

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

Lecture No. # 20
Case Studies – III

Good morning and welcome to this the lecture number 20 of the course on, “stochastic hydrology”. If you recall, in the last lecture, we were discussing the case study number three, which is on the monthly stream flows of the Cauvery River at the KRS reservoir. So, what we did in that case study up till now is we plotted the time series; we computed the correlogram and plotted the correlogram.

We also looked at the partial auto correlations and the power spectrum. Then we went on to build the ARMA type of models for both synthetic generations of the stream flows as well as for one time step ahead forecasting. What did we do in that? We considered several candidate models of the ARMA family. We constructed these candidate models based on the information that we obtained specifically form the correlogram and the partial auto correlation.

And then we considered several of these candidate models. Corresponding to each of the candidate models, we estimated the parameters by considering half the data. We had 40 years of data, which means 480 values we had for the monthly time series; we considered the first 240 values to calibrate the model. By calibration of the model, I mean estimation of the parameters. We use the armax function of the matlab to estimate these parameters. You can use this several of any of the other algorithms that are available in standard packages, computer packages, you can estimate the parameters associated with the ARMA models.

Remember, that we were working for this particular case study, we were working with the standardized flows, and then we estimated the likelihood values for each of the candidate models, picked that particular model, which gives the maximum likelihood values among those models that we were considering. The model that is chosen based on the maximum likelihood value is meant for long term synthetic generation of the data,

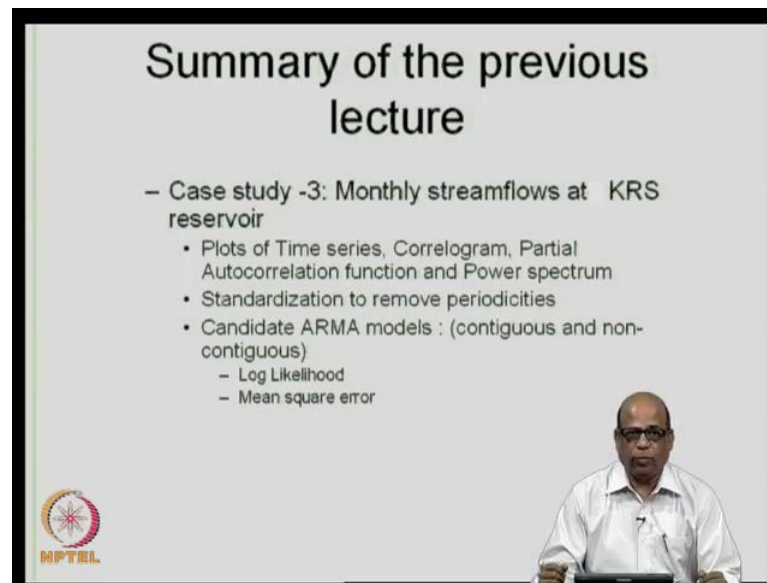
which are typically used for assessment of risk, assessment of reliability of the particular system, for fixing up the capacity of the reservoir and so on.

For all normal purposes of planning and operational decisions that we often come across in water recourses, we use the long term synthetic generation of the data. Typically, we generate for 100 years, 150 years, even 200 to 300 years and so on depending on the type of application that we are interested in. Then we also chose a particular model, based on the minimum mean square error. The mean square error criteria, we use for short term forecasting and typically we are looking at short term one time step ahead forecasting. We considered both contiguous as well as non-contiguous models and for both the contiguous as well as non-contiguous models.

We did this exercise and picked up models corresponding to contiguous models, based on the maximum likelihood criteria and the non-contiguous models based on the maximum likelihood criteria and then we compare the two. If the maximum likelihood criteria are quite close to each other, the maximum likelihood values, corresponding to the chosen model from the contiguous type of models and the non-contiguous type of models. If the maximum likelihood values are quite close to each other then we choose that particular model, which has minimum number of parameters. So, this is the Principle of Parsimony; if two models are satisfying more or less the same objective, then we choose that particular model which has lower number of parameters.



Similarly, for forecasting also, we did the exercise both for contiguous as well as non-contiguous models. In the case of forecasting, the simplest model, namely the ARMA 1 0 model for contiguous type of models and ARMA 2 0 models for the non-contiguous models, they surface up to be the best models, in terms of the mean square error values. In today's lecture, what we then do is we have chosen one model corresponding to the maximum likelihood criteria.

(Refer Slide Time: 05:44)



Summary of the previous lecture

- Case study -3: Monthly streamflows at KRS reservoir
 - Plots of Time series, Correlogram, Partial Autocorrelation function and Power spectrum
 - Standardization to remove periodicities
 - Candidate ARMA models : (contiguous and non-contiguous)
 - Log Likelihood
 - Mean square error

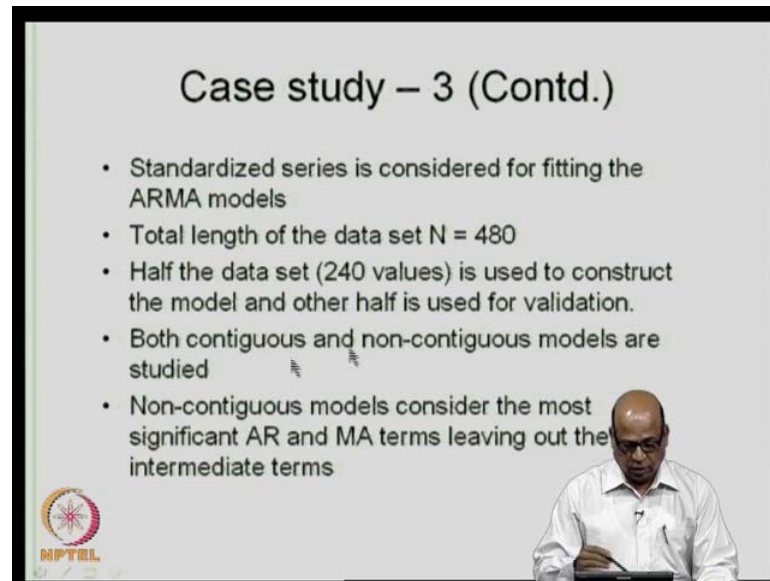
This model turned out to be ARMA 4 0 model for the Cauvery monthly stream flows, and we also choose one model corresponding to the minimum mean square error. Now, this model turned out to be the ARMA 1 0 model for the contiguous type of models that headed the minimum mean square error.

Now, once we have identified which model we would like to use for the specific purpose, we need to validate this model. Remember, all of this exercise we did on the first half of the data. We had 480 values, all of this exercise we did only on the first 240 values, except of course, first part where we identify the periodicities etcetera, where we plot the time series and spectral density and so on. That was done on the entire time series, but once we come to the identification of the model, we focus only on the first half of the data.

Now, once we choose the model, we apply this model to the remaining half of the data. In this particular case, to the remaining 240 values, and obtain the residuals and then we do the analysis on the residuals, to examine whether the residual series that we so obtain by applying the model chosen to the validation data set, in this particular case it is 240 values, the remaining 240 values, obtain the residual series. Then we do the test on the residual series for validation. So, typically the validation test will consist of the test to examine whether the residual series has 0 mean or the mean is insignificant.

Then whether the residual are devoid of any periodicities, whether there are any periodicities present in the residual series and whether the residual series that we so obtain is uncorrelated or it constitutes a white noise. So, these are the three types of test and I have discussed earlier how we carry out the test. Today, we will apply these testes to the particular case study that we are considering.

(Refer Slide Time: 08:26)



The slide is titled "Case study – 3 (Contd.)" and contains a list of five bullet points. In the bottom left corner, there is a circular logo with a star-like pattern and the text "NPTEL" below it. In the bottom right corner, there is a small inset image of a man in a white shirt sitting at a desk, looking at a laptop.

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

Just to recapitulate what models that we chose for the particular case study, we will just go through this. These are all the details that I have already discussed; half the data set is used to construct the model and the other half is used for validation. So, this is what I just mentioned.



(Refer Slide Time: 08:41)

Case study – 3 (Contd.)

Contiguous models:

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

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



This is the summary results for the contiguous model, this is the way we calculate the likelihood values. This is the variance of the residuals when we are applying for the calibration period and that is how we get the likelihood values. You choose that particular model, which gives you the maximum likelihood value. In this particular case it is ARMA 4 0. Similarly, for the non-contiguous models this is what I have discussed in the last lecture, so let me go a bit fast.



(Refer Slide Time: 09:08)

Case study – 3 (Contd.)

Non-contiguous models*:

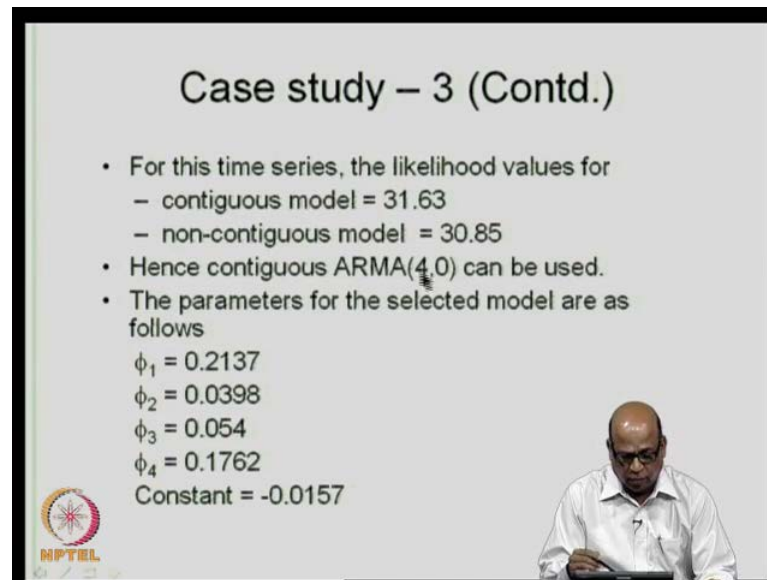
Sl. No	Model	Likelihood values
1	ARMA(2,0)	28.52
2	ARMA(3,0)	28.12
3	ARMA(4,0)	28.21
4	ARMA(5,0)	30.85
5	ARMA(6,0)	29.84
6	ARMA(7,0)	29.12
7	ARMA(2,2)	29.81
8	ARMA(2,3)	28.82
9	ARMA(3,2)	28.48
10	ARMA(3,3)	28.06
11	ARMA(4,2)	28.65

* The last AR and MA terms correspond to the 17



So, ARMA 5 0 turns out to be the best model in terms of a likelihood values. Now, in the contiguous models, if you recall, the last term we included as the twelfth lag, so the last AR and MA terms correspond to the twelfth lag. This I have explained in the previous lecture.

(Refer Slide Time: 09:35)



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

NPTEL



Now, what we will do is...Another minor point here that when we did all this exercise, the parameters of these models would have been estimated. So, finally, for the chosen model, we give you the parameters like phi 1, phi 2, phi 3, and phi 4. This ARMA 4 0 model, it is also typically written as AR 4. When MA model MA term is absent, we simply write it as AR 4 model and this is a constant term.

(Refer Slide Time: 10:12)

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, similarly, for forecasting model, we chose the minimum mean square, the model that gives you the minimum mean square error. In this particular case, it was ARMA 1 0 or AR 1 model.

(Refer Slide Time: 10:28)

Case study – 3 (Contd.)

Non-contiguous models:

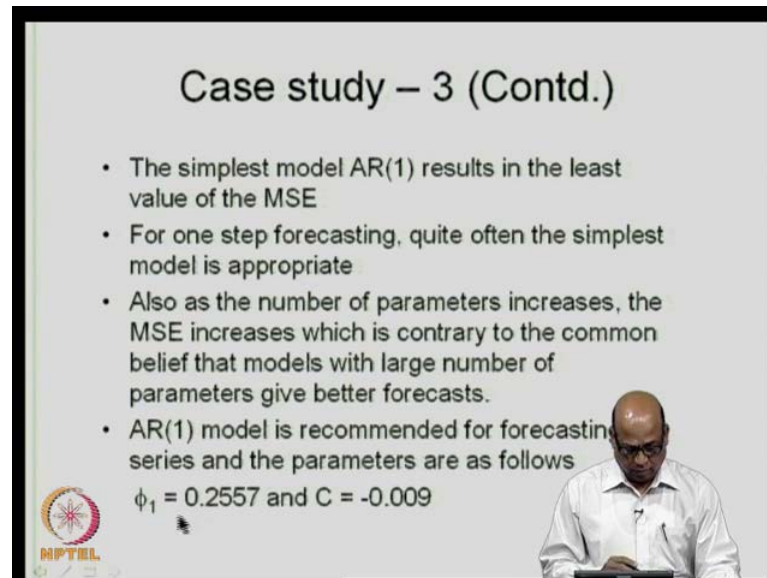
Sl. No	Model	Mean square error values
1	ARMA(2,0)	0.96
2	ARMA(3,0)	1.89
3	ARMA(4,0)	2.84
4	ARMA(5,0)	3.79
5	ARMA(6,0)	4.74
6	ARMA(7,0)	5.7
7	ARMA(2,2)	2.42
8	ARMA(2,3)	1.99
9	ARMA(3,2)	2.52
10	ARMA(3,3)	1.15
11	ARMA(4,2)	1.71



And for the non-contiguous model, it turns out to be ARMA 2 0 model. Again, the second term, in this particular case, will correspond to the twelfth lag. So, ARMA 2 0 model for the non-contiguous model will be X_t is equal to $\phi_1 X_{t-1}$ plus $\phi_2 X_{t-12}$, the twelfth lag is considered here for forecasting model.

Now, both of this yields the same, almost the same mean square error value, so you can use one of the two and then as I said, we use that particular model, which has lower number of parameters. In this particular case, it will be ARMA 1 0 model.

(Refer Slide Time: 11:08)



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 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 photograph of a man in a white shirt and glasses in the bottom right corner, appearing to be presenting.


So, these are some of the details. So, for the ARMA 1 0 model, we again give the parameters phi 1 as 0.2557 and C is minus 0.009. Remember, that we are applying these models to the standardized series and that is why you get this kind of magnitudes.

(Refer Slide Time: 11:25)

Case study – 3 (Contd.)

- Validation tests on the residual series
 - Significance of residual mean
 - Significance of periodicities
 - Cumulative periodogram test or Bartlett's test
 - White noise test
 - Whittle's test
 - Portmanteau test
- Residuals,
$$e_t = X_t - \left(\sum_{j=1}^{m_1} \phi_j X_{t-j} + \sum_{j=1}^{m_2} \theta_j e_{t-j} + C \right)$$

Residual Data Simulated from the model



Now, we will go to the validation test. So, as I mentioned, we have applied, we have calibrated the model based on the first half of the data. Now, we will apply these models; compete with their parameters to the second half of the model. What do I mean by that? Let say 240 values, first 240 values you have used to calibrate the model and therefore, the parameters have been estimated. The models have been fixed and this model we start applying for the next 240 values.

So, we generate as I discussed in the last, perhaps lecture number 18, we generate the data for the next 240 values and then compute the residuals. How do I compute the residuals? This is e_t is equal to X_t . We have the data; minus this is how the model has been simulated. You know the ϕ_j 's, you know the θ_j 's and similar to this we calculate the estimated value for X_t based on the number of parameters that you are using. For AR 4 model we would have four models, so 4 AR parameters, so it was from j is equal to 1 to 4. We do not have any m parameters, so this summation we do not consider plus there is a constant term, which we have estimated earlier.

So, this is from the simulation from the model. So, this is a simulated data, this is an actual available data and that is how we calculate the residual. So, we now form the residual series by using this expression. On the residual series, the validation tests are

done. First we examine whether the residual series we so obtain, has a zero mean or the mean is not significant. Then we also examine whether there are any significant periodicities that are present in the residual series. For the significant periodicities we also do the Bartlett's test. This Bartlett's test is also for significance of periodicities. Then we examine whether the residual series in fact is uncorrelated, for that we generally do two tests; Whittles test and Portmanteau test. Remember, that I have mentioned Whittles test is preferred over the Portmanteau test.

(Refer Slide Time: 14:07)



Case study – 3 (Contd.)

Significance of residual mean:

Sl. No	Model	$\eta(e)$	$t_{0.95}(239)$
1	ARMA(1,0)	0.002	1.645
2	ARMA(2,0)	0.006	1.645
3	ARMA(3,0)	0.008	1.645
4	ARMA(4,0)	0.025	1.645
5	ARMA(5,0)	0.023	1.645
6	ARMA(6,0)	0.018	1.645
7	ARMA(1,1)	0.033	1.645
8	ARMA(1,2)	0.104	1.645
9	ARMA(2,1)	0.106	1.645
10	ARMA(2,2)	0.028	1.645

$$\eta(e) = \frac{N^{-1/2} \bar{e}}{\hat{\rho}^{1/2}}$$

$\eta(e) \leq t(0.95, 240-1)$:
All models pass the test

For the case study we will see now, what kind of test that we do. First, is the significance of residual mean and if you recall, we formulate the statistic eta as N to the power half e bar y rho cap to the power half; the rho cap is the residual variance.

Remember, here the moment you compute the residual series first obtain the variance as well as the mean of the residuals, simply obtain the mean and the variance. These will be useful for any of the statistical test that we are doing on the residual series. So, e bar is a residual mean and rho cap is the residual variance, N is the number of data. Now, in this particular case N will be 240, and e bar is the residual mean and rho cap is the residual variance. Therefore, you can get eta e, which is the statistic, these are the values. Although we have chosen a particular model, you know typically you can do these tests only for the particular chosen model, but just for completeness sake, I am giving you for

all the candidate models that we have considered. These are the data for all the candidate models, so these are the type of statistic value that you get.

Now, the model passes the test for significance of residual mean by passing I mean that the residual mean is insignificant or we can approximate the residual mean to be 0 that is what I mean by the model test or passes the test. If the eta value that is so obtained is less than, let say you are fixing the 95 percent of confidence limit, so $t_{0.95, 240}$ minus 1, this 240 is the N that we are considering, So, $N - 1$, $0.95 N - 1$, this can be either 0.95 or 0.99, typically we consider 95 percent confidence level.

So, as you can see, this value remains constant, because it only depends on the confidence limit and the N value; both of which are constant, so this is 1.645. If the statistic that you calculate is less than this particular value, then the particular model passes the test, as you can see all of these values are less than the associated t value and therefore, all the models pass the test. Although, we are interested in this particular model ARMA 4 0; ARMA 4 0 is the model that we have chosen that passes the test but, typically you can do this test for all the candidate models just as an exercise.

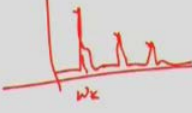


Now, we will go to significance of periodicity. So, this is a order in which we carry out the test. Typically, first you test for the residual mean, make sure that the residual mean it is non-significant in the sense that it can be approximated to be 0. Once you are satisfied that the residual mean is 0 or can be approximated to be 0, then you go to significance of periodicity. Why we did this for all the models? Let say that the chosen model ARMA 4 0 did not pass the test, and then you go to your again likelihood values and see which is the next best model. Then look for that particular model, let say that after ARMA 4 0 you had, let us look at which is other model that you had for.

See, this was 30.85, the next value seems to be 29.81, which is ARMA 2 2. Let us say that this model did not pass one of the tests, then you go to the next model, next available model in terms of the likelihood values and see whether that model test passes the test. That is why we calculate this t statistic as well as the likelihood values for all the candidate models. So, in this particular case, ARMA 4 0 model passes the test and therefore, we progress further.

(Refer Slide Time: 18:55)

Case study – 3 (Contd.)

Significance of periodicities:

$$\eta(e) = \frac{\gamma_k^2 (N-2)}{4\hat{\rho}_1}$$

$$\gamma_k^2 = \alpha_k^2 + \beta_k^2$$
$$\hat{\rho}_1 = \frac{1}{N} \left[\sum_{t=1}^N \{e_t - \hat{\alpha} \cos(\omega_k t) - \hat{\beta} \sin(\omega_k t)\}^2 \right]$$
$$\alpha_k = \frac{2}{N} \sum_{t=1}^n e_t \cos(\omega_k t)$$


Once, we are satisfied that the residual mean passes the test for its significance, in fact, the lack of significance then we go to the next test, which is significance of periodicity. Again, we formulate the statistic and this statistic is gamma k square N minus 2 by 4 rho 1 cap; rho 1 cap is now not the residual variance in this particular case. Now, rho 1 cap is the corresponding to the periodicity for which we are testing. We compute the rho 1 cap like this.

Remember, omega k is the omega value for the particular. Let say we are testing for periodicity number one, in our case it is periodicity of 12 months. Then omega k is that particular value of the periodicity. To make it clear, let say that we had our original spectrum like this and this is the omega k, corresponding to the first periodicity, omega k corresponding to the second periodicity etcetera, so this test that we are just discussing now, is carried out for one periodicity at a time. So, you pick up the first periodicity examine whether that periodicity is still present in the residual series. Then go to the next periodicity, examine whether that is still present in the residual series and so on.

So, like this one after the other you carry out the periodicities significance **I am sorry** one after the other, you carry out the test for the periodicities. So, this is for that particular k, so you can calculate gamma k, where alpha k is given by this and beta k is given by this; omega k is known. Therefore, e t is the residual series, t is equal to 1 to N and N is in this particular case 240.

Alpha k and beta k are obtained from here, so this is alpha k corresponding to that particular k and similarly, this is beta k corresponding to that particular k. Remember here, t is the time period, so this is going from 1 to N; small n and capital N are analogues, they are the same and in this case it is 240, so you can get gamma k square. So, once you know gamma k square and rho one cap is estimated like this and you can get eta because you know N.



(Refer Slide Time: 21:40)

Case study – 3 (Contd.)

$$\beta_k = \frac{2}{N} \sum_{t=1}^n e_t \sin(\omega_k t)$$

$2\pi/\omega_k$ is the periodicity for which test is being carried out.

$\eta(e) \leq F_{\alpha}(2, N-2)$ – Model passes the test


So, eta is obtained like this; 2π by ω_k is the periodicity for which the test is being carried out. Let us say you are testing for 12 months, then you can obtain the associated ω_k . This is typically obtain from your spectral analysis. Now, if this statistic value that you calculate is less than the F alpha 2 N minus 2, then it passes the test. Two is the degrees of freedom and in N minus 2, N is the number of data in this case and is again 240, alpha is the significance level. Typically, we use 95 percent value corresponding to this.

(Refer Slide Time: 22:16)

Case study – 3 (Contd.)

Significance of periodicities:

Sl. No	Model	η Value for the periodicity				$F_{0.95}(2,238)$
		1 st	2 nd	3 rd	4 th	
1	ARMA(1,0)	0.527	1.092	0.364	0.065	3.00
2	ARMA(2,0)	1.027	2.458	0.813	0.129	3.00
3	ARMA(3,0)	1.705	4.319	1.096	0.160	3.00
4	ARMA(4,0)	3.228	6.078	0.948	0.277	3.00
5	ARMA(5,0)	3.769	7.805	1.149	0.345	3.00
6	ARMA(6,0)	4.19	10.13	1.262	0.441	3.00
7	ARMA(1,1)	4.737	10.09	2.668	0.392	3.00
8	ARMA(1,2)	6.786	10.67	2.621	0.372	3.00
9	ARMA(2,1)	7.704	12.12	2.976	0.422	3.00
10	ARMA(2,2)	6.857	13.22	3.718	0.597	3.00



We do this significance of periodicity test on all the models, all the candidate models. The first periodicity is 12 months, second is 6, third is 4 and fourth is three months. Associated with those particular omega k values, we calculate these statistic values. Remember, again that these are all done for the residual series, so on the residual series we calculate these eta values and the associated F statistic value, which is a critical value is 2 comma 238, n is 240, so this becomes 238 and this is the 95 percent significance value.

These values of $F_{0.95}$ can be picked from any standard tables of F distribution. Any of the standard text books will provide this F value associated with the degrees of freedom and N minus 2, and that comes out to be 3. As you can see all the models pass the test for all the periodicities, indicating that none of these periodicities is significant in the residual series. We have done the test of periodicities one at a time, so first we considered the first periodicity, which is corresponding to 12 months, next 6 months, 4 months and 3 months. We conclude that all the model pass the test.

But as I discussed in the one of the earlier lectures, may be lecture number 16, 17 or 18 that instead of doing the test on periodicities, one at a time, we would prefer to test all the periodicities at a time. That means we have the residual series available with us. Now, with this residual series can we examine whether there is any significant periodicities present at all in the residual series? Now, this is best achieved by the

Bartlett's test or the Cumulative periodogram test. Now, the cumulative periodogram test which I will discuss now will apply to case study number three, which is Cauvery River flows.

(Refer Slide Time: 24:49)


Case study – 3 (Contd.)


Significance of periodicities by Bartlett's test :
(Cumulative periodogram test)

$$\gamma_k^2 = \left\{ \frac{2}{N} \sum_{t=1}^N e_t \cos(\omega_k t) \right\}^2 + \left\{ \frac{2}{N} \sum_{t=1}^N e_t \sin(\omega_k t) \right\}^2$$

$k = 1, 2, \dots, N/2$

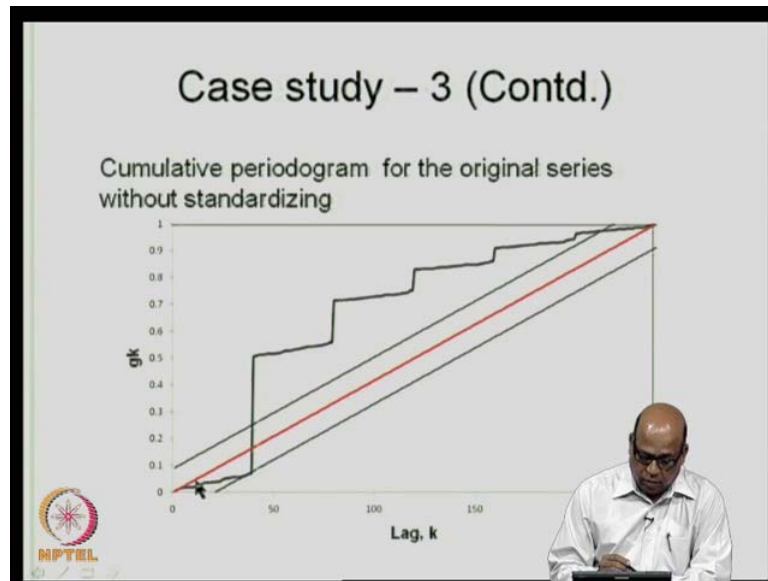
$$g_k = \frac{\sum_{j=1}^k \gamma_j^2}{\sum_{k=1}^{N/2} \gamma_k^2}$$





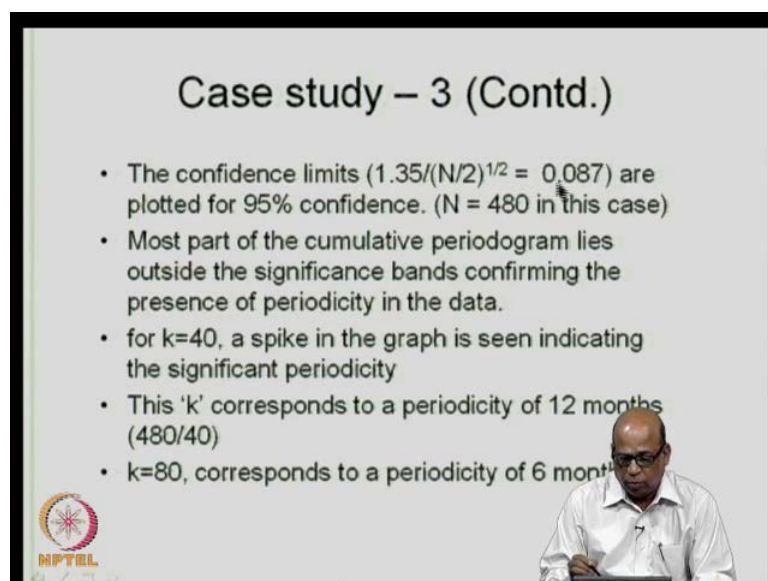
If you recall in the Bartlett's test what we do again we formulate gamma k square, which is two by N e t cos omega k t plus 2 by N e t sin summation, sin omega t whole square. So, this is gamma k square. this we do for k is equal to 1 to N etcetera, N by 2 then we calculate the cumulative periodogram, g k is equal to, by considering g k, g k is equal to gamma j square, j is equal to 1 to k, let say k is equal to 2. This goes from 1 to 2, divided by gamma k square; k is equal to 1 to N by 2, the whole range, so this is normalized, gamma j gamma j square. So, g k is what we require for constructing the periodogram, so g k versus k. The plot of g k versus k is called as the cumulative periodogram, so you know how to calculate gamma k for a given k. You know now how to calculate g k.

(Refer Slide Time: 26:07)



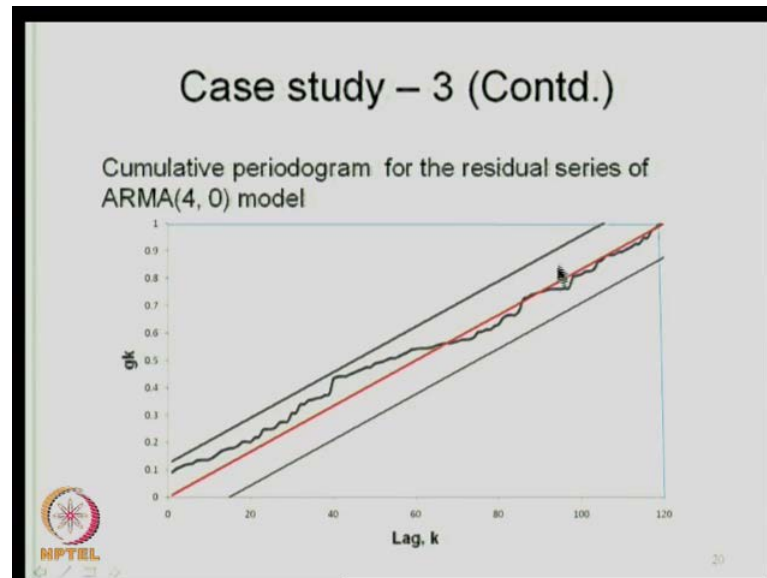
From this now, let say that we calculate the g_k for the original time series; that means the original, all the 480 values, monthly values, we compute g_k and then plot for lag k versus g_k . Typically, we go up to N by 2; lag k we go up to N by 2, in this particular case it is 480 by 2 which is 240. Then what we do is the 0.00 here and N by 2 and 1, so g_k the maximum value is 1, so N by 2 in this particular case it is 240, so 240 and 1. We joint that those two particular point by a line around this line we form the confidence bands.

(Refer Slide Time: 27:11)



Now, this confidence band if you recall is 1.35 by the root N by 2 . This 1.35 corresponds to 95 percent value, so 95 percent confidence value is 1.35 by root N by 2 . This comes out to be 0.087 .

(Refer Slide Time: 27:28)



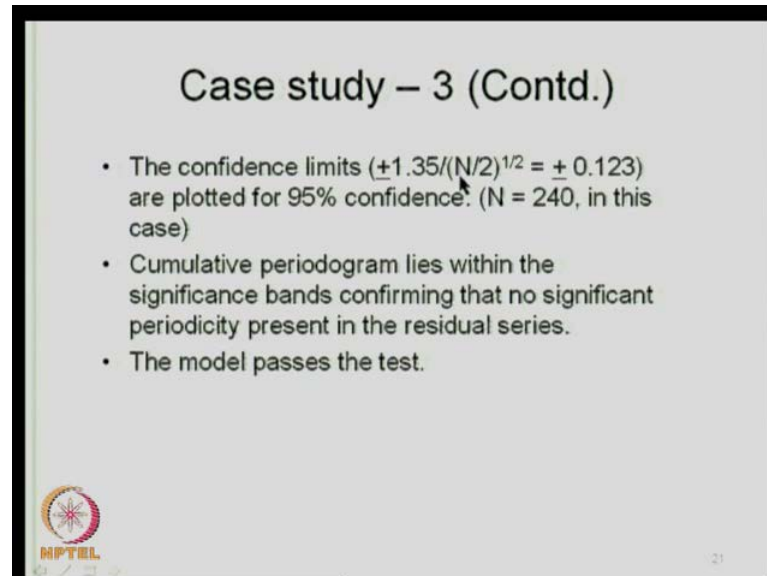
Now, this 0.087 , we draw a line at each of these points and form a band. Now, this gives you the band of confidence for this particular test. On either side, we draw that line and as you can see for the original monthly time series where we had 480 values. If you conduct this test for the original time series, you see that most of the values lie above this band; beyond this band, indicating that there are significant periodicities present. In fact, you see a first spike corresponding to a value of 40, so 480 by 40 that corresponds to a periodicity of 12 months.

Similarly, you can capture periodicities associated with each of the spikes. So, this is cumulative periodogram. Now, the same test we carry out on the residual. Remember, this test that I just mention now, this cumulative periodogram, is not a test on the residuals. I am just showing this to indicate that if you have periodicities this is how the cumulative periodogram will appear, whereas, you do not want any periodicities to be present in the residual series. So, we do the test on the residual series and look at for the ARMA 4 0 model how does the cumulative periodogram look like.

Now, we now consider 240 values, because we are calculating the residual on the remaining half, which is one half of the data, which is 240. So, the number of values that


we consider for lag k is N by 2, which is 240 by 2, which is 120. So, from 120 to 1 and 0, you draw this line and around this line you form the confidence limit.

(Refer Slide Time: 29:43)



Case study – 3 (Contd.)

- The confidence limits $(\pm 1.35/(N/2)^{1/2} = \pm 0.123)$ are plotted for 95% confidence. (N = 240, in this case)
- Cumulative periodogram lies within the significance bands confirming that no significant periodicity present in the residual series.
- The model passes the test.

 21

How do we get the confidence limits? Now, 1.35 by root N by 2 and N in this case is 240. Therefore, you get a band of plus minus 0.123; by plus minus, I mean it is on either side either side of this particular line red line that is drawn here. Then we form the cumulative periodogram. How do we get the cumulative periodogram? We apply this now on the residual series. So, you get g_k on the residual series and g_k versus k on the residual series, this is the plot of the cumulative periodogram, this dark black line shows the cumulative periodogram.

If the entire cumulative periodogram lies within this confidence band then there are no significant periodicities present in the data. As you can see, the cumulative periodogram in this particular case of the residual series lies completely within the confidence bands. Therefore, there are no significant periodicities present in the data and therefore, the residual series corresponding to ARMA 4 0 model passes the test for periodicities. When I say it passes the test for periodicities, it means that there are no significant periodicities present in the residual series.

We conducted two tests for the testing, whether there are significant periodicities present in the residual series or not. Among the two tests, this particular test, which is called as the Bartlett's test or the Cumulative periodogram test is preferred, because you can do

the test in one go for all the periodicities, whereas, in the earlier case, we did one periodicity at a time. You do the test for the first periodicity, remove that periodicity from the data then do for the next one, remove that periodicity if it is significant and then go to the next one and so on.

So, you did one at a time in the earlier case, whereas, in the cumulative periodogram test you simply plot the cumulative periodogram, you draw the confidence limits, the band of confidence limits, and then for all the periodicities at single time, you can examine whether the model passes the test or not. So, this test is preferred over the earlier test. We now examine for the significance of mean, we satisfied ourselves that the model that is chosen ARMA 4 0, does not have significant mean. The residuals resulting from that residual series does not have significant mean, which means that the mean is zero mean and can be approximated to 0.

We also satisfied ourselves that there are no significant periodicities present in the data present, in the residual series, by doing the cumulative periodogram test and the test that I just discussed earlier. Now, we have to look at another important assumption that we made in building the model that the residual series is uncorrelated. This is also called as the test for white noise. That means the residual series that we obtain from applying the model should constitute a white noise, in terms of absence of any correlations in the residual series.

So, this is what we do by two tests; Whittles test and Portmanteau test, both of which require the covariance matrix; covariance matrix of the residual series. So, once we constitute the residual series, we form the covariance matrix of the residual series. Now, based on the covariance matrix, we define a statistic for the Whittles test and a different statistic for the Portmanteau test and then based on this statistic, we formulate based on the covariance matrix, we examine whether the series is in fact uncorrelated.

(Refer Slide Time: 34:15)


Whittle's test for white noise

White noise test (Whittle's test):

- This test is carried out to test the absence of correlation in the series.
- The covariance r_k at lag k of the error series $\{e_j\}$

$$r_k = \frac{1}{N-k} \sum_{j=k+1}^N e_j e_{j-k} \quad k = 0, 1, 2, \dots, k_{\max}$$

- The value of k_{\max} is normally chosen as $0.15N$

 Ref. Kashyap R.L. and Ramachandra Rao.A, "Dynamic stochastic models from empirical data". Academic press, New York, 1976 23

So, let us see what the Whittles test is. This I have discussed earlier, but just to recapitulate essentially what we do is we formulate the covariance rho k; the covariance r k at lag k of the error series e t; error or residual series is the same. So, this we go from k is equal to 0, 1, 2 etcetera till k max. We typically consider the maximum k as 0.15 N; N is the number of data points. In this particular case it will be 240. So, the covariance is taken as e j minus e j minus k. How did we define covariance? It is the expected value of X t minus mu into expected value of X t minus k minus mu. So, this was your r k for the original time series X t, but since we have verified that the mean of the residual series is in fact 0, so I write the covariance as r k is equal to 1 by N minus k e j into e j minus k, because the residual mean is 0 and j goes from k plus 1 to N.


So, k is, let say you are calculating for two, k is equal to 2, you will have N minus 2 here and N is 240 and this goes from 3 to N which is 240. Typically, we calculate this for up to k max and k max is typically taken as 0.15 N. I am going through the text book Kashyap and Rao, dynamic stochastic models from empirical data. Aside, I would like to mention that Professor Kashyap is an electrical engineer and has contributed significantly to stochastic models, to theory of stochastic models and Professor Rama Chandra Rao is a civil engineer. He is a hydrologist both of them were at Perdue University and have contributed significantly through this particular classical text book called Dynamic Stochastic Models from Empirical Data. There are large numbers of examples from the electrical engineering and there are also significant numbers of

examples from hydrology in this particular text book. I would encourage all of you to have a look at this particular text book.

(Refer Slide Time: 36:52)

Whittle's test for white noise

- The covariance matrix is

$$\Gamma_{n1} = \begin{bmatrix} r_0 & r_1 & r_2 & \cdot & \cdot & r_{k_{\max}} \\ r_1 & r_0 & r_1 & \cdot & \cdot & r_{k_{\max}-1} \\ r_2 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{k_{\max}} & r_{k_{\max}-1} & \cdot & \cdot & \cdot & r_0 \end{bmatrix}_{k_{\max} \times k_{\max}}$$


23

Now, you have formulated the covariance. Now, once you have the covariance matrix, this is gamma n 1, you have the entire covariance matrix, k max into k max size. We formulate gamma n 1 minus 1 by taking out the last row and the last column. So, you have gamma n 1 as complete covariance matrix of size k max by k max, then you take out the last row and the last column to formulate gamma n 1 minus 1.

(Refer Slide Time: 37:28)

Whittle's test for white noise


- A statistic $\eta(e)$ is defined as

$$\eta(e) = \frac{N}{n1-1} \left(\frac{\hat{\rho}_0}{\hat{\rho}_1} - 1 \right) \quad n1 = k_{\max}$$

Where $\hat{\rho}_0$ is the lag zero correlation =1, and

$$\hat{\rho}_1 = \frac{\det \Gamma_{n1}}{\det \Gamma_{n1-1}}$$

The matrix Γ_{n1-1} is constructed by eliminating the last row and the last column from the Γ_{n1} matrix.



24

Then we define the statistic η as N divided by $n - 1$, where n is the k max into $\rho_{\text{naught cap}}$ by $\rho_{\text{one cap}} - 1$, where $\rho_{\text{naught cap}}$ is the lag 0 correlation which is one. And $\rho_{\text{one cap}}$ is determinant γ_{n-1} ; this is γ_{n-1} , you take the determinant of this by determinant γ_n . As I said γ_{n-1} is formulated by taking out the last row and the last column. So, the matrix γ_{n-1} is constructed by eliminating the last row and the last column from the γ_n matrix, so you can get $\rho_{\text{one cap}}$. Once you get $\rho_{\text{one cap}}$, you know $\rho_{\text{naught cap}}$ which is one and you know $n - 1$, which is k max typically 0.15 into N , you know N which is 240 in this particular case and therefore, you will get η .

(Refer Slide Time: 38:41)

Whittle's test for white noise

- The statistic $\eta(e)$ is approximately distributed as $F_{\alpha}(n1, N-n1)$, where α is the significance level at which the test is being carried out.
- If the value of $\eta(e) \leq F_{\alpha}(n1, N-n1)$, then the residual series is uncorrelated.

Once you get η , you compare this η with $F_{\alpha}(n - 1, N - n - 1)$, $n - 1$ in this particular case, it is k max. That difference is the degrees of freedom and $N - n - 1$. If the η you have calculated is less than this critical value of F and α is typically chosen as 0.05 or something, then it passes the test.

(Refer Slide Time: 39:14)

Case study – 3 (Contd.)

Whittle's test - white noise $\eta(\epsilon) = \frac{N}{n1-1} \left(\frac{\hat{\rho}_0}{\hat{\rho}_1} - 1 \right)$

F_{0.95}(2, 239)

	n1 = 73	n1 = 49	n1 = 25	
Model \ η	η	η	η	
ARMA(1,0)	0.642	0.917	0.891	
ARMA(2,0)	0.628	0.898	0.861	
ARMA(3,0)	0.606	0.868	0.791	
ARMA(4,0)	0.528	0.743	0.516	
ARMA(5,0)	0.526	0.739	0.516	
ARMA(6,0)	0.522	0.728	0.493	
ARMA(1,1)	0.595	0.854	0.755	
ARMA(1,2)	0.851	1.256	1.581	model fails
ARMA(2,1)	0.851	1.256	1.581	
ARMA(2,2)	0.589	0.845	0.737	

NPTEL

So, for the case study now, let us carry out the Whittles test. So, you know how to formulate n now. I have not given all the data details of the covariance matrix and so on. I leave it as an exercise for you. I will give you the final results here. So, the final results for the Whittles test, we do it for all the candidate models, although we can do it only for the particular model that we have chosen namely ARMA 4 0. you could have done only for ARMA 4 0, but not necessarily only as an exercise, I mentioned you do it for all the candidate models, so that if the particular model that you have chosen does not pass the test then you can go to another model.

Typically, we do this test for all the candidate models that you have chosen. Now, we consider n 1 variously, as 73, 49, and 25 and so on. So, you can do it for several n 1. Typically, you can chase it up to 0.15 N. In our case N was 240; 240 we have considered to be half the data, so you could have gone up to 0.15 N. So, you consider for various test just to see where the model fails. So, as you can see these are the eta values that you obtain for different n 1 value. So, let us say n 1 is equal to 73, I take like this then corresponding to that particular n 1, and you will have F 0.95 and 239. 239 being N minus n 1, **I am sorry**, there is a mistake here, this is not F 0.95 239.

So, this is what you get from the table. This is actually what I have written here that is F alpha n 1 N minus n 1. So, this is F alpha n 1 comma N minus n 1. So, these are the values that you get; 1.29, 1.39 and 1.52, if these values that you are getting of the

statistic are less than the particular F alpha values then the model passes the test. As you can see most of the models pass the test, but when you come to ARMA 1 2 model, here 1.581 which is more than 1.52, for both ARMA 1 2 as well as ARMA 2 1 model, these models fail the test, whereas, all other models pass the test and we are concerned with ARMA 4 0 model, which passes the test.

So, this is the white noise test to examine whether the residual series e t that we have constitute a white noise in, as much as the correlations are all insignificant. That means there is no significant correlation present in the data that is the Whittles test, which is to examine whether the series constitutes a white noise.

(Refer Slide Time: 42:51)

Case study – 3 (Contd.)

Portmanteau test for white noise: $\eta(e) = (N - n_1) \sum_{k=1}^{n_1} \left(\frac{r_k}{r_0} \right)^2$

$\chi^2_{0.95}(k_{max})$	kmax = 48	kmax = 36	kmax = 24	kmax = 12
Model	η	η	η	η
ARMA(1,0)	31.44	33.41	23.02	14.8
ARMA(2,0)	32.03	34.03	24.47	15.17
ARMA(3,0)	30.17	32.05	21.61	13.12
ARMA(4,0)	20.22	21.49	11.85	4.31
ARMA(5,0)	19.84	21.08	11.75	4.14
ARMA(6,0)	19.64	20.87	11.48	3.79
ARMA(1,1)	29.89	31.76	22.24	12.76
ARMA(1,2)	55.88	59.38	48.37	39.85
ARMA(2,1)	55.88	59.38	48.37	38.85
ARMA(2,2)	28.62	30.41	20.38	11.25

 model fails

We have another test, which is called as a Portmanteau test, which again uses r_k , which is a covariance matrix. We go from k is equal to 1 to n_1 . This is simply r_k by r_0 , which is ρ_k if you recall. The r_k is the covariance at lag k and r_0 is covariance at lag 0, which is a variance itself and k is equal to 1 to n_1 ; n_1 is the maximum lag. So, n_1 and k_{max} are similar. Now, this test uses the chi square value, the statistic η follows chi square distribution, so at 95 percent significance value with k_{max} as argument, you get the critical values of chi square as 65, 50.8, 36.0 and 21.0 for various values k_{max} that I have considered.

And the η values that you are obtaining here N is 240, n_1 is the k_{max} here, so for this it is 48, for this 36 and so on. So, you can get η values, the η values here must be less

than the associated chi square value. For these two pass the portmanteau test and in this case, for k_{\max} is equal to 48, all of them pass the test and these two of them fail the test. Again, two of them fail the test, so ARMA 1 2 and ARMA 2 1 fail the test for lower values of k_{\max} and all other models pass the test.

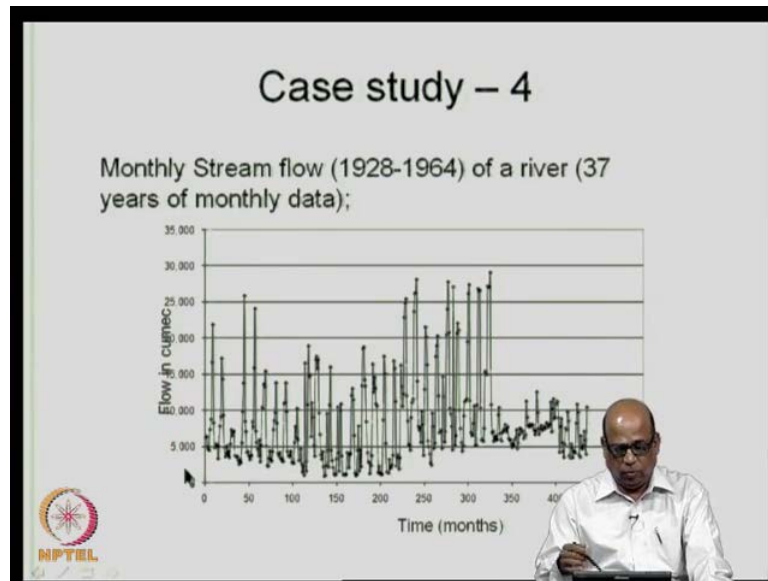
Again, between Portmanteau and Whittle test, we prefer the Whittle test that means suppose there are conflicting results between Portmanteau and the Whittle test then you go with the results of the Whittle test. So, that completes the case study number 3, which is on monthly stream flows for Cauvery River at KRS reservoir. Let me just quickly recapitulate what all we did for case study number three, because this is the most extensive case study that I would discuss in this particular class particular course.

All the other case studies that I will discuss subsequently, we will simply summarize what kind of test we do and so on. We simply summarize the results. So, in the case study three our aim was to finally, build a model, time series model for the monthly time series, both for long term simulation of the data, long term synthetic generation of the data, as well as for term forecasting one time step ahead forecasting.

So, we build the model by considering a number of candidate models. We estimated the parameters by taking half the data and then we did the test by constructing the residual series, by applying the candidate models to the remaining half of the data. Corresponding to each of these residual series, we did the test to examine whether the residual series has a zero mean, whether it is devoid of periodicities. All the periodicities are removed and whether the residual series is uncorrelated.

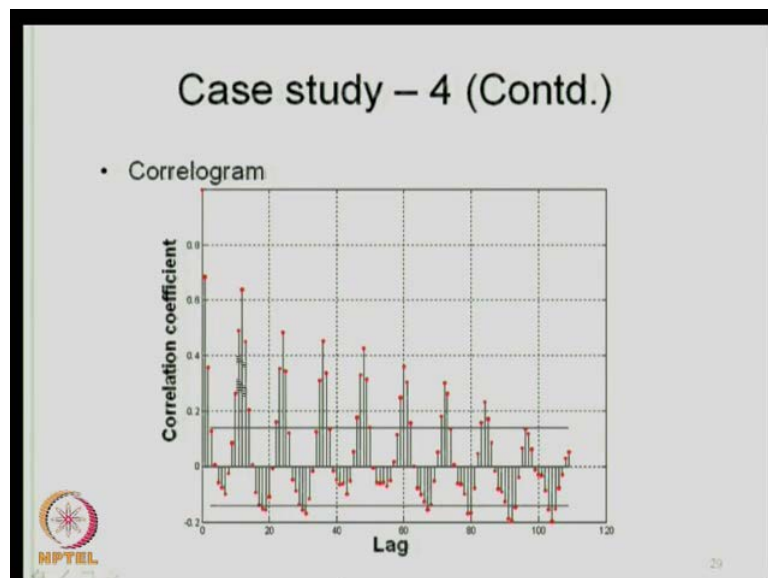
So, these three test we did and then finally, chose ARMA 4 0 model for synthetic generation and ARMA 1 0 model for forecasting of the data. Now, similar case study I will discuss now, but we will not spend as much time as we did on the first case study. Just to give you a variety of the time series that we come across in the hydrology. Next again, we will take monthly stream flow, but not in the monsoon region. See, the Cauvery River typically comes in a region of the country of India, where it is essentially driven by the monsoon patterns, so the monsoon flows contribute to the Cauvery River, so it behaved in a particular manner so we chose those particular types of models.

(Refer Slide Time: 47:24)



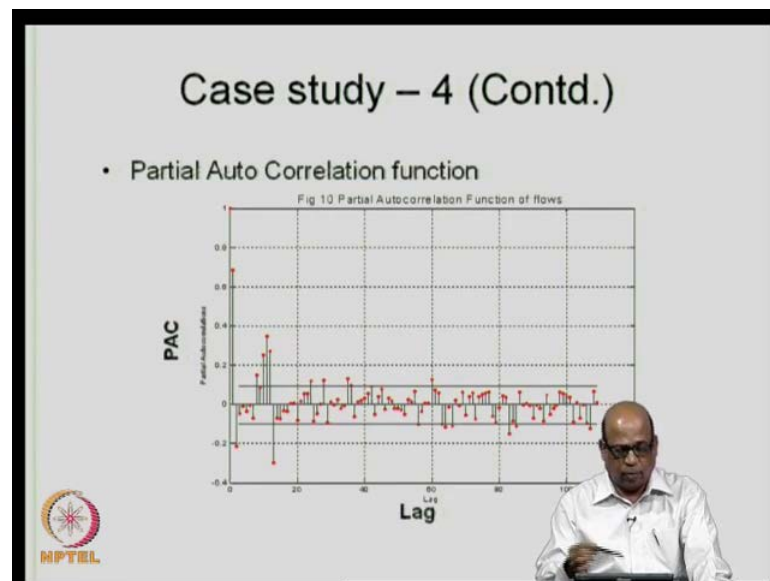
Now, I am taking a stream flow data. This is taken from one of the U S Rivers and the data is directly downloaded from the U.S site and this is the monthly data from 1928 to 1964; 37 years of monthly data is available, so about 4444 values are there. So this is how the time series looks. You can compare the time series of this type of stream flow with any other Indian rivers that we consider. May be you can compare it with the Cauvery River and so on. So, you will see the type of difference is that you may get.

(Refer Slide Time: 48:04)



Then we construct the correlogram and correlogram shows periodicity, but there are significant. There are large numbers of periodicities, which are insignificant, as you can see. These are the bands of 2 by \sqrt{N} on either side, that is plus minus 2 by \sqrt{N} . As you can see most only the positive correlations are significant, most of the negative correlations are insignificant in this particular case. Although this does indicate that there are periodicities present here.

(Refer Slide Time: 48:36)



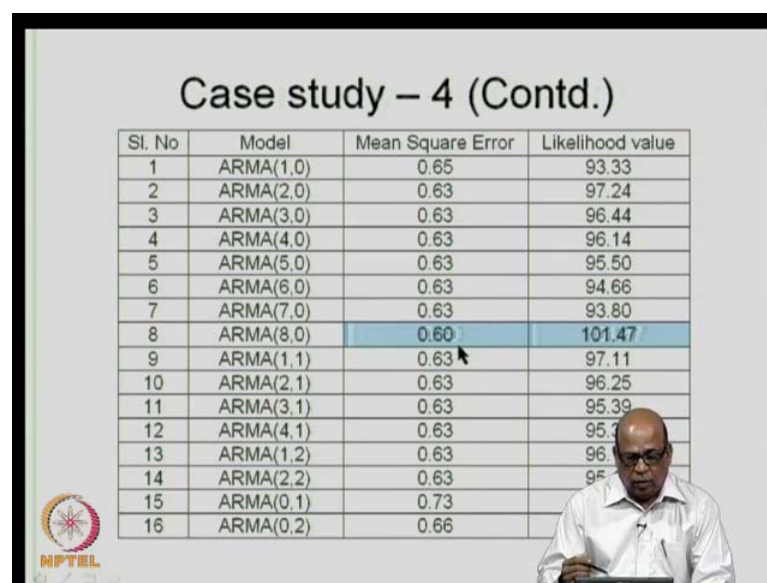
Then we do the **spectral analysis** the partial auto correlations; the partial auto correlations again, there are certain partial auto correlation significant here. The first one is significant and similarly, there are several of them significant. The auto correlogram also shows that there is decay; there is a slow decay in the auto correlations and the partial auto correlations show up some partial auto correlation, which means that if you are build a AR model, you can pick up these partial auto correlations and then pick those particular terms.

Then we go on to build a large number of candidate models like ARMA 1 0 etcetera. It goes on up to ARMA 8 0 and then typically we take up to 1 and 2 of the MA terms and 1 and 2 of the MA terms. So, these are purely moving average models. These are ARMA models where you have both AR as well as MA terms. Here you have only AR terms and we compute the parameters. We estimate the parameters here phi 1 to phi 8 and theta 1 and theta 2, so for ARMA 8 0 model you have all the 8 AR parameters, but no MA parameters here. So similarly, you get all the parameters corresponding to AR and MA terms here.

(Refer Slide Time: 50:54)

Case study – 4 (Contd.)

Sl. No	Model	Mean Square Error	Likelihood value
1	ARMA(1,0)	0.65	93.33
2	ARMA(2,0)	0.63	97.24
3	ARMA(3,0)	0.63	96.44
4	ARMA(4,0)	0.63	96.14
5	ARMA(5,0)	0.63	95.50
6	ARMA(6,0)	0.63	94.66
7	ARMA(7,0)	0.63	93.80
8	ARMA(8,0)	0.60	101.47
9	ARMA(1,1)	0.63	97.11
10	ARMA(2,1)	0.63	96.25
11	ARMA(3,1)	0.63	95.39
12	ARMA(4,1)	0.63	95.7
13	ARMA(1,2)	0.63	96.1
14	ARMA(2,2)	0.63	95.5
15	ARMA(0,1)	0.73	95.5
16	ARMA(0,2)	0.66	95.5



Once we calibrate the models based on half the data, we then calculate both the means square error as well as both the mean square error as well as the likelihood. This is a log likelihood value. As you can see here, the ARMA 8 0 model turns out to be the best, both in terms of the mean square error as well as in terms of the likelihood value, much unlike what we saw in the Cauvery river data In Cauvery river data, we got ARMA 4 0 for synthetic generation whereas, ARMA 1 0 for our short term forecasting. Whereas, in this particular case, which is a US stream, you get ARMA 8 0 as the best model, both for synthetic generation as well as for short term one time step ahead forecasting.

(Refer Slide Time: 51:47)

Case study – 4 (Contd.)


Significance of residual mean:

$$\eta(e) = \frac{N^{-1/2} \bar{e}}{\hat{\rho}^{1/2}}$$

\bar{e} is the estimate of the residual mean
 $\hat{\rho}$ is the estimate of the residual variance

- All models pass the test

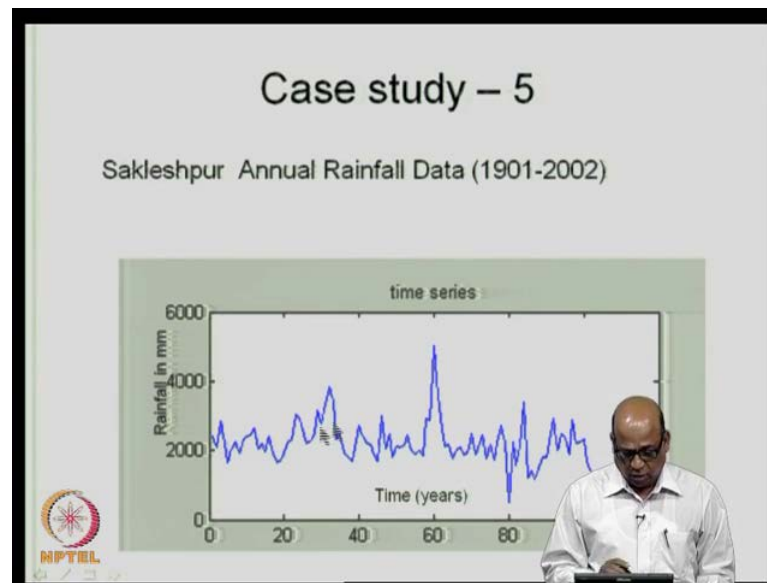
Model	Test t	
	$\eta(e)$	$t(\alpha, N-1)$
ARMA(1,0)	1.78672E-05	1.645
ARMA(2,0)	6.0233E-06	1.645
ARMA(3,0)	4.82085E-05	1.645
ARMA(4,0)	-3.01791E-05	1.645
ARMA(5,0)	6.84076E-16	1.645
ARMA(6,0)	3.0215E-05	1.645
ARMA(7,0)	-6.04496E-06	1.645
ARMA(8,0)	5.54991E-05	1.645
ARMA(1,1)	-0.001132046	1.645
ARMA(2,1)	-0.002650292	1.645
ARMA(3,1)	-0.022776166	1.645
ARMA(4,1)	0.000410668	1.645
ARMA(1,2)	-0.000837092	1.645
ARMA(2,2)	0.002631505	1.645
ARMA(0,1)	0.022950466	1.645
ARMA(0,2)	0.019847826	1.645



Then we do the residual test that is test for on the residual series. So, first we check whether the mean is significant. We formulate this statistic eta e and then compare it with a t alpha N minus 1. In this case N exactly 444, 37 years of values and then you get eta e values, which are very small compare to the N value. They typically come to e to the power minus 5 and so on, so this passes the test. All the models pass the test and typically we are interested in ARMA 8 0; ARMA 8 0 passes the test indicating that the residuals do not have a significant mean or this mean of the residual can be approximated to be zero.

Similarly, we do the entire test. I am not indicating all the details of other test. You can do it as an exercise and then you satisfy that all the tests are **I am sorry** I have not given you the data, so you would not be able to do those test yourself. So, I am indicating here that similar to the residual mean we do the test that I just discussed for the Cauvery case study and then ensure that all the models that we have chosen namely ARMA 8 0 passes all the test.

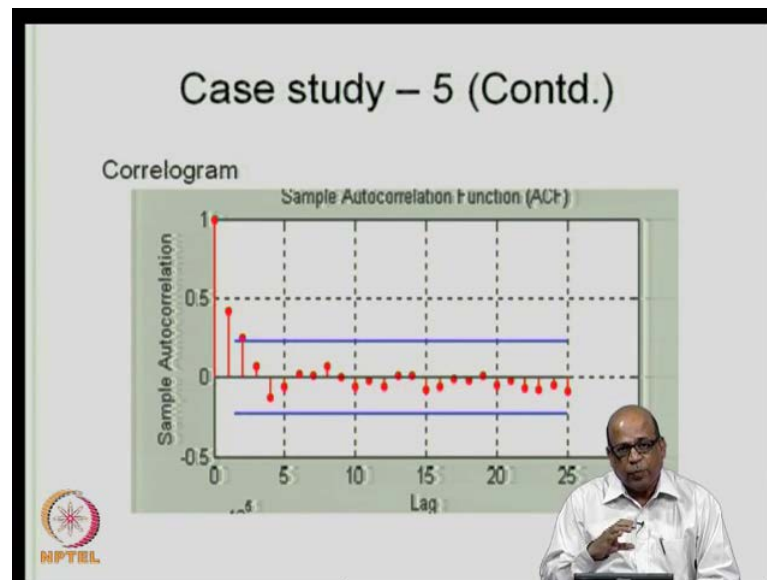
(Refer Slide Time: 53:11)



Now, this is another type of time series that we are considering. the first one, which is case study number three, we considered the Cauvery river flows, which are essentially driven by the monsoon pattern and that is monthly time series. Then we went to another monthly stream flow time series where we consider a western river that is a river in the US and then we did the example of fitting the time series. We came up with a model of ARMA 8 0 or AR 8 0; AR 8.

Now, we look at another type of time series, which is Sakleshpur annual rainfall data. Sakleshpur is in the Western Ghats of the nation of India, Western Ghats, typically the monsoon patterns here are much different compared to any other region in the country, so we consider this annual rainfall data. So, the annual rainfall data in the Sakleshpur region for 1901 to 2002, 102 years of data is available, so you see the time series. Remember, this is annual data. So, you are not getting any, you are not seeing any particular pattern here, so you get the time series plot like this.

(Refer Slide Time: 54:43)



Then we look at the partial auto correlation, this is auto correlation, ACF, and the partial auto correlation and then we go on to build the models and then look at the type of models that we get for synthetic generation as well as one time step ahead forecasting.

As I discuss earlier in one of the earlier lectures, annual rainfall can be modeled as using one of these ARMA type of models, if we are satisfied that, if we transform it to make sure that there is no stationarity; there is no non-stationarity present in the data, whereas, if you go down on time scales, let say that form annual you come to monthly rainfall then from monthly you come to daily rainfall etcetera, there are significant non-stationarity that will be present which are not easily removed, even if you do differencing. Typically, on the daily rainfall data, let say even if you do the first order differencing, second order differencing etcetera the non stationarities will still remain.

So, when you want to apply the ARMA type of models, first you must make sure that you have transformed the original time series to another time series, which is stationary. These models are only for stationary time series. So, the case study number 5 which I am dealing with is annual rainfall series in a high intensity rainfall region, our namely the Western Ghats region of India, we will see how this behave when we apply the time series models.

We will discuss and we will continue this discussion in the next lecture. So, essentially I have dealt with today mainly two case studies; case study number 3, which I started in

the last lecture and we completed the case study number 3 in some detail, where I explained all the test that we do on the residual series. Then we quickly run through another case study, case study number 4, which is a monthly rain monthly stream flow time series of a U S river, United States River, which the data can be downloaded from the US GS side. And then we have just introduced another case study. We have just looked at the annual time series of Sakleshpur data. This case study, I will continue discussion on this particular case study in the next lecture. So, thank you for your attention. We will continue the discussion in the next lecture.