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 14: Seasonality and SARIMA Model

Hello all, welcome to this new video in this course on time series modeling and forecasting using R. Before we start with anything new, just to give you a brief overview of what we covered in the last session. Currently, we are exploring the idea of non-stationarity, right? Because again, if you remember, almost 90% to 95% of the practical applications that you see in time series are non-stationary or they have a non-stationary kind of behavior. So analyzing any non-stationary aspects, such as trends, seasonal variations, cyclicalities, or any other pattern, becomes all the more important. So, in this session, we will discuss one more non-stationary time series model, which is called SARIMA.

SARIMA is a very natural extension of ARIMA. Probably in the last couple of lectures, we have explored enough about the model, which is ARIMA. If you remember, the full form of ARIMA is called auto-regressive integrated moving average. ARIMA is capable of handling any trends in the model. Depending on whether you have a linear trend, quadratic trend, or cubic trend, one can actually difference the model those many times.

If you want to difference the model once, then the integration power would be 1. So that middle order in the ARIMA model or the ARIMA structure would be 1. Obviously, all three orders correspond to P, D, Q. P and Q correspond to the AR order and the MA order, respectively, while the middle order, which is D, stands for the number of times one has to difference the series to achieve stationarity. Now, slightly on the other side, if you extend this ARIMA model to a SARIMA model.

So, how do we kind of bring in the idea of seasonality? So, SARIMA is nothing but seasonal ARIMA, or in other words, seasonal autoregressive integrated moving average. So, there should be some amount of trend in the data, as well as some aspect of seasonality in the data. So, before we move ahead with the SARIMA model, a few other

points regarding non-stationarity once again. So, in general, the non-stationary behavior of any time series could be due to multiple reasons.

So, for example, it could be due to trends or, let us say, unit roots. So, just a very brief introduction or kind of a revision about unit roots. So, in the last session, we talked about unit roots. So, if you remember, what we mean by unit root is that if you solve for that polynomial that you get from the ARIMA structure, and if at least one of the roots happens to be exactly equal to 1, then we have a situation called a unit root situation. And again, we saw that unit roots kind of amount to non-stationarity.

Because, if you remember, we argued that unit root What happens is that the current time series value depends on the past value plus some function of random errors. So, in a given long term or in a longer horizon, such a series won't converge to anything or won't converge to its mean. So, in that sense, the series would be non-stationary. So, the presence of unit roots.

The third is seasonality. So the idea of this session would be more about seasonality because we will expand upon that SARIMA kind of model structure. And then the last one is changing variance. So, of course, if you see some patterns in the variance or the variance is not constant, then again it can amount to some degree of non-stationarity in the data. Now, again, a very quick refresher about what exactly do you mean by seasonality.

So, seasonality implies that firstly the mean of the observations is not constant, right, and it evolves according to a cyclical pattern. So, we studied some properties about seasonality in the last session if you remember. So, let us say if you have fixed frequencies and the series is kind of oscillating, then we will say that you have some seasonal effect or you have seasonality in the picture, okay? And if you remember, the other idea was cyclicality, right? So, cyclicality means that the repetitions are not constant.

So, the repetitions or the frequency of the repetitions may vary, okay? So you have two different ideas. So seasonal or seasonality and then cyclicality, okay? All right. So, in fact, seasonality could be of multiple types, right?

So, we will briefly discuss what those types are before we kind of go to the SARIMA model, okay? So, these are all the types of seasonality. So, seasonality could be

deterministic. So, what do you mean by that? So, for example, if the seasonality is constant in the same month over different years.

So, let us say if you have data and the aspect or the amount of seasonality is the same or constant in the same month over different years, then we will say that the seasonality aspect or the seasonality factor is deterministic. Now, in general, even as we took in the last session. So, let us say the period of seasonality would be some s, okay? Now, again, remember what do you mean by the period of seasonality. So, the period of seasonality means after how much frequency you see the repetitions, okay?

So, let us say the period of seasonality is s, then we actually get something like this structure, okay? So, let us say s_t to the power s within brackets, okay? If it happens to be exactly the same as some s_t plus k_s to the power s, then in such a situation, we will say that seasonality is deterministic. And then this could work for any case. So, let us say k could be plus or minus 1 or plus or minus 2 or plus or minus 3 and so on and so forth.

$$S_t^{(s)} = S_{t+ks}^{(s)}, \ \ k = \pm 1, \pm 2, \pm 3, \ldots$$

So essentially, just if you look at this equation one more time, what exactly you mean is that if the current value or the current factor of the seasonality at time point t happens to be exactly the same as the factor of seasonality at time point, let us say t plus ks. So, t plus ks, what we are doing is we are shifting the lag. Right, and rather than simply taking t plus k as we did earlier, we are kind of taking t plus ks because s stands for the seasonal factor, right? So, we can take one example: let us say if k is 1 and if you have monthly, let us say, temperature data, then generally s is 12, right? So, what do you mean by this equation? So, this would reduce to something like st and then within brackets s. If this thing happens to be equal, then st plus k into s would be 12, right?

So, this would be t plus 12 and then s. So, again, if you go to this structure, essentially what you mean is that let us say the value in January of a particular year, right? So, let us say 2023, okay? So, the value you have in January of 2023, if that value happens to be the same in January of next year, so let us say 2024, right? Because you have a shift of 12 here.

So, January of a particular year and then January of the next year. Then what would happen? If such a situation arises, then you have a deterministic kind of seasonality. Now, one can actually extend this. So, let us say this could be equal to S and then S obviously, and then let us say if you increase this K to 2 now.

So, let us say if K is 2, then what would happen? This would be nothing but T plus 24, right? And so on. Right. So here, clearly, one can see that you have gaps of, or you have jumps of, 12 between any two consecutive points. So, let us say you start here, then T plus 12, then T plus 24, then the next one would be T plus 36, etc.

OK. So, in this case, we will say that you have a deterministic kind of seasonality. So, this is the first type. Then, the next type is what would happen if seasonality evolves over time as a stationary process. So, what do you mean by that?

So, let us say again the factor of seasonality is STS, and then assuming that STS is a stationary factor oscillating around some mean. So, let us say mu S. So, this is the actual equation. So, STS happens to be equal to some mu S plus VT, and what exactly do you mean by VT? So, let us say VT is some process. Such that the expectation of VT is 0.

$$S_t^{(s)} = \mu^{(s)} + v_t, \;\; E(v_t) = 0$$

Okay. Such that the expectation of VT is 0. And then the idea of bringing this VT is nothing but. So, VT is a stationary process and sort of brings some variability in this factor of STS. Right.

Because STS does not depend on this mean entirely. Right. So, you have this additional VT portion here. So, since you are adding this extra VT term. So, VT sort of brings in some variability.

Okay. And why exactly is VT stationary? Because VT is stationary because the expectation is zero. So, the expectation is constant and so on and so forth. Or for that matter, this is more like an assumption.

So, you can write down the assumption. So, you're kind of assuming that VT is stationary. But nevertheless, this VT, which is a stationary process, sort of brings in a kind of variability in the seasonality factor or the value of seasonality. So, such an equation will say that you have a certain seasonality for sure, but that seasonality kind of fluctuates or evolves over time as a stationary process. Now, the third kind of seasonality could be that seasonality evolves over time as a non-stationary process.

So, earlier we saw how it evolved as a stationary process, and now we will see how seasonality evolves over time as a non-stationary process. So, here again, we have this equation if you take note of this. So, we have STS, which is the seasonal factor, and what we are doing is we are creating a random walk out of that. So, can you see the structure

here and then kind of resemble the structure of a random walk? So, in a random walk, what exactly do you have?

$$S_t^{(s)} = S_{t-s}^{(s)} + v_t, \;\; E(v_t) = 0$$

So, you have something like this, right? So, y t equals y t minus 1 plus some error term, right? So, this is a very, very typical random walk. So, similarly, here what we are doing is we are finding out the current seasonality factor, and if that seasonality factor sort of depends on s t minus s plus some stationary process. But since this is a random walk, seasonality itself is a non-stationary process in this case.

So, in this case, STS itself may follow a non-stationary process. So, say a random walk, as this equation tells you. Now, again, you have this error term here, which is VT. So, again, the idea is that we are kind of assuming that VT has an expectation of 0, and then VT again sort of brings in some variability in that seasonality factor. But thus, what would happen is if you apply some seasonal differencing here.

So, thus, applying some sort of a seasonal differencing actually corrects seasonality in all three cases. So, the third one, the one prior to this, and the first one. So, the idea is one should always apply some appropriate seasonal differencing. Now, again, remember what exactly do you mean by seasonal differencing? So, seasonal differencing means you want to apply some differencing at a particular lag, right?

So, let us say lag s. So, differencing at lag s would be something like y t minus y of t minus s, directly, okay. So, you apply some sort of a suitable seasonal differencing, depending on the data, of course, right, to sort of correct seasonality in all three cases or in all three seasonality types, okay. So, this is a brief idea as to different kinds of seasonality that one can actually encounter. So, seasonality can evolve as a non-stationary process, seasonality can evolve as a stationary process, and if you go back a couple of slides, the first one was if you have a deterministic sort of a seasonality. Okay, so now, in this slide, what we will do is we will cover a very brief idea about a SARIMA model.

Okay, so now again, SARIMA stands for seasonal autoregressive integrated moving average. So, seasonal autoregressive integrated moving average, and then, as explained in the beginning of this lecture itself, a SARIMA model has different parts to it, right? So, you should have an AR part, there should be some MA part, and the overall model should be integrated. So, there should be some trend aspect, and there should be some seasonality aspect as well, right? So, hence, SARIMA.

Now, we have two points to note here, which are again important. We have covered this multiple times: if you have a trend aspect, right? So, one can actually convert any trend to a stationary series by taking regular differencing, right? So, in this case, how would the differencing look like? So, one can actually take something like yt minus yt minus 1, which is nothing but given by nabla yt.

And so on. So, depending on the strength of the trend. So, if you have a quadratic trend, cubic trend, or to the fourth power or some other power, then, depending on how strong the trend is, one has to difference the series those many times. But all such differences are nothing but regular differences. Because yt minus yt minus 1 is, in fact, nabla yt, and that is nabla yt is a regular difference.

On the other hand, if you want to convert any seasonality to stationarity, then one has to take seasonal differences, okay? And in the last slide, we saw what you mean by seasonal differences. So, seasonal differences, again, a quick refresher, mean nothing but yt minus y of t minus s directly, okay? So, this would be something like a lag s kind of difference, right? So, take the current value and then subtract y of t minus s from that, all right?

Okay. Now, how do you actually arrive at the particular SARIMA model, right? So, let us say the initial yt that you had earlier, you make a small transformation to that. So, let us say zt is a new process, and how are you getting to zt? So, zt is nothing but through this polynomial applied on yt or through this operator applied on yt.

$$Z_t = (1 - B^s)^D (1 - B)^d Y_t,$$

Where D is the number of seasonal differences (usually 0 or 1)

And d is the number of regular differences (usually ≤ 3)

Now, if you look at this operator for a second. So, what exactly do you mean by that? So, the first chunk in this operator is 1 minus B to the power s, right? Whole to the power capital D, and then the second chunk is 1 minus B to the power small d. So here you see a lot of notations. So, you see small d, then you see capital D, and eventually you see S. So, all of you know what S means.

So S is nothing but the period of seasonality. So, if you have monthly data, then S would be 12. If you have quarterly data, then S is 4, and so on. Now, what about capital D and small d? So capital D is nothing but the number of seasonal differences.

And then, this number is usually 0 or 1. So, how many times does one have to difference the data on a seasonal basis? And generally speaking, for capturing any practical data sets or practical problems, one need not go beyond 1 with respect to capital D. And at the same time, what do you mean by small d? So small d is nothing but the number of times one has to take the regular differences.

So, this one is seasonal difference. So, something like yt minus yt minus s, and then capital D kind of tells you how many times you are taking such differences. And then, on the other hand, small d is nothing but the number of regular differences. So, something like yt minus yt minus 1, which is nothing but nabla. Okay.

So, to sort of differentiate between the kinds of differencing one has to perform, we are kind of using small d and capital D. So, essentially, the idea of the SARIMA model is to model regular and seasonal impacts separately and then sort of combine them into a model multiplicatively. So, even here, you see that we are kind of multiplying these two factors. So, one of the factors is seasonal, which is the first one. The other factor is non-seasonal. So, incorporating both these individually and then sort of combining them in a form of multiplication.

So, this is the idea of a SARIMA model. Now, how exactly does a SARIMA model look? So, the equation is not at all easy, but then just to understand, the SARIMA model has lots of individual parts as you see in this equation. So, the entire left-hand side and the right-hand side are a combination of seasonal and non-seasonal parts both. So, again, if you focus here on the LHS, right.

So, you see capital phi, by the way, this notation is capital phi, and then the subscript is capital P, and then B to the power S, and then small phi or regular phi, and then the subscript is small p, right. So, essentially, what happens is the SARIMA model is kind of a combination of AR structure of a seasonal order, then AR structure of a non-seasonal order, then MA structure of a seasonal order, and then similarly MA structure of a non-seasonal order. So, can you sort of digest this idea that the SARIMA model, since it is a combination of AR and MA, right? So, there have to be two orders, right? So, P comma Q, but at the same time, since you are controlling both for the trend and seasonality, then one has to differentiate in multiple ways.

So, seasonal differencing and then non-seasonal differencing, which you also call regular differencing. So, hence we see this capital D and then small d here, right? And so, capital phi corresponds to seasonal orders, small phi corresponds to non-seasonal orders, capital theta. So, this is capital theta, by the way, corresponds to seasonal MA orders, and then small theta corresponds to non-seasonal MA orders, right? So, essentially speaking, what you have is you have AR structure and MA structure for both seasonal and non-seasonal orders, and hence we see four coefficients.

$$\Phi_P(B^S)\phi_p(B)(1-B^s)^D(1-B)^d Y_t = \Theta_Q(B^S)\theta_q(B)e_t,$$

Where,

$\Phi_P(B^S) = 1 - \Phi_1 B^S - \cdots - \Phi_P B^{S^P}$ is the seasonal AR operator of order P

$\phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B$ is the regular AR operator of order p

$\Theta_Q(B^S) = 1 - \Theta_1 B^S - \cdots - \Theta_P B^{S^Q}$ is the seasonal MA operator of order Q

$\theta_q(B) = 1 - \theta_1 B - \cdots - \theta_q B$ is the regular MA operator of order q

So, capital phi, small phi, capital theta, and then small theta, right? And then, depending on whether you are having an AR kind of ordering or MA kind of ordering, we are using P or Q. So, P stands for AR orders, right, as usual, and then Q stands for MA orders. The only extra thing here is that we are using capital P and small p because capital P stands for seasonality, and small p stands for regular AR order. Similarly, capital Q stands for seasonal MA order, and then small Q stands for non-seasonal MA order, right? So, hopefully, maybe you can pause the video and then just digest this equation one more time as to what exactly is happening here, right? So, again, just to summarize very quickly, any SARIMA model has these four paths essentially, right?

So, capital phi P is the seasonal AR operator of order capital P, and small phi p is the regular AR operator of order small p. So, when we say regular, this is non-seasonal, by the way, right? Similarly, capital theta is the seasonal MA operator of order capital Q, and then small theta is the regular MA operator of order small q. So, essentially speaking, any SARIMA model has to be defined using two individual triplets of ordering. So, small p, small d, small q, and then capital P, capital D, capital Q, and then towards the end, we can actually specify the period of the seasonality, which is S. So, again, if you have monthly data, this S would be nothing but 12. So, hopefully, this is clear.

So, we will take up an example now, which is what you see in front of you. So, So, here, what exactly do you see in this table? Here is the structure of the data. So, let us say you have months given by columns. So, the first month, then the second month, third month, all the way up to the twelfth month.

So, let us say January, February, March, April, up to December. So, this would be my December month, and then this would be my January month. And then, at the same time, the first-row sort of gives you the years. So, let's say you have R number of years. So, you have monthly data over how many years?

So, over a small R number of years. And one can actually structure the data in a tabulated form. So, the first month, then the first-year value is Y1. First month, second year value is Y13 because once 12 of the observations have been exhausted, then the 13th observation corresponds to January again. So, for that matter, the entire first column corresponds to data from January over all these R years.

And subsequently, the second column corresponds to data from February over all these R years. And similarly, the last column corresponds to data from December over all these R years. So, Y12, then Y24, then the next one would be Y36. And then, in general, the last one is Y12R. So, from this data structure, a structure of the data, we can actually formulate this SARIMA model.

Okay. Okay. So here, firstly, you should understand that the period of seasonality is 12, clearly, right? Because since you have monthly data, the period of seasonality is 12, okay? And then, how exactly are you writing down the models?

$$\Phi_P(B^{12})\phi_p(B)Y_t = \Theta_Q(B^{12})\theta_q(B)e_t$$

And let us say capital Phi of capital P and then B to the power of 12. So why 12? Because S is 12, right? So, you have to raise it to the power of 12. And then small Phi, small P, and then B operated on YT, right?

And on the right-hand side, we have the MA ordering. So, capital theta capital Q again the same thing. So, B to the power of 12 and then small theta small Q B operated on the error. So, towards the end, you see the single error term which is EP, which is nothing but the error term. So, essentially, what is happening is the entire left-hand side of the model stands for the AR structure comprising both seasonal as well as non-seasonal ordering, and the entire right-hand side of the equation corresponds to the MA structure, which is both seasonal as well as non-seasonal.

Now, again, remember that capital Greek letters, so capital phi or capital theta, they correspond to seasonal ordering. Similarly, capital P and capital Q, while the regular phi and then regular theta or small phi and small theta along with small p and small q, they correspond to non-seasonal order, okay. So, in fact, one can actually write down such a model from this data structure, okay. Now, we are almost towards the end of this session. So, we will talk briefly about what exactly are the pros and cons of, let us say, ARIMA and SARIMA models, right.

So, each model has some advantages and some disadvantages, right. So, what could be some pros of, let us say, the ARIMA model or a particular SARIMA model, right. So, firstly, any ARIMA or SARIMA model is easy to understand and interpret, right, because if you have both trend and seasonality which is working on the same series at the same time, then how are you kind of combining both the aspects, right? So, seasonal and non-seasonal, then how are you kind of combining trend and seasonality in general, right? So, that idea is slightly easy to understand and interpret, okay?

So, the simplicity and interpretability of the models are there, okay? And then the other pro is limited variables. So, you have fewer parameters to choose and estimate. So, again, if you go back a slide, for example here, right? So, the set of parameters that need to be estimated are nothing but all these big phi's, all these small phi's, all these bigger thetas, and then all these smaller thetas, isn't it right?

So, once you are able to estimate all these parameters, then your entire SARIMA model is kind of estimated, right? So, even if you have an unknown population SARIMA model to start with, once you estimate all these individual parameters, then your complete SARIMA model is kind of estimated. So, these are some pros, and the last one, what about the cons, right? So, let us say exponential time complexity.

So, when the values of p and q increase, there are equally more coefficients to fit, hence increasing the time complexity. So, as one sort of increases p and q, then one can actually imagine that the equation becomes more and more difficult, right? Second one is complex data. So, there can be a possibility where your data is too complex and there is no optimal solution for p and q. So, if your actual practical data itself is so complex that the algorithm does not converge to a particular value of p or q. Last one is the amount of data needed.

So both the algorithms, be it ARIMA or SARIMA, require considerable data to work on, especially if the data is seasonal. So, for example, using 3 years of historical demand is

likely not to be enough because 3 years would be a short life cycle product. But let us say if the outlook is to estimate on a longer scale to obtain a good forecast, then even if you bring in seasonality, such shorter periods are not enough. So, these are some pros and cons of the models, which are ARIMA and SARIMA. So, I think in the next session, we will take up some practical ideas of, let us say, fitting some ARIMA model on a dataset and so on and so forth, and then see how it works out.

Thank you.