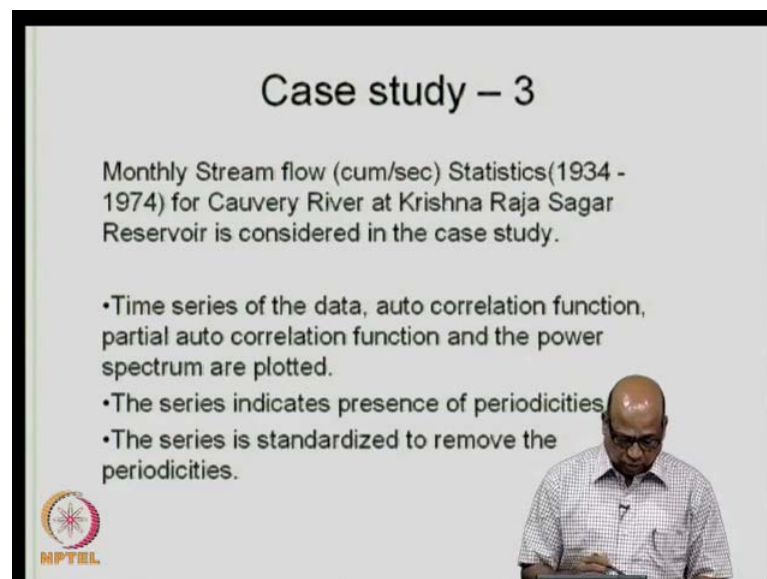
The spectral analysis shows up, which are this periodicities and we identified in the last lecture that you have periodicities corresponding to 12 months, 6 months, 4 months and 3 months. This is typically true of most of the monsoon driven stream flows where you will have regular regularities in the flows, because these flows are essentially produced especially in the peninsular region of the country. These are essentially produced by the rainfall and the rainfall in the monsoon regions has a regular pattern. Therefore, you will typically get a 12 month periodicity, 6 month periodicity and perhaps in certain cases 4 months' periodicities.

This need not be true for all the types of stream flows that you may have globally, across the globe, because it is the way the stream flows are produced. Perhaps, if you have a continuous small intensity rainfall all through the year, like it happens in some of the European regions, then you may not get so significant periodicities corresponding to 12 months and 6 months and so on. Therefore, the information that the data is giving us is important with respect to the region in which we are present and the particular process, the particular physical process by which these variables are generated. What I mean by that is here in the daily monthly and annual rainfall, perhaps this may be monsoon driven rainfall or at certain daily levels it may be just a convective type of rainfall and so on.

So, the periodicities that you examine or the periodicities that come up and the correlations, lag correlations that come up, will all depend on how these are actually generated and what kind of relationships that exist with respect to the process that has taken place in the earlier time periods. We will continue with the third case study that is the stream flow, monthly stream flow at KRS reservoir. Remember, in all earlier cases that I discussed in the last lecture, we did not go on to build the models, we just looked at the spectral density, we looked at the correlogram, we also looked at the partial autocorrelation and then try to get what is the best information that we can extract from these analysis.

So, now we will proceed with the case study three, and see how we build the models and how we test the models for validity or validation tests we carry out. So, these are case studies on ARMA models.

(Refer Slide Time: 08:55)



**Case study – 3**

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

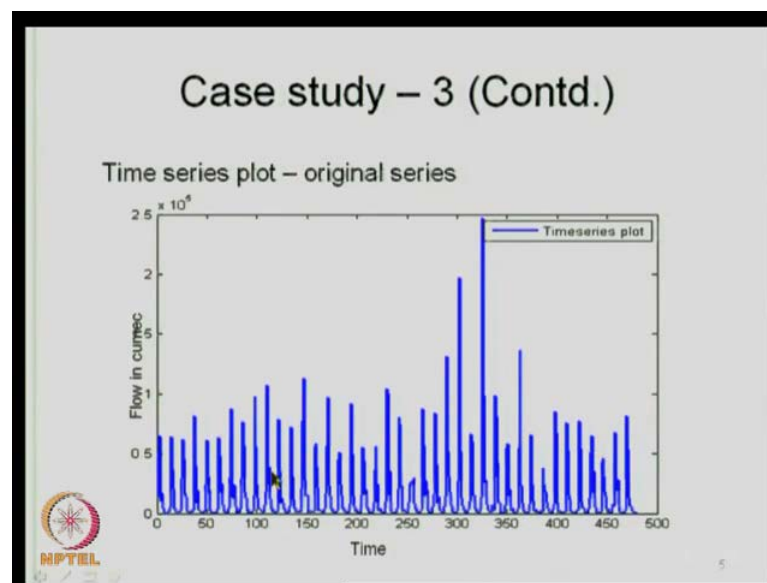
- Time series of the data, auto correlation function, partial auto correlation function and the power spectrum are plotted.
- The series indicates presence of periodicities
- The series is standardized to remove the periodicities.

NPTEL

We are considering case study 3 where we are looking at the monthly stream flow in cubic meters per second, it also denoted as cumec, and we are taking the period 1934 to 1974. This is for Cauvery River at Krishna Raja Sagar reservoir. Now, the time series of the data autocorrelation function, partial autocorrelation function and power spectrums are plotted and I showed it in the last lecture, but for completeness, we will again see those plots.

Now, this time series indicates presence of periodicities. We would then ask the question how would we remove these periodicities and one of the ways of doing that will be keep on differencing the series. consider the first difference, second difference, third difference etcetera and keep examining the different series for the presence of periodicities, until you are satisfied that all the periodicities have been removed from the data. Another way of doing it, as I discussed in one of the earlier lectures, is that you can use standardization. So, we use the standardization. As standardization, I mean  $Z_t$  you can construct a new series  $Z_t$  is equal to  $X_t$  minus  $\bar{X}$  by  $S$ .

(Refer Slide Time: 10:28)

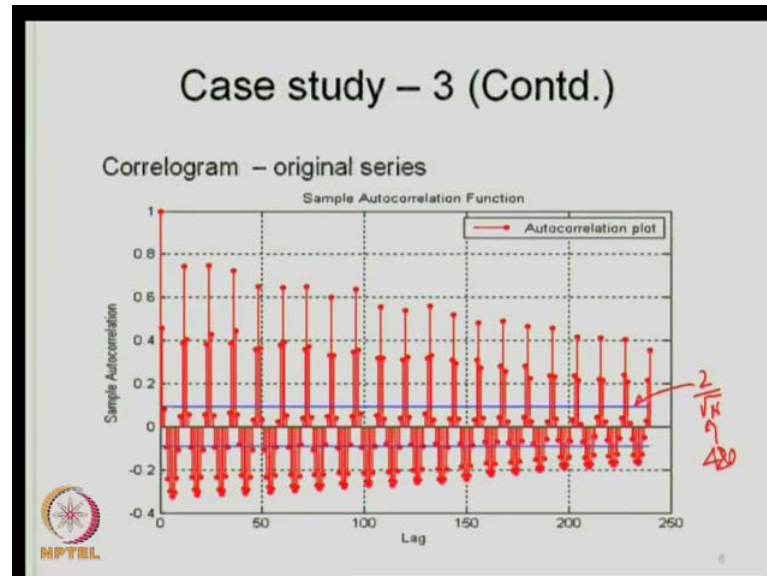


So, that is how you construct the standardized time series. I will just come to that again. So, this is how the time series look like. Always the first two step, when we do time series analysis is once you have the data on the particular process. In this particular case, we are considering the stream flow. Simply plot the data as a time series, so time on the x axis and the particular variable on the y axis. The time series plot shows certain type of regularity here, like this it is going in a regular manner.

So, it shows that there is certain kind of inherent regularity or periodicity present in the data, which as I said can be expected in most of the monsoon driven stream flows. Typically, you get you get high flows during the monsoon periods and very low flows during non monsoon periods and so on. So this keeps on oscillating like this now this

information contained comes up more clearly when we plot the correlogram, so we plot the correlogram.

(Refer Slide Time: 11:38)



Correlogram shows a sinusoidal variation like this and it is going on in a periodic manner. So, it shows that there is definite periodicity present in the data, also in most of the correlograms; you will observe especially the hydrologic time series especially when you are considering the flows monthly flows etcetera.

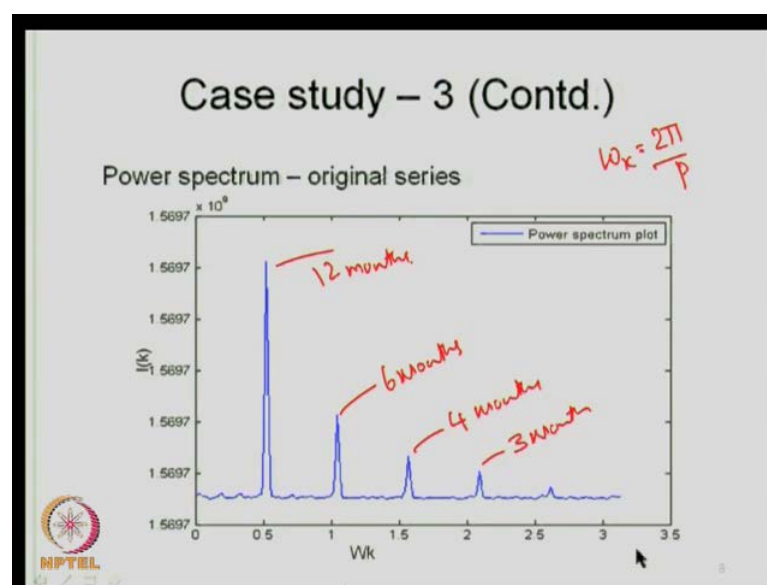
There is a slow decay in the correlogram. As you can see here, there is a slow decay in the correlogram. Although most of the correlations are still significant, as you can see these are the significant bands. How do we plot these significance bands? It is plus or minus 2 by root N, so this significant band is simply 2 by root N and on either side. So, this 2 by root N, this N is a number of data. In this particular case, we had 480 values for the Cauvery data, so 480, so 2 by root N and this will give you the band.

What does the correlogram indicate? The correlogram indicates that there is a periodicity present in the data. Now, can we say there is a periodicity associated with 12 months because it has been oscillating around 12 months? Let us say the first 12, first 6 months, it goes, then 12 months it goes up and so on, like this it is oscillating. So, you suspect that there may be a periodicity corresponding to 6 months, there may be a periodicity corresponding to 12 months and so on, because there is a regular oscillation that you are seeing and it is in fact sinusoidal type of oscillation.

To confirm, that there indeed is a periodicity present in the data and to look at which are these periodicities that are being thrown up from the data, we do the spectral analysis. So, recall that in the spectral analysis, we convert the time domain information, the time series into a frequency series, in the frequency domain. And then we analyze the same time series,  $X_t$  in the frequency domain, and then look at the distributions of the frequencies or the variance that exist in different bands of frequencies. Let us not lose the site of what we are doing in the spectral analysis; especially we are looking at the time domain having as being constituted of a large number of frequencies. Then we are looking at the variance content or which of these frequency bands contribute more to the variance that has been observed in the data.

That shows the periodicity that means there may be certain frequencies at which there is a significant contribution to the variance and then there are certain other frequencies where there is much less contribution to the frequency. If you have a purely random sequence then the distribution will be more or less uniform. That means there is no single frequency, which contributes more to the variance than any other frequency, in the case of purely random frequencies. In fact, if you consider white noise as we saw in the last lecture and also as we see in the case studies that we consider, if you have your white noise then the spectral density will be more or less scattered and it will not show up any significant frequencies around which the variance is contributed significantly.

(Refer Slide Time: 15:45)

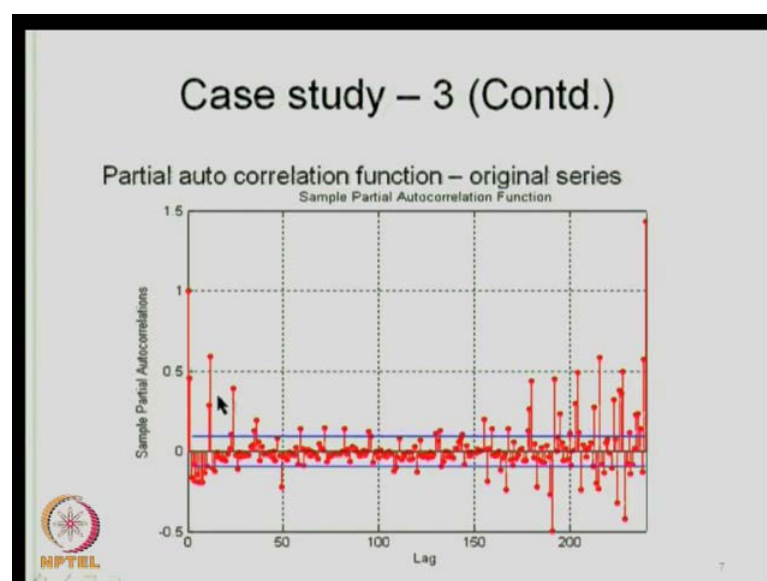


So, let us look at how the frequency diagram looks like. Now, this is for the monthly time series; this is the original time series of the Cauvery River. You get a periodicity corresponding to  $\omega_k$  and are equal to somewhere around 0.52 or some such thing.

This in fact corresponds to a periodicity of 12 months. How do we get it? We get it from  $\omega_k$  is equal to  $2\pi$  by  $P$ , we will simply do  $2\pi$  by whatever value that you get, here 0.52 or some such thing. When you do the tabulation, you will know this. So, this is around 0.52 and therefore, the periodicity comes out to be 12 months. Similarly, the periodicity corresponding to this spike is 6 months and this is 4 months and this is 3 months and so on.

What was not so obvious in the correlogram? In the correlogram, we only noticed that yes there may be certain periodicity in the data, because it is oscillating in a sinusoidal manner and there is regularity in this oscillation; it is also decaying slowly with respect to lag. Therefore, we suspected that there is a periodicity present in this data. Now, this information comes up most strongly in the spectral density. So, when you plot the spectral density, you will see that there are spikes present in the spectrum. So, these spikes correspond to a periodicity of 12 months. We convert the  $\omega_k$  that you get here, these are denoted both  $\omega_k$  as well as  $\omega_k$ , this corresponds to 12 months, 6 months, 4 months and 3 months.

(Refer Slide Time: 17:40)





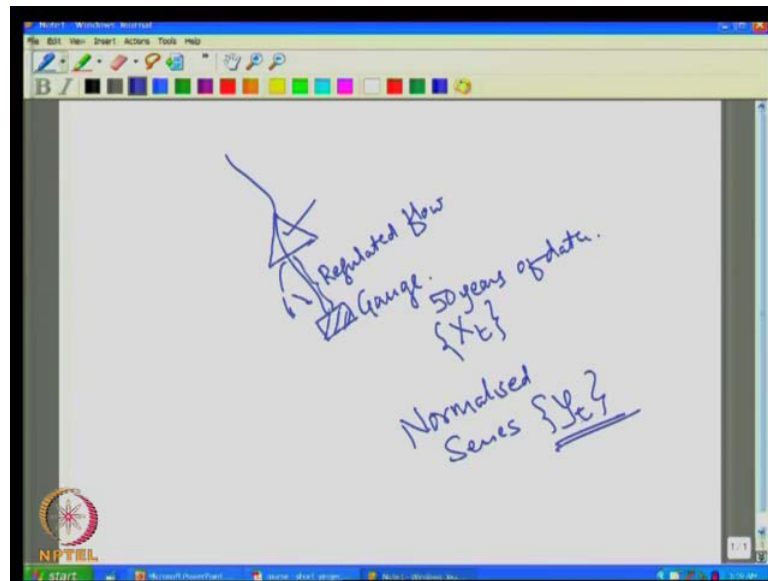
We also look at the partial autocorrelation function. So, this is for the original series. You see that there are several partial auto correlations that are significant, typically when you are looking at the partial autocorrelations, we look at initial lags. These further away points are not so important when you are looking at the partial autocorrelations. We may be looking at let say up to first 50, first 60 and so on, so there are certain, in this particular case, 480 values are there. So, you can look up to about 15 percent of the total data, so you look at these partial autocorrelations you see that there are partial autocorrelations, which are also significant.

How do we use the partial information provided by the partial autocorrelation? As I mentioned, in the identification of the models, you look at both the correlogram as well as the partial autocorrelations to see whether the model can be approximated with the AR model or a MA model. If you have the correlogram slowly decaying, which in this case we have observed that it is and you have some significant partial autocorrelations present, then you may suspect that you may model it with an AR model.

And the number of terms of AR model corresponds to the number of significant partial autocorrelations. When you have seen that the correlogram is showing a slow decay, it may be an exponential type of decay or as we saw in this particular case, it may be a sinusoidal form of decay. Now, this is information that we get from the data analysis. So, typically in any of the data analysis we do this. That we plot the time series first, we see if there are any periodicities that are indicated in this. In fact, in the earlier case study that we considered the moment we plotted the time series, the annual stream flow case study that you can look back it immediately showed that there is a trend.

So, the time series plot itself gives us an important insight into the type of process that we are dealing with and therefore, the first step always is to look at the time series. Now, when we are considering the stream flows, you must remember that the stream flow models that we are building have to be done on what are called as the normalized stream flow series, which means that if you have any regulated contribution to the stream flow. Let us say that you are looking at a gauge site at which you have observed the data. But, this gauge site has been governed by the data, as this gauge point has been governed by let us say that or let me show the particular feature that I am just mentioning.

(Refer Slide Time: 20:46)



Let us say you have the data at this particular location. This is the stream gauge, and this stream gauge has given you last 50 years of data. You may have a reservoir here. Let us say you may have a reservoir like this and what you are getting here is regulated flows from the reservoir. What do I mean by that? You are having a control flows let down into the stream and this gauge is measuring only the control flows perhaps in addition it may have its own catchment here. It has its own catchment, so you are getting a regulated flow plus the catchment flow.

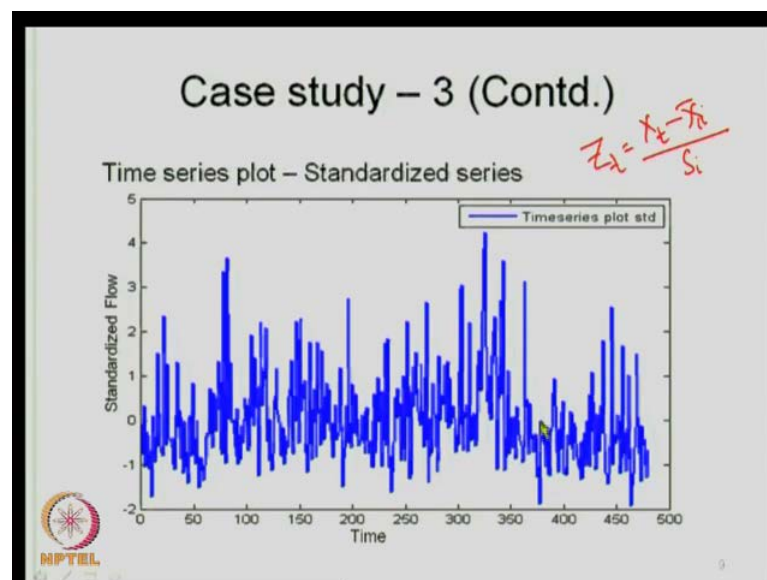
So, when we are doing the time series analysis, we must make sure that the effect of this upstream structure is removed in some sense. There are ways of doing it perhaps in the applications part; I may discuss some of these methods. You remove the effect of the upstream structure, construct the time series. This was observed time series, which is from the regulated flows and plus the catchment area you construct a normalized time series,  $Y_t$ , which choose this  $X_t$ , which has been observed and then you remove the effect of this storage here, storage structure, and then account for the flows that would have naturally occurred and that would have been observed at this location.

You normalize the series and then on this normalized series  $Y_t$ , you do the time series analysis. Always students tend to do this mistake especially during the projects that they take up as part of their academic degrees and so on. That whatever gauge flow is available they start doing the time series analysis but, in the case of stream flow,

particularly you must be careful in looking at what exactly is the gauge measuring; whether it is measuring the natural flows or is it measuring the natural flows plus certain component of control flows.

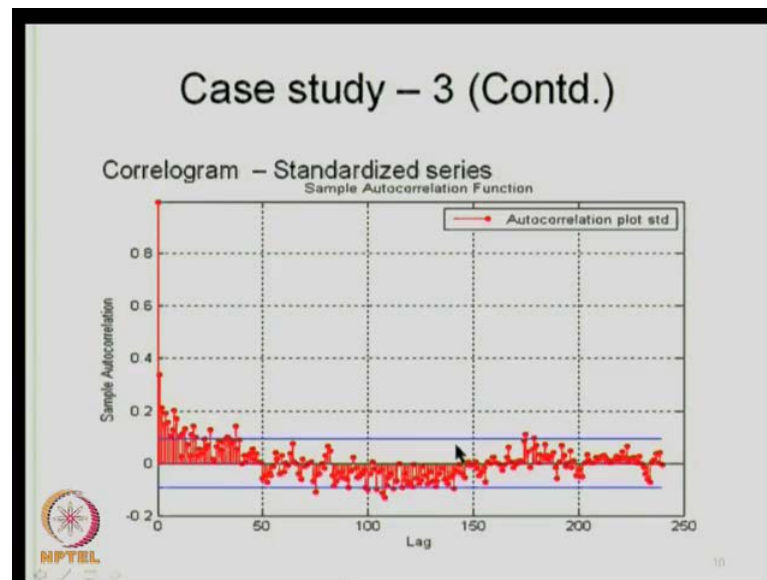
So, this must be kept in mind when we are doing the time series analysis. So, let us come back to this, so if you do not have any major control structure upstream of this and you look at the time series, then this is how typically it looks for monthly time series in the monsoon regions. All right now, let us progress we saw that these were the power spectrum indicated 12 months, 6 months, 4 months, periodicities and so on.

(Refer Slide Time: 24:05)



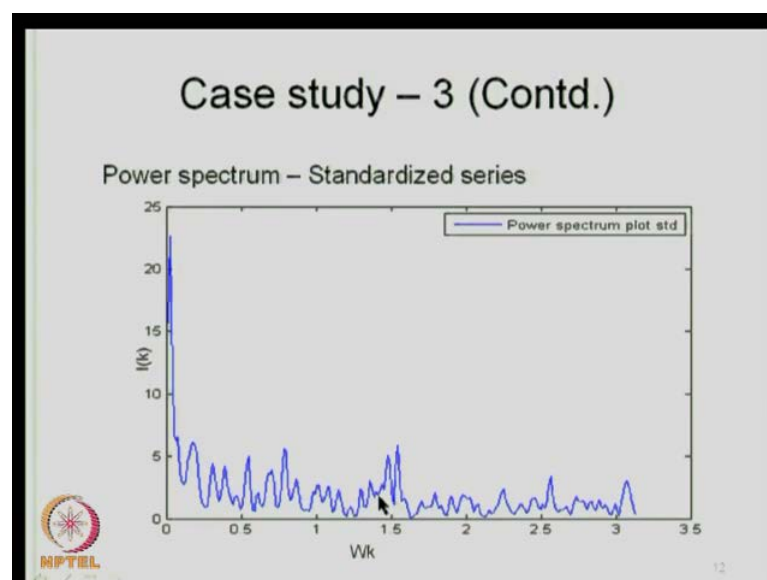
Now, let us say that we want to examine the effect of standardization on the periodicities. So, this is the original time series, this original time series now we standardize. We standardize by like I mentioned last time you take  $Z_t$  is equal to  $X_t$  minus  $\bar{X}_i$  by  $S_i$ , where  $\bar{X}_i$  is the mean of the flows of the month to which time period  $t$  belongs. Similarly,  $S_i$  is a standard deviation associated with that month. So, this is how we construct  $Z_t$ . So, this is what I have plotted here is  $Z_t$  with respect to time. So the standardize time series immediately shows that this is much more random compare to what we had earlier. This looks to be a random series. Then we look at the correlogram associated with this.

(Refer Slide Time: 25:01)



So, the correlogram again we put the same bands of plus minus 2 by root N. These bands remain the same because N remains the same. You see that most of the correlations become insignificant except for the first few correlations and the periodicities that were so permanently shown up by your earlier correlogram of the original time series, they are absent in the correlogram and this is confirmed also by the power spectrum.

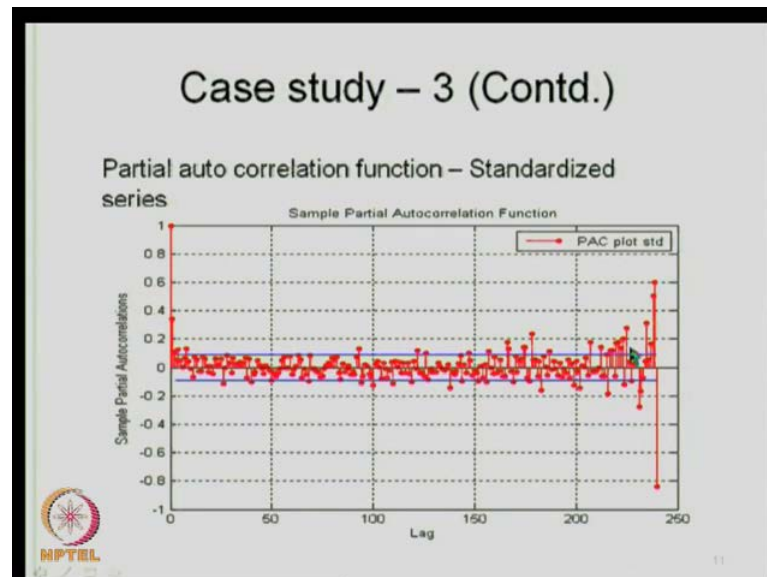
(Refer Slide Time: 25:35)



So, if you look at the power spectrum for the standardize time series as I mentioned, it shows that there are no periodicities here except that for the first one, and first two,

perhaps. There are no periodicities here, which are more significant compare to any other periodicities, which means that all the frequencies contribute more or less uniformly to, all the bands of frequencies contribute more or less uniformly to the variance and the partial autocorrelation looks something like this.

(Refer Slide Time: 26:08)



Most of partial autocorrelations are also insignificant, except towards the very end of the spectrum that we are looking at here. Now, this is information that we get. Now, the summary of this information is that we may feel that there are autoregressive terms.

(Refer Slide Time: 26:22)

**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

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-12} + \epsilon_t$$

13

That may be important, because we are looking at the correlogram being slowly decaying and there are certain partial autocorrelations, which are significant. Therefore, we may suspect that there may be some auto regressive terms here. But similarly, you may also have moving average terms, because the partial autocorrelations also may show certain decay in this particular case and therefore, now we want to start constructing the models.

As I mentioned, we will not adopt the classical approach of identification of the model, based on the partial autocorrelations spectral density and the correlograms and so on and then validation and then parameter estimation and validation. What we will do in studies is that from all of these analysis that we have done, we know how many of AR terms may be present and how many of MA terms may be present. Let us say that you are looking at the correlogram and then you are seeing that correlations at lag 12, keep on turning out to be significant.

At lag 24 it is significant, at lag 6 it is significant and so on, also correlations at lag 1, 2, 3 may be significant. So, you may want to consider those terms, AR terms associated with this particular lags. Then you may also have MA terms, which are shown by your partial autocorrelations. So, you pick let say 2 or 3 MA terms, typically in hydrologic applications, as I mentioned in one of the earlier lectures, you go up to 6 or 7 AR terms then 2 or 3 MA terms for monthly time series more or less this will be adequate.

So, you now construct a candidate set of candidate models and then do all the analysis on the candidate models. Let us see how we do in this particular case? So, this model building can be done on the original series or some transformed series. You may consider either log transform series or you may consider a square root transform series and so on. So, we consider the standardized series. Now, because essentially it has to be a stationary time series, so we consider the standardized series. The total length of data set is 480 or you have 24 years of data, so 480 values, I am sorry you have 40 years of data, so you and it is a monthly dataset, so you have 480 values.

We consider half of this data for calibration of the model that means parameter estimation and essentially parameter estimation, because we would have fixed the structure of the model itself and then the remaining half we use it for validation. So, 240 values, we use it for validation. We consider both contiguous and non-contiguous

models, this case. Like I mentioned, the non-contiguous models, for example, you may have AR 2 for non-contiguous model, which considers  $X_{t-1}$  and perhaps  $X_{t-12}$ . You know that there is a significant correlation still present in the standardized series at lag 12, so we would like to put lag 12 there. But, in the contiguous model, if you want to put lag 12, what you will do?  $\phi_1, \phi_2, \phi_3$ , etcetera, up to  $\phi_{12}$ , you have to go, which means a number of terms you are increasing number of parameters are increasing.


Whereas, in the non-contiguous model, if you want put, let say lag one and lag 12, just the two of them. So, you write the ARMA model as  $X_t$  is equal, to let me write that. You will write the ARMA model as in the case of non-contiguous model. You will write this as  $X_t$  is equal to that is  $\phi_1 X_{t-1} + \phi_2 X_{t-12}$ , we will take plus  $\epsilon_t$  or something, and this is a noise term. So, instead of taking all the intermediate term, we consider only those terms, which we feel are significant. So, like this we construct the non-contiguous models. In fact, for consistency you may call it as  $\phi_{12}$ , so  $\phi_1$  and  $\phi_{12}$  only may be important here, but that is a matter of notation.

Similarly, for MA terms also, what we may do is we may consider the first term as well as a twelfth term or first two terms and the twelfth term, first three terms and the twelfth term and so on like this. By doing this, what we are achieving is we are making sure that the number of parameters in the model, which need to be estimated from the data and the data length, is fixed. Therefore, if you consider large number of parameters, the parameter estimation becomes less reliable. As you have large number of parameters to be estimated with the same length of data, then the models becomes much more cumbersome and therefore, less reliable. So, we will always go with the least number of parameters and this is what is generally called as Principle of Parsimony. So, in time series analysis we always go with Principle of Parsimony.

(Refer Slide Time: 32:33)

### Case study – 3 (Contd.)

- For example, a non-contiguous AR(3), with significant dependence at lags 1, 4 and 12, the model is written as
$$X_t = \phi_1 X_{t-1} + \phi_4 X_{t-4} + \phi_{12} X_{t-12} + e_t$$
- Similarly the moving average terms are also considered and a non-contiguous ARMA(3, 3) is written as
$$X_t = \phi_1 X_{t-1} + \phi_4 X_{t-4} + \phi_{12} X_{t-12} + \theta_1 e_{t-1} + \theta_4 e_{t-4} + \theta_{12} e_{t-12} + e_t$$



So, as I mentioned for example, a non-contiguous AR 3 model, with significant dependence at lags 1 4 and 12, this we may write it as  $X_t$  is equal to  $\phi_1 X_{t-1}$  plus  $\phi_4 X_{t-4}$ , plus  $\phi_{12} X_{t-12}$  plus the noise term. The intermediate terms here are  $\phi_2, \phi_3$  we have left out and similarly,  $\theta_2, \theta_3, \dots, \theta_{11}$ , we have left out. We are simply looking at those particular terms, which indicates significant dependence through their lag correlations and we include only those particular terms.

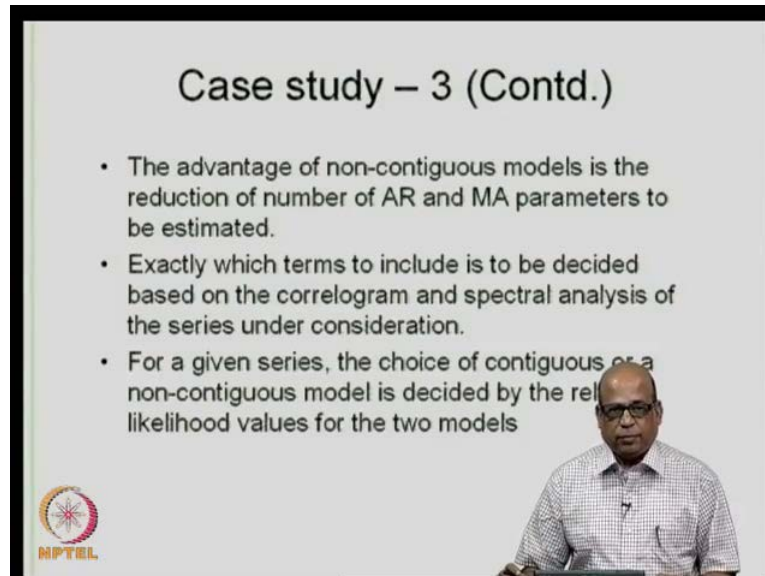
Similarly, we do that for the moving average terms also. We write the moving average terms for an ARMA 3 3 model. Let say where we are considering same lag 1 4 and 12. Then we write  $\phi_1 X_{t-1}$ , plus  $\phi_4 X_{t-4}$ , plus  $\phi_{12} X_{t-12}$ , plus the same dependences we take  $\theta_1 e_{t-1}$ , plus  $\theta_4 e_{t-4}$ , plus  $\theta_{12} e_{t-12}$  plus the noise term. So, this is how a non-contiguous ARMA 3 3 model is written. What we would have done for a contiguous ARMA 3 3 model? we would have said  $\phi_1 X_{t-1}$ , plus  $\phi_2 X_{t-2}$ , plus  $\phi_3 X_{t-3}$  plus  $\theta_1 e_{t-1}$  plus  $\theta_2 e_{t-2}$  plus  $\theta_3 e_{t-3}$ , so we would have gone continuously up to the last term there and add the noise term.

We in the candidate model selection what we do is we construct both contiguous models as well as non contiguous models depending on which terms we feel are significant and which terms need to be indicated, included in the model, we include those terms, those AR as well as MA terms and then construct the contiguous models also. So, for this set



of identified candidate models, we do the parameter estimation then we do also the validation and we compute the likelihood values associated with each of these models.

(Refer Slide Time: 35:04)



The slide is titled "Case study – 3 (Contd.)" and contains three bullet points. In the bottom right corner, there is a small inset image of a man in a checkered shirt and glasses, and the NPTEL logo is in the bottom left corner.

- The advantage of non-contiguous models is the reduction of number of AR and MA parameters to be estimated.
- Exactly which terms to include is to be decided based on the correlogram and spectral analysis of the series under consideration.
- For a given series, the choice of contiguous or a non-contiguous model is decided by the relative likelihood values for the two models

Let us do that and see which model we select in the contiguous model as I said exactly, which terms to include is to be decided based on the correlogram and spectral analysis of the series under consideration. So, correlogram typically shows up some of the significant lags. So, look at the correlogram then include those particular terms now. let say you had a series of a set of contiguous models and another set of non-contiguous models, which one do you chose? As I mentioned, you compute the likelihood values for each of these models then look at the maximum likelihood value from each of these.

So, one model comes up from the contiguous and another model comes up from the non contiguous. Now, this is now a matter of judgment. You may either look at the likelihood value itself and if there is a significant difference between the maximum likelihood value of the non contiguous models compared to the maximum likelihood value of the contiguous models, then you have to choose that particular model either the contiguous or the non-contiguous, which gives you whose likelihood value is more than the other one. Whereas, if they are quite close to each other, if the maximum likelihood values of the candidate models of the models that you have selected are close to each other, then you may go with the particular model which has lesser number of parameters.


(Refer Slide Time: 36:43)

### Case study – 3 (Contd.)

$L_i = -\frac{N}{2} \ln(\sigma_i) - n_i$

Contiguous models:

Sl. No	Model	Likelihood values
1	ARMA(1,0)	29.33
2	ARMA(2,0)	28.91
3	ARMA(3,0)	28.96
4	ARMA(4,0)	31.63
5	ARMA(5,0)	30.71
6	ARMA(6,0)	29.90
7	ARMA(1,1)	30.58
8	ARMA(1,2)	29.83
9	ARMA(2,1)	29.83
10	ARMA(2,2)	28.80
11	ARMA(3,1)	29.45



Let us see what happens in this particular case. We compute the likelihood value for each of the models. These are candidate models that we considered for the contiguous case; ARMA 1 0, 2 0, 3 0, 4 0, etcetera up to 6 0. Then we start including the MA terms 1 2, we go up to maximum of two terms, then we go up to AR 3 terms, in the case of ARMA 3 1. So, these are the candidates' models we consider. We went up to AR 6; because your correlogram showed that there may be certain correlations, which are significant, so that gives us an idea. See here, you are seeing standardized series; there are quite a few significant correlations so you have gone up to AR 6.

Similarly, the partial autocorrelations show that there may be some significant partial autocorrelation up to about 6, so you want to include the AR up to AR 6. In fact, you can go up to AR 10 and so on. Just to make sure that you do not miss the actual model but, typically the MA terms, we consider for only up to 2 or maximum 3 in several cases. Then you go up to ARMA 3 1. So, this is how we constructed the candidate models. In fact, ARMA 3 2 also can be one of the models, so for each of these models these are contiguous models. So, you know how to formulate the models for each of these models. We first estimate the parameters and then you also get the likelihood values.

Let say that the likelihood value is this, is actually the log likelihood value, you have minus  $N$  by  $2$  log  $\sigma_i$  minus  $n_i$ ;  $n_i$  is a number of parameters. Say for example, here it is three, and in this case it is again three, in this case it is 4. So, the total number of

parameters  $\sigma^2$  is the variance of the residual. For any of these models, you would have computed the parameters. Then in the calibration period, you would have got the residuals and that residual variance is  $\sigma^2$ , so  $n$  is total number of values. In this case, it is 480, so you know for every model you can get the likelihood value. Like this you get the likelihood values for each of these models.

The model that gives you the maximum likelihood value is the model that is chosen. So, in this particular case 31.63 happens to be associated with ARMA 4 0 and that is also the maximum values. So, we say that ARMA 4 0 model is chosen for long term synthetic generation of the data for the Cauvery river flows at Krishna Raj Sagar as far as contiguous models are concerned.

(Refer Slide Time: 40:09)

**Case study – 3 (Contd.)**

Non-contiguous models\*:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-12} + \epsilon_t$$

Sl. No	Model	Likelihood values
1	ARMA(2,0)	28.52
2	ARMA(3,0)	28.12
3	ARMA(4,0)	28.21
4	ARMA(5,0)	30.85
5	ARMA(6,0)	29.84
6	ARMA(7,0)	29.12
7	ARMA(2,2)	29.81
8	ARMA(2,3)	28.82
9	ARMA(3,2)	28.48
10	ARMA(3,3)	28.06
11	ARMA(4,2)	28.65

\*: The last AR and MA terms correspond to the

Now, we will do this also for the non-contiguous models. We construct the non contiguous models, every time taking the last term to be  $X_{t-12}$  or  $\epsilon_{t-12}$ . That means the twelfth lag; we are taking it as significant, so the last AR and MA terms correspond to the twelfth lag. Let me see if we have, let say that I want to write ARMA 2 0 non contiguous model with the twelfth lag. How do I write this? I will write this particular model as non-contiguous model. This is non contiguous model, so I will write this as  $X_t$  is equal to  $\phi_1 X_{t-1}$ . The first term I keep it the same, the second term I will put it as  $X_{t-12}$ .  $X_{t-12}$ , you can use it as  $\phi_{12}$  or  $\phi_2$ , in this

particular case, we can go as phi. Let me rewrite this so this is how you write the non contiguous model.

So, in the non-contiguous model, we are essentially writing only and we are taking only those terms, which we feel are significant and are important. So, this is how we write the ARMA 2 0 models, similarly, you can write ARMA 2 2 model,  $\phi_1 X_{t-1} + \phi_2 X_{t-12} + \theta_1 X_t + \theta_2 e_{t-1} + \theta_2 e_{t-12} + \epsilon_t$ . So, this is how we write the non-contiguous model. Once, you write these models you can go to your parameter estimation algorithms and estimate the parameters and apply these models to the calibration data. in this particular case, the first 240 values we apply get the residuals, get the variance of the residuals and then calculate the likelihood value.

So, for the non-contiguous models, the likelihood values turn out to be like this 28.5 to 28.5 etcetera. So, the ARMA phi 0 of the non-contiguous type of models comes out with a likelihood value of 30.85, which is the maximum likelihood value. So, we either choose now the non-contiguous ARMA 5 0 model or the contiguous ARMA 4 0 model. Now, the likelihood value that is present here is 31.63 and this is 30.85. What are the terms that this model will have? It will have all of those ARMA 4 0 models that the contiguous model had plus it will have a additional term corresponding to the twelfth lag that is the way I have formulated the non contiguous models.

So, this has the first 4 terms;  $X_{t-1}$ ,  $X_{t-2}$ ,  $X_{t-3}$ ,  $X_{t-4}$  and the twelfth term  $X_{t-12}$ , and that leads to a maximum likelihood value of 30.85 and there is not too much of a difference between these two maximum likelihood values. So you can go with either contiguous or non-contiguous. But typically, we keep in mind the Principle of Parsimony of Parameter Parsimony, which means that you go with the lower number or go with the model which has lower number of parameters and therefore, we chose ARMA 4 0 models, which also has a higher likelihood value in this particular case.

So, for the data generation then for the Cauvery river flows, for monthly time periods, we choose the ARMA 4 0 model.

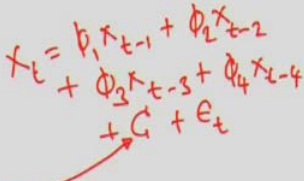

(Refer Slide Time: 44:19)

### Case study – 3 (Contd.)

- For this time series, the likelihood values for
  - contiguous model = 31.63
  - non-contiguous model = 30.85
- Hence contiguous ARMA(4,0) can be used.
- The parameters for the selected model are as follows

$\phi_1 = 0.2137$   
 $\phi_2 = 0.0398$   
 $\phi_3 = 0.054$   
 $\phi_4 = 0.1762$   
Constant = -0.0157

$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \phi_4 x_{t-4} + C + \epsilon_t$



Let see what, for this time series, I will just summarize this. The parameters of the selected model are 0.2137. This is you have 4 parameters phi 1, phi 2, phi 3 and phi 4. So these are the parameters and then you may have a constant, which is minus 0.0157. Now, this constant, if we write the model in another form where we typically write  $X_t$  similar to your regression, you may want to add a constant to that, so in this particular form what we do is  $\phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + \phi_4 X_{t-4} + \text{constant} + \epsilon_t$ .


Now, this is what we write and this is a constant term. So, these are the 4 parameters phi 1 phi 2 phi 3 phi 4 and you know the data and therefore, all these terms are known and therefore, you can apply this to your calibration time period and get the residual associated with that.

(Refer Slide Time: 45:40)

### Case study – 3 (Contd.) Forecasting Models

Contiguous models:

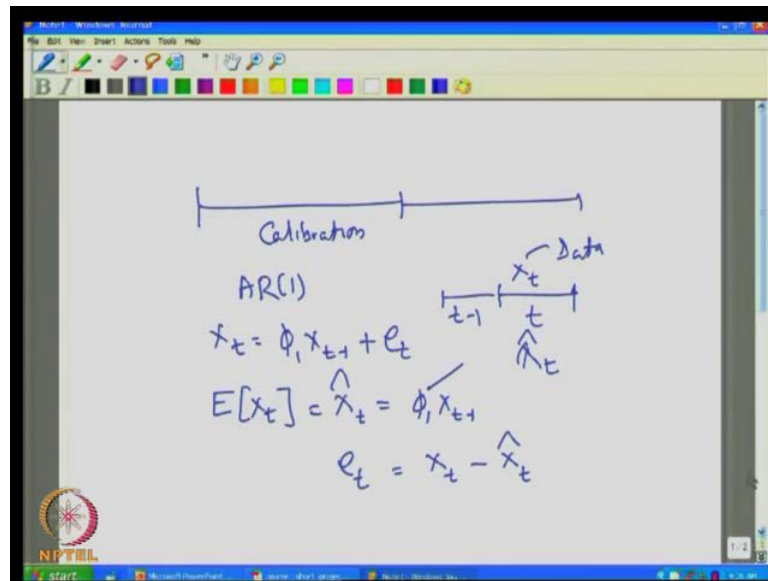
Sl. No	Model	Mean square error values
1	ARMA(1,0)	0.97
2	ARMA(2,0)	1.92
3	ARMA(3,0)	2.87
4	ARMA(4,0)	3.82
5	ARMA(5,0)	4.78
6	ARMA(6,0)	5.74
7	ARMA(1,1)	2.49
8	ARMA(1,2)	2.17
9	ARMA(2,1)	3.44
10	ARMA(2,2)	4.29
11	ARMA(3,1)	1.89



Now, we will look at deciding, which is a good model for forecasting, the models that we just discussed are for long term synthetic generation of data, so you can use this particular model and then generate the data over a long period of time, let say 50 years 100 years 150 years and so on. You can use this particular model. Remember that this is a time series; it has no relevance for the individual months, so this simply generates time series of the flows without regard to the particular month in which they belong. In fact, we have done this for the standardized series actually. It generates the standardized series. From the standardized series you should be able to get back to the flows.

Now, we will see if we are interested in forecasting; that is monthly forecast models for the stream flows, then what we do is we compute the mean square errors rather than the likelihood values. So, we choose the model, apply the model to the calibration time period and then we use let us say this one I have discussed earlier. But, for the completeness sake, let us go through this again and see how we compute the mean square error for this.

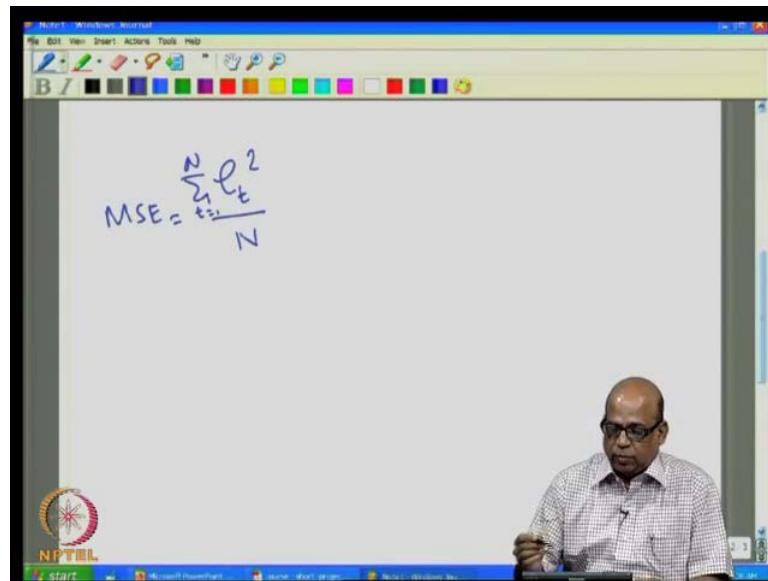
(Refer Slide Time: 47:13)



This is a calibration period. So, let say that you are talking about AR 1 model, so we are writing this  $X_t$  is equal to  $\phi_1 X_{t-1} + e_t$  and when we are doing the forecasting. We write this as we take the expected value of  $X_t$  and this we indicated as  $\hat{X}_t$  is equal to  $\phi_1 X_{t-1}$  and the expected value of this will be 0.

So, if we want to apply AR 1 model for forecasting, we look the calibration, we look up the calibration period, we would have estimated  $\phi_1$  and then apply this for the forecasting period. Let say you apply this for time period  $t$  starting with time period  $t-1$ , you also have the data here and you have the forecasting here, so this is  $\hat{X}_t$  as obtained from this. So, this would have been estimated now.  $X_t - \hat{X}_t$  will give you the residual and from this we obtained the residual variance. So, essentially then what we are doing is the errors that we get from here. So, error  $e_t$  will be equal to  $X_t - \hat{X}_t$ . This is the error then we get the mean square error so from this we get the mean square error.

(Refer Slide Time: 49:13)



So, we write mean square error as  $e_t$  square by  $N$ ,  $t$  is equal to 1 to  $N$ . This is how we get the mean square error. So, associated with each of the models, we know now how to compute the mean square error. Now, that is what we look at, which means for each of the models of the contiguous type, we again consider up to 6 0 and then up to 3 1, 1 1, 1 2 etcetera up to 3 1.

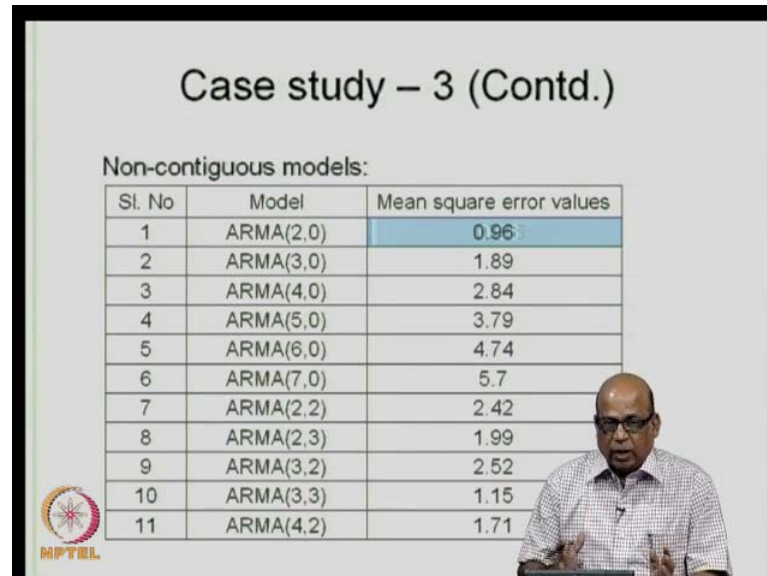
We compute the mean square error associated with this and choose that particular model, which results in the lowest mean square error. So, in this particular case, ARMA 1 0 which is AR 1 model comes up as the best model out of those considered here. Typically, when we are looking at normal processes like monthly stream flows, seasonal stream flows, etcetera, which are smoothen processes in some sense, and the best model is often the first model. It is that is only you are looking at  $X_{t-1}$  and then the dependence on  $X_{t-1}$  and that itself comes out to be the best model; however, if you had done this exercise for the rainfall, which we will do, subsequently in some other case studies.

If you do it for rainfall time series, this may not turn out to be the best model for forecasting. In fact in MA, in rainfall processes, often we may have to consider MA, MA terms, second order, third order, moving average terms. We may have to consider in the rainfall time series. So, this is for contiguous models. Let us now formulate the non



contiguous models following the same principle same method that we use for the non contiguous models for data generation process.

(Refer Slide Time: 51:18)



Case study – 3 (Contd.)

Non-contiguous models:

Sl. No	Model	Mean square error values
1	ARMA(2,0)	0.96
2	ARMA(3,0)	1.89
3	ARMA(4,0)	2.84
4	ARMA(5,0)	3.79
5	ARMA(6,0)	4.74
6	ARMA(7,0)	5.7
7	ARMA(2,2)	2.42
8	ARMA(2,3)	1.99
9	ARMA(3,2)	2.52
10	ARMA(3,3)	1.15
11	ARMA(4,2)	1.71

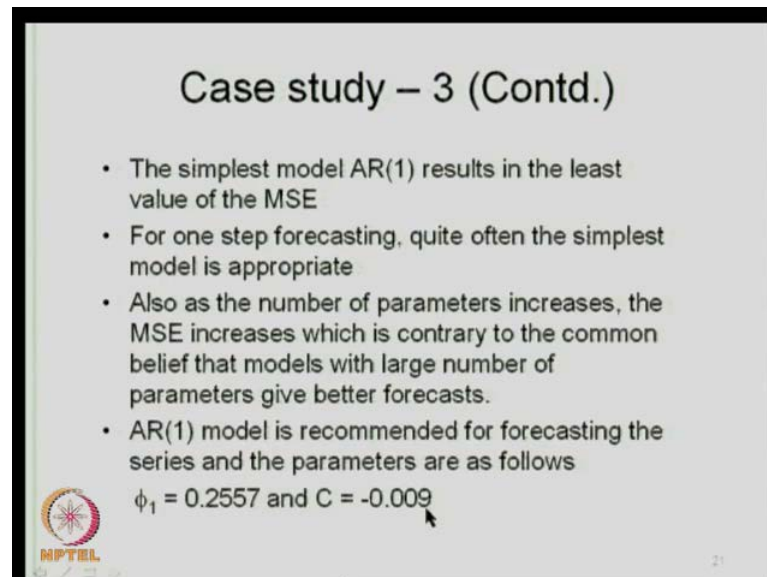
So, we form the non contiguous models what is ARMA 2 0 in this case? You will have  $X_{t-1}$  and  $X_{t-12}$ . Similarly, ARMA 7 0 will have  $X_{t-1}$ ,  $X_{t-2}$  etcetera  $X_{t-6}$  and  $X_{t-12}$ . So, the last term of any of these models will correspond to a lag of 12, both in terms of AR terms as well as in terms of the MA terms. So, MA when we have two terms, the first term will be  $e_{t-1}$ , the second term will be  $e_{t-12}$ , so that is how we construct the non-contiguous models.

So, these are the candidate models that we consider, and then we formulate the models, so you will know how many parameters. The parameters are estimated by going through any of the algorithms, typically we use the mat lab algorithm armax as I showed in one of the earlier classes.

Once you know the parameters, we apply these models to the calibration time period, which is a first 240 values, get the errors associated with that, and then calculate the mean square error that is a mean square error value that comes up here. So, these are the means square error values and again here you get ARMA 2 0, as the best model, which gives a mean square error of 0.96, corresponding to the earlier contiguous models that gave a mean square error of 0.97.

So, the mean square errors are more less the same in both the cases. So, you can choose either ARMA 2 0, which is a non-contiguous, which will have two parameters  $\phi_1$  and  $\phi_2$ ,  $X_{t-1}$  and  $X_{t-2}$ , or you may choose ARMA 1 0 model, which is  $\phi_1 X_{t-1}$ .

(Refer Slide Time: 53:30)



The slide is titled "Case study - 3 (Contd.)" and contains the following text:

- The simplest model AR(1) results in the least value of the MSE
- For one step forecasting, quite often the simplest model is appropriate
- Also as the number of parameters increases, the MSE increases which is contrary to the common belief that models with large number of parameters give better forecasts.
- AR(1) model is recommended for forecasting the series and the parameters are as follows

$\phi_1 = 0.2557$  and  $C = -0.009$

The slide also features the NPTEL logo in the bottom left corner and a small number "21" in the bottom right corner.

Now, this is how you select models for long term synthetic generation, as well as for short term forecasting models. For one time step ahead forecasting as I mentioned, often the simplest model is appropriate, because you are doing a time series forecasting and as long as the processes that we are considering are smoothen process.

For example, monthly stream flow, seasonal stream flows etcetera, then the simplest model will work well for the forecasting as a number of parameters increases, the MSE increases. In this particular case, as you can see here the MSE generally increasing as number of parameters increase; as 1.89, 2.84 etcetera for the non-contiguous models.

Similarly, for contiguous models 1.9 to 2.87 etcetera, it goes up to 6, then again here 1 1, you have two parameters, then it keeps on, not necessarily, but by enlarge it keeps increasing. So, in this case you have 4 parameters which is 1.89. So in general it appears to be the MSE value appears to increase in this particular case for, as a number of parameters increase. The AR 1 model is recommended for forecasting in this particular case and the parameters for this are  $\phi_1$  is equal to 0.2557 and C is equal to minus 0.009.

So, what did we do? We built two models now, by considering a large number of candidate models, we selected one model, which was ARMA 4 0 which is AR 4, for long term synthetic generation of the data, and we choose AR 1 the contiguous AR 1, which only considers  $\phi_1 X_{t-1}$ , as the models suitable for forecasting. So, this is how we select from among a number of candidate models which model is the best for that particular purpose. For long term synthetic generation we considered the maximum likelihood and for short term one time step ahead forecasting we considered the minimum mean square error, and then we choose these two models. the parameters have been fixed when we choose the particular models, the parameters have been already estimated.

Now, what we do is we apply these so chosen models on the remaining part of the data and get the residuals. The residuals that you so obtain must satisfy the conditions that we set forth earlier, namely that the residuals must have a zero mean. They should all be uncorrelated and they should be devoid of periodicity. So, these are the test that we conduct once we choose the particular model. So, the models that we choose must pass all these test only then we recommend that particular model for the purpose for which it is meant.

So, we will continue the discussion through other case studies. First complete this case study, in the next lecture, as soon as we start and see how we carry out the test on the residuals, how we construct the residuals and how we carry out the test on the residuals.

So, in today's lecture then we essentially continued with the case study number three that we started in the last class and we built these models, both for long term synthetic generation as well as for short term one time step ahead forecasting. We will continue the discussion in the next class.

**Thank you** for your attention.