Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 03

Lecture 11: Non-Stationary Time Series

Hello all. So, welcome to this new lecture in the course on Time Series Modeling and Forecasting using R. So, now, if you vaguely remember in the last section, we kind of played around with a practical example in R. And now it's a new week, So, we'll start with a slightly different idea, which is how do you capture non-stationarity? Because remember one thing: Almost all the practical examples that you see around, be it sales data, stock price data, or, let's say, temperature data, etc., would be non-stationary, right? Because there will be some trend component, or there might be some seasonality, or some cyclicality, or in general, there might be some patterns, right?

So, one can't expect that the practical dataset would be stationary to start with, right? So, analyzing any non-stationary data becomes all the more important, right? Okay. All right. So, exactly what we spoke about just a short while back is kind of summarized here. So, by the way, the first step in the analysis of any time series data is to plot the data.

Right. So, for example, if you kind of plot the data like that. So, if the data shows some trend, then one can actually identify that the dataset or the time series is not stationary. So, if the time series behaves as if it has no fixed mean level, for example, even here, the mean is not fixed. The mean keeps on changing.

For example, initially, the mean is at that level. Then the mean level changes to that level, right? And then, eventually, the mean could be somewhere here. So, as long as the mean is not fixed, one can say that the data set is not stationary, okay? But then many times, however, although there is no fixed mean, one can actually see that parts of the time series are local.

So, what we call local parts of the series display a certain kind of homogeneity in the sense that local behavior is observed. Or on shorter intervals. So, what do you mean by that? So, for example, here, let's say if I draw a fresh plot, right? So, let's say overall, there might not be a trend as such, but let's say initially you have a trend which is upward, then probably downward, right?

And then probably upward again for a while, and then probably downward again, and then it continues like that, right? So, here you see that even though you don't have an overall upward trend or a downward trend, in local behavior or local intervals, for example here, and then probably here, and then probably here, you see that in smaller local intervals one can observe some trends. And then this is exactly what we mean by a non-stationary kind of series because the mean is changing throughout, right? So, initially, the mean could be somewhere here, let's say.

Then the mean changes to that level, right? Then again, it comes down to that level. Then, it probably increases to that level again. So, overall, again, even if you don't have an overall trend or overall seasonality, the time series may possess some patterns or some replications in a shorter span or a shorter interval, all right? So, this is a small definition of when you can say that a series is stationary or not stationary, right?

Okay, so, now analyzing any non-stationary series, one has to do a lot of refinements. So, probably from all the previous lectures, we recall that one idea to correct non-stationarity is transformation. So, let me just write it down that one can actually correct non-stationarity by transformation also. So, one can correct non-stationarity by transformation. But now, here we'll see two operators, right?

And so, the idea of these two operators is that if you operate these operators on any time series process yt, it again helps to kind of correct some of the non-stationary ideas. And I think both the operators are kind of easily defined. So, the first one is called the backshift operator. So, what do you mean by the backshift operator? So, the backshift operator, notation-wise, is capital B.

And then, this backshift operator is a really important operator when it comes to defining the time series lags. Now, the definition of this backshift operator is kind of easy. So, if you apply this B on any time series process Yt, for that matter. So, the resultant should be Yt minus 1, and hence the name backshift. So, what we are doing is, you are kind of shifting the time span by one time span in the history.

So, if you apply this backshift operator B on, let us say, Yt, it should produce Yt minus 1. Now, again, you may wonder what would happen if you apply this backshift operator a number of times. So, let us say twice. So, one can do that. So, B square of Yt.

$$BY_t = Y_{t-1}$$

$$B^2Y_t = Y_{t-2}$$

.

.

.

$$B^dY_t = Y_{t-d}$$

So, what do you mean by B square of Yt? So, B square of Yt means nothing, but if you apply this backshift operator twice. So, essentially, this B square of Yt is nothing but if you are applying this backshift operator twice on Yt. So, something like B on B of Yt, right, something like that. So, B of Yt is Yt minus 1, and then B applied on Yt minus 1 is Yt minus 2, OK? So, hence, if you apply this B on Yt minus 1, then again you should backshift the time span by 1 unit, and then the answer should be something like Yt minus 2.

So, the first operator is really easy. So, the name is the backshift operator, and if you apply this b on any time series or current time series value y t, the resultant or the answer should be y t minus 1. And then now I can keep on extending this same idea, right. So, one can actually apply this backshift operator a number of times. So, let us say d times, right?

So, b to the power d applied on y t is nothing but y of t minus d. So, you kind of reduce the time span by those many units. So, this is our first operator. And then the second operator is slightly more important. It's called the differencing operator.

And then the idea is that by suitable differencing of any non-stationary process, we are able to convert that process to a stationary process. And by the way, this technique was due to these two people, which are Box and Jenkins. So, by the way, one small point to note here is that Box and Jenkins kind of gave us a lot of different ideas. For example, the differencing operator, and then Box and Jenkins have their own methodology by which one can actually analyze a practical time series dataset, right? So, how do you convert non-stationary to stationary?

Then, how do you model any practical time series processes and so on and so forth, okay? Now, what exactly do you mean by the differencing operator? So, firstly, notation-wise. So, by the way, you have an inverted triangle kind of notation. If you see this, you have an inverted triangle kind of notation, and then in Greek letters, this is called nabla.

So, it is called the nabla operator. So, n-a-b-l-a. So, let me rewrite this. So, it is called nabla in Greek. All right, and then what happens is if you apply this nabla operator on any time series process yt, So, in general, nabla to the power d applied on yt is nothing but the dth difference of yt, okay, for all t.

$\nabla^d Y_t$ is the d^{th} difference of Y_t for all *t* where,

$$\nabla Y_t = Y_t - Y_{t-1}$$

$$\nabla^2 Y_t = \nabla(\nabla Y_t) = \nabla(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}$$

So, for example, the first difference would be what? So, the first difference would be if you apply this nabla once. So, nabla applied on yt, okay. So, nabla applied on yt is nothing but yt minus yt minus 1. And then here, as the name suggests, there should be some differencing going on. So, this is essentially the differencing that the differencing operator gives you.

So, what you are doing is taking the current value, which is yt, and subtracting one previous value from that. So, yt minus yt minus 1. So, we covered two operators. So, the backshift operator and the nabla operator. And again, as we saw earlier, one can keep on applying this nabla operator a number of times.

So, let us say nabla square of yt, right? So, nabla square of yt would be applying nabla twice. So, nabla applied on nabla yt, ok. So, by the way, this thing is nabla yt. If you apply nabla on top of that, then wherever you have t right, you basically replace that by t minus 1, right? So, for example, here, the resulting idea would be something like yt minus 2 yt minus 1 plus yt minus 2 because if you apply this nabla operator on yt minus yt minus 1, you would get something like this.

So, yt minus yt minus 1 and then minus, as per the definition, you kind of reduce to t minus 1. So, yt minus 1 minus yt minus 2. And then one can combine and get this result. So, here nabla operated on yt is hence called a differenced process. So, initially, you start with yt and then apply this nabla operator on yt to get the differenced process.

$$\nabla Y_t = (1 - B)Y_t$$

Now, one very evident question that might come to your head is: What is the relationship between NABLA and B? So, is there any relationship at all between the backshift operator B and the differencing operator NABLA? And the answer is simple. So, maybe you can pause the video for a second and then try to prove it on your own. But then, we saw that B is nothing but what?

So, B applied on YT, or rather, I should write it this way. So, B applied on YT is nothing but y t minus 1. Right? Okay?

And y t minus 1 also, appears in the NABLA definition, which is exactly here. Right? So, essentially, NABLA applied on y t is nothing but y t minus B of y t. Isn't it? Right? So, y t minus B of y t because B of y t is nothing but y t minus 1.

And then, from here, I can combine 1 minus b in a bracket operated on y t. So, this is the relation. So nabla is kind of similar to 1 minus b. So, nabla is similar to 1 minus b. And then this is exactly what you saw here. So, nabla operated on y t is nothing but 1 minus b collectively operated on y t. So, hopefully, if you understand the ideas behind these two operators, let us say the backshift operator and the differencing operator.

So, the differencing operator is a really powerful tool to kind of convert any non-stationary trend that you have essentially and then convert that to a somewhat mild trend. The idea behind applying a differencing operator is to reduce the amount of trend or reduce the amount of non-stationarity that is present in the model. Okay. So, now we'll take up a couple of examples just to understand how you would make a non-stationary process stationary. For example, the first example you see is a linear trend, and probably all of you must have seen this in our last lectures when we presented some very basic time series models. And there, we actually saw that you have two ways of concluding that this process is not stationary.

$$Y_t = bt + S_t$$

Where, S_t is a stationary process.

So, firstly, you have a trend component, obviously. And if you find out the mean, the mean itself would depend on T. And then if the mean is depending on t, the process has to be non-stationary. Anyways, but then again, if you recall, this is the structure of the linear trend. So, y t equals b t plus s t. Now, here, what exactly is s t?

So, s_t is assumed to be some stationary process. So, s_t in itself is completely stationary, but then obviously y_t is not stationary because it depends on t, through this term, b_t. So, a couple of questions. So, is Y_t stationary? Obviously, the answer is no. Y_t is not stationary.

And then, how would you make it stationary? So, how would you make it stationary? Now, if you think about this problem for a second, applying any transformation would not work here. So, let us say if you apply a square root transformation or if you apply some log transformation, it would not work because the same transformation would appear on the right-hand side, and that necessarily would not ignore the idea of t there. So, what could be done?

Let $W_t = \nabla Y_t$. Then,

$$W_t = (bt + S_t) - (b(t-1) + S_{t-1})$$

$$= b + S_t - S_{t-1}$$

So, here, the nabla operator comes to the rescue. So, what one can do is one can apply nabla on y_t, right? So, if you apply nabla on y_t, what would you have? So, this would be b_t plus s_t minus b and then t minus 1 plus s_t minus 1. Now, again, blindly apply the formula of nabla.

So, nabla applied on yt is what? So, yt minus yt minus 1. So, you take yt as it is and then subtract one prior time point of yt. And here, one can clearly see that I can actually cancel out a few things. So, I can cancel out bt minus bt from here, right?

So, what is remaining is nothing but st and then plus b and then minus st minus 1. And this clearly is stationary because b is constant and st minus st minus 1 is a function of a stationary process, which has to be stationary, okay. So, the answer is if you apply a nabla operator once on yt, you are able to kind of reduce the non-stationarity and then make it stationary in a sense. And this is exactly what is seen here also. So, let us say if you start with wt. So, we are defining wt to be nabla yt. So, wt would be yt minus yt minus 1, and eventually, we will end up getting something like that, which is b plus 1.

st minus st minus 1, which is stationary, okay. Now, another example: let us say quadratic trend. So, what would happen if the trend is kind of stronger? So, initially, we started with a linear trend, and then now the next kind of trend could be quadratic. So, this is the function.

$$Y_t = bt^2 + S_t$$

So, y_t equals b_t square plus s_t, right? Now, again, as before, s_t is a stationary process. So, again, the same couple of questions could be asked: Is y_t stationary, and if not, how would you make it stationary? Now, again, the clear answer is that y t is not stationary, of course, because y_t depends, through this term, on t square, which is a function of time, okay? So, how would you make it stationary?

So, we will see. So, if you apply nabla once, would that make it stationary? Probably not. Because now the trend is much stronger, right? So, if you apply this nabla on y_t once, it won't work out, right?

Because nabla of y_t would be what? So, b_t square plus s_t minus b of t minus 1 square, and then in a bracket, let us say, and then plus s_t minus 1, okay? This is nabla operated once. So, if you apply nabla on y_t once, One can see that I am not able to kind of cancel out the t terms.

I am not able to cancel out the t terms. So, bt squared plus st minus b of t minus 1 whole squared plus st minus 1. So, what could be done here? So, I can essentially apply the nabla operator twice. So, how about this?

So, nabla squared applied on yt. So, nabla squared applied on y t possibly would reduce my y t to a stationary process, ok, and then we will see how. So, in the next slide it gives an explanation. So, let us say you are defining W t to be nabla squared of y t. Then W t would be something like that, ok.

So, y t minus 2 y t minus 1 plus y t minus 2. So, this is the definition of nabla squared y t, right? And then if you replace y t here, then eventually you see that nabla squared y t happens to be that guy, which is 2 b plus a function of a stationary process, which has to be stationary because b is a constant. So, the underlying conclusion is that depending on how strong your trend is, whether you have a linear trend, a quadratic trend, or a cubic trend, right? So, depending on the power of the trend, one has to difference the process that many times.

For example, if you want to reduce a linear trend, then one should apply nabla once. If you want to reduce a quadratic trend, then one should apply nabla twice. If you want to reduce a cubic trend, then one should apply nabla thrice, and so on. So, in general, if the trend is a polynomial of order k, right? So, let us say if you have a really curved kind of trend, something like almost exponential, something like that.

Let $W_t = \nabla^2 Y_t$. Then, $W_t = Y_t - 2Y_{t-1} + Y_{t-2}$. Thus,

$$W_t = b[t^2 - 2(t-1)^2 + (t-2)^2] + S_t - 2S_{t-1} + S_{t-2}$$

$$= 2b + S_t - 2S_{t-1} + S_{t-2}$$

So, obviously, nabla being applied once, twice, or let us say thrice will not do the job. So, let us say if k is bigger than 3. So, if k is, let us say, 6 or 7. But what could be done is we can actually difference the process that many times for stationarity. So, hence we saw that the differencing operator is really powerful in reducing any amount of trend that one might have.

So, let us say linear trend, quadratic trend, cubic trend, or for that matter, any higher power also, okay? Now, again, we will see a small example, which is a random walk, and I am pretty sure that you must be comfortable with the random walk now, right? So, now, again, if you recall, this is the structure of the random walk. So, y t equals y t minus 1 plus e t. Now, long back in one of my earlier lectures, we kind of postulated why a random walk is not stationary.

$$Y_t = Y_{t-1} + e_t$$

And then again, there we kind of gave a proof also. But then just by looking at probably some of the simulations, if you remember, of a random walk. So, how does a random walk look? So, a random walk is purely non-stationary. So, it would look something like that, maybe, where you see an overall upward trend or a downward trend also, right?

But then, can you kind of prove that a random walk is not stationary by using a slightly different idea? Can you do that? So, let's say yt equals yt minus 1 plus et. And then the question is, why is it non-stationary? The question is, why is it non-stationary?

So, there is a small hint for you. So, think about restructuring this model in terms of a particular AR1 kind of model structure. So, if you remember AR1, so AR1 looks like this, by the way. So, y t equals let us say some c plus phi 1 of y t minus 1 plus e t. So, this is nothing but my AR1, is it not? This is a pure AR1 process.

So, can you somehow connect these two models? And one can actually do that because, let us say, if you assume c to be 0, right? So, firstly, there is no c here. So, c is 0. And what is the value of my phi 1?

So, phi 1 happens to be 1 here, right? Because phi 1 is nothing but the coefficient attached to yt minus 1, which is, and then since you do not have any coefficients here, the value of phi 1 has to be 1. So, by kind of restructuring this AR1 process, I can actually obtain a random walk model. And then, if you remember, when we discussed the topic on AR1 and its properties as to when would an AR1 be stationary and so on. If you vaguely remember, any AR1 model would be stationary whenever this

coefficient phi 1 is strictly between minus 1 and 1, right? So, we had a small argument there that whenever the coefficient phi 1 is strictly between minus 1 and 1, the process would be completely stationary, right? But in this situation, we do not have that because phi 1 is exactly equal to 1, right? So, phi 1 is not strictly less than 1, and hence this random walk happens to be not stationary, right? Make sense?

So, this is a slightly different argument. So, if you're kind of sitting in a time series course, one should be comfortable enough to kind of restructure the model as and when possible to a slightly different kind of model and then argue accordingly. So, if you start with a random walk, one can actually restructure this into a possible AR1 kind of model structure. And then, since the coefficient value is exactly equal to 1, the model has to be non-stationary because for any AR1 model to be stationary, the coefficient needs to be entirely between minus 1 and 1, which is not happening in this case.

Now, one evident question is, what about nabla y t, though? So, we will see what happens. So, nabla y t is what? So, nabla y t is nothing but y t minus y t minus 1. So, again, if you come back to this random walk structure.

So, here we will see that y t minus y t minus 1 if you do is nothing but e t. And e t is nothing but entirely stationary, right? Because e t is a random error or a white noise kind of term, okay? So, nabla y t happens to be stationary, but y t itself, hence it is a random walk, is not stationary. So, essentially, what is happening is, let us say this is my y t, right?

So, this is one simulation of a random walk, but if you difference this once, right? If you difference this once, for example, if you take something like ∇Yt, then the picture would be different. So, ∇Yt is hopefully stationary, something like that. So, this is a small example of a particular case where you are trying to analyze a random walk. Now, we will kind of shift attention to seasonal models.

So, down the line, probably we will spend some time on modeling seasonality using a particular ARIMA structure, which is SARIMA, by the way. So far, we have not introduced ARIMA also. So, we will do it subsequently. But then, a quick idea as to how you would apply the differencing operator if seasonality is present. Now, a few very quick examples of seasonality just to recall.

$$Y_t = m_t + s_t + S_t$$

So, let's say airline passengers, sales data, or let's say accident data. So, all these examples have to have some seasonality present. Or, let's say rainfall data, temperature data, humidity data, et cetera. And then, the classical decomposition model is given by that model. So, let's say your time series value Yt could be broken down into three parts.

So, mt, small st, and big st. So, what is MT? So, MT is a pure trend component. Small st is nothing but a pure seasonal component with a particular period. So, period D. So, we will talk about what you mean by this period very soon.

And then lastly, as I have seen before, this capital ST is a purely stationary process. So, my YT is broken down into a trend component, a seasonality component, and a purely stationary component. Now, differencing at lag D. So, a slightly different idea. So, differencing at lag D and how exactly things are changing here. So, let us say, what do you mean when you apply the lag D difference?

$$W_t = \nabla_d Y_t = (Y_t - Y_{t-d})$$

$$= (m_t - m_{t-d}) + (S_t - S_{t-d})$$

So, we call this difference the lag D difference. The notation is slightly different. So, let us say Wt is nabla subscript D rather than power. So, nabla subscript D operated on Yt. So, the lag D difference is nothing but the difference between the current value and the Dth lag value.

So, yt minus D. So, again, if you remember, a quick revision. So, nabla yt was what? So, nabla yt was nothing but yt minus yt minus 1. So, essentially, my nabla yt is nothing but lag 1 difference. So, this is called lag 1 difference.

So, and then, if you want to take, say, nabla 2 as a subscript of yt. So, what is this? So, this is nothing but my lag 2 difference. And then, the lag 2 difference would be what?

So, you are taking a difference between yt and yt minus 2 directly. And so on. So, lag 3 difference, lag 4 difference. So, in general, what do you mean by lag D difference? So, the lag D difference is you are taking a difference between yt and its dth lag.

So, yt minus yt minus d. So, let us see what would happen if you take a lag D difference of the classical decomposition model that we saw earlier. So, what you would have is something like this. So, MT minus MT minus D plus the difference between the stationary version. So, a particular lag D difference kind of completely ignores or reduces the seasonality aspect.

So, as a small comment, a lag D difference removes the seasonality of period D. So, here, since you do not see any small st minus small st minus d because small st minus small st minus d would be kind of 0. So, if one wants to correct seasonality, then one has to apply a lag d difference. If one has to correct a trend or reduce a trend, then one can actually apply either nabla or nabla square or nabla cube or, in general, nabla to the power d. So, hopefully, the idea is clear.

So, on one hand, you have nabla to the power d, and on the other hand, you have nabla subscript d, which is called lag d difference, all right. So, here we are clearly seeing that if you want to reduce seasonality, then one should actually take a lag d difference on the time series process y2. Now, what exactly do you mean by this period d? So, any seasonal time series comes with some particular repetitions, right. So, let us say temperature data.

So, now, if you see the temperature data, if you try to draw the plot of temperatures, it would be seasonal, obviously, right? So, let us say all the peaks are nothing but summer months. So, let us say June of a particular year, then June of the next year, then the June of the next year, and so on. So, all these could be June or July months, and all the troughs could be, let us say, December or January. So, I will write down D for December.

And then the plot is, So, on the x-axis, you have all the years, and on the y-axis, you have the temperatures. So, let's say this is in 2000, then in 2001, then 2002, etc. So, here, if you see, you see a repeating pattern because it has to be repeating, as the time series is seasonal. But how exactly is it repeating? So, let's say, you have a repetition which is kind of occurring after every 12 months, isn't it?

Because here, you are seeing peaks which are repeating themselves after every 12 months. So, June of a particular year and June of the next year, So, 12 months have gone, you see a similar kind of temperature, and similarly, all the troughs, right? So, for temperature data, one can say that you have a seasonality of period 12 because you see repetitions after every 12 months. So, in general, what do you mean by period D now? So, period D means that the seasonality is being repeated after every D periods rather than 12.

Does that make sense? So, I will give you one more question. So, what do you mean by, let us say, the seasonality of period 4? So, the seasonality of period 4 could mean that something is repeated every quarter. So, let us say in January, then the next thing is in April, then the next thing is in July, and the last thing is in October.

So, if something is repeated every quarter, then the period is 4, or if something is repeated every 12 months, then the period could be 12. So, depending on how the repetitions occur in a seasonal time series, one can actually define the period of seasonality. And the last statement here is that, depending on what period of seasonality you have, one has to apply the lag D difference to remove that seasonality aspect. So now, the last thing we will discuss is the definition of the ARIMA PDQ process. So, how do you define the ARIMA process?

So, by the way, the full form of ARIMA is auto-regressive integrated moving average. So, autoregressive integrated moving average. So, this ARIMA process has two components. So, an AR component and an MA component. But then there has to be some trend in the data also.

So, ARIMA is a classic example of a non-stationary series to start with. And any ARIMA process comes with three orders or a triplet of orders. So, P, D, Q. And then the definition is kind of easy if you read this multiple times. So, what exactly is an ARIMA process?

So, any non-stationary time series process yt follows a particular ARIMA p, d, q model if the dth difference. So, by the way, this is not lag d difference, this is dth difference. So, nabla to the power d. If the dth difference, which is nothing but produced by differencing yt d times, is a stationary ARMA p, q process. So, what do you mean by that?

So, let us say, initially, you have a time series y t and y t is assumed to be modeled using an ARIMA process, all right. And then, since y t is not stationary, it has to have a strong trend component. So, let us say what we do is we difference this d times, all right. So, we apply NABLA d times and then the resulting series, which is let us say w t in this case. So, w t happens to be a stationary ARMA process with orders p, q.

And then you have an assumption here that this process should be stationary. So essentially, what you are doing is starting with some ARIMA model and differencing that ARIMA process d times. So, as to get a stationary ARMA process of orders p, q, then this initial y_t that I have. So, initially y_t has to be modeled using an ARIMA (p, d, q) process. So, here again, just to summarize, the idea is kind of easy: the p order stands for the autoregressive order, the q order stands for the moving average order, and then d stands for how many times you want to difference.

So, I will take a simple example. So, let us say you start with y_t and then y_t is assumed to be non-stationary with some trend component. So, you have some strong trends, ok. And then let us say if you apply nabla two times, alright. So, if you apply nabla two times and then the resulting series w_t happens to be a stationary ARMA process with orders 2, 1, let us say, right.

Now, just pause the video for a second and then tell me what could be the model for the initial time series y_t. So, what is the answer? So, again, just to summarize. So, since you started with y_t and then you applied the differencing operator twice to obtain a stationary ARMA process with orders 2, 1. So, my initial y_t process should be ARIMA with orders 2, 2, 1.

So, why 2, 2, 1? Because the AR order is 2, and the MA order is 1 from this ARMA 2, 1. And then, since you have different this twice, the middle order is also 2. So, probably in the next session, we will revise this ARIMA process one more time if you want. And then, probably subsequently, we will try to extend this to a particular seasonal ARIMA process.

It is called SARIMA. And then we will talk about its properties and probably some examples. So, as to understand this ARIMA process and the subsequent forecasting in an easy way. Thank you. Thank you.