Time Series Modelling and Forecasting with Applications in R

Prof. Sudeep Bapat

Shailesh J. Mehta School of Management

Indian Institute of Technology Bombay

Week 01

Lecture 02: Examples of Time Series Data

Hello all, to this second lecture of the course on time series forecasting with applications in R. So, just to give you a brief overview of what we talked about last time in the first lecture We discussed lot of examples of time series coming from different areas, if you remember, such as finance or climatology or ecology or environmental sciences, etc. And we also talked about different data types. So, cross-sectional data or time series data or panel data.

So, panel data, if you remember, is nothing but a combination of cross-sectional data and time series data. So, now in this lecture we will focus more on, we will elaborate on what we talked about last time by bringing in some other examples. But I think the very first thing we will discuss is why time series? So, this entire area of time series, why is it important? Okay So, the first point is it helps to understand the underlying trend and patterns.

So, like we discussed last time, trend is a very useful friend of all of us. So, trend means what? So, trend is nothing but an upward or downward movement. So, let's say if you observe the Google stock price from the last lecture. So, initially the stock price trended.

Then in between you didn't see any trend. And down the line there was again a trend. So, how to analyze or how to understand the underlying trend in the data or all the patterns in the data? This is the first aim. The second aim is to predict the likelihood of any future events.

So, a simple example could be again linking to the Google stock price. So, let's say if somebody is interested in forecasting the Google stock price for the next month or forecasting the Apple stock price for the next year, then they have to apply some time series analysis. The third one is making inference about the underlying probability model. which governs the series. So, if you want to understand the underlying probability model, so let's say how are all the observations or the time series observations xt, xt plus one, xt plus two related. So, can you put forward some distribution on that or not?

So, if one is more interested in understanding the underlying probability model governing the series, then one has to apply some time series analysis. And the last one could be obviously to develop some confidence intervals or hypothesis testing on the model parameters. So again, just to summarize before we proceed, I'll give you how exactly people write down time series observations or what exactly is a time series data notation. So, different people kind of prefer different notations. The first one is something like this,

$$X_t, X_{t+h}, X_{t+2h}, X_{t+3h}, \ldots$$

So again, if you see here, the t is the same time point, so the underlying time point. And h is what? So, h is nothing but, if you look closely here, h is nothing but the gap between any two observations. So, how much exactly is the gap between the first observation and the second observation is nothing but h. Similarly, how much is the gap between the second observation and the third observation?

Again, it is nothing but h and so on. So, if you notice here, what exactly h is? So, h is nothing but the time between any two observations. Or this is again a different kind of a notation that some people prefer,

$$X(t), X(t+h), X(t+2h), \ldots$$

So, throughout this entire course, we'll kind of stick to one notation, which is the first one, just to be simplistic. All right. So, X_t, X_{t+h}, X_{t+2h} in the subscript. Now, again, just to elaborate very quickly on what exactly is h and what exactly is 1 by h. So, 1 by h is what? So, 1 by h is nothing but the sampling frequency.

So, how frequently are you collecting the observations? or how frequently are you sampling from the data set? all right. So, h stands for the gap between the observations and 1 by h stands for the sampling frequency. So, I think it is time to elaborate a bit more on these two notations. So, I will give you a simple example.

So, let us say we can talk about daily temperature data. ok? Now here let's say we talk about X_1 or let's say X_t, then the next one could be X_{t+2}, then let's say X_{t+4}, etcetera, all right? So, assume that you're talking about daily temperature levels, but how exactly are you collecting the observations? You're not collecting the observations each day, but you are kind of skipping one day if you notice here.

So, the first observation is let's say today. Then the next observation is not tomorrow, but then day after tomorrow because it's t and then t+2. So, t+1 is missing here. All right. So, how exactly are you collecting the observations is if you see the gap between any two observations.

So, how much is that? So, the gap between any two observations or which is nothing but h in our notation is nothing but two here. So, you are collecting one observation let us say today, then the next observation would be day after tomorrow, then the next observation would be 2 days after that and so on. Alright. So, in this context the value of h is, 2 and now if you talk about value of 1/h. So, value of 1/h is nothing but 1/2 or 1 half. So, how would you elaborate on this 1/h aspect?

So, 1/h means that that you are actually collecting one observation every two time points which is nothing but the sampling frequency Right. So, h is what again so h is nothing but the gap between any two observations which is two and one by h is nothing but the sampling frequency so you are collecting observations once every two time points. Hope this is clear. So, the difference between h and 1/h, now again, I think these two points are more like summary points. Just to summarize what we discussed in the previous lecture and now is that, hence, order of all the observations is very important.

So, are the observations ordered or are they not ordered. So, order is very important. For example, here X_t is the first observation. X_{t+2} has to be the next observation. X_{t+4} has to be the next one.

So, you can't mix and match the ordering. So, one can't say that X_{t+2} is today's temperature. That won't be correct. Right? So, X_t is today's temperature. X_{t+2} is day after’s. X_{t+4} is 2 days after that. And so on. So, order is really important here. And hence observations are also dependent. And then dependent on what again? So, dependent on this time frame t. Okay.

I think now we'll discuss lot many examples just to understand the fact that where all can time series be applied. So again, if you remember in the last lecture, we discussed a list of different examples coming from different areas. Just to summarize it quickly let's say areas like finance or student enrollment, education, sales data, climatology, temperature, rainfall etc. But I think now we are in a position to kind of see how the data set behaves, particularly. So, the first one is a very interesting example. It gives you carbon dioxide levels in the atmosphere.

So, over all these years. So, if you see here, we are starting roughly in 1960 and then we are kind of ending somewhere around 2020, here, which is somewhere here. And this is exactly how the CO_2 levels in the atmosphere are progressing. So, couple of points which immediately come to mind is that you have an upward trend here. So, you have an upward trend which is very evident from this plot. And there is something extra here, which is, you can see repetitions, which is called a seasonality. So, the other one, apart from trend is called a seasonality.

So, you have both trend and seasonality. So, carbon dioxide levels in the atmosphere might be seasonal due to some underlying phenomena. So, trend and seasonality are one of the two aspects of any time series which are really important. So, in order to understand, in order to analyze, any time series data one has to actually observe the underlying trend and the underlying seasonality as well. The second example that we have is on oil spot price in dollar per barrel.

So, these are crude oil prices and then on the X axis again you see the time frame. So, you see all the years here. So, 1, 2, 3, 4 up to let us say 25 which is somewhere here. And this is exactly how the oil prices behave. Now, one important thing to note here is that if you look at this overall picture, you don't see any major trend which is there overall as opposed to the previous example.

So, the CO_2 levels in the atmosphere, you had a clear overall trend. But here one can't see any overall trend. So, for example, one can observe trends in some small batches. So, one can see trend here or one can see a downward trend possibly here or one can see a small upward trend again here. Right. But one cannot see an overall trend and seasonality is also not present. Right.

So, there is no overall trend and there is no seasonality in this example. Now, the third example is again coming from India. So, we have SGST revenues of India. So, these are the state GST values, the tax values and then revenues generated from SGST in India over all these months. So, roughly we are starting in somewhere in August of 2017 to let us say August of 2020 which is here.

And again, this is a very typical example of how the time series data of SGST revenues behaves. So, again here one can see that one might have an overall mild trend, right. So, earlier the revenues were not that much. So, somewhere around let us say 22,000 mark. And then down the line, you can see that the revenues are kind of increasing in recent years, let's say somewhere around 28,000 mark. So, one can actually observe a very mild trend which is there present overall.

Again, the next example is very interesting, which gives you the Delhi's air quality index between 2017 to 2020. And immediately one can see that you have a seasonal time series. So, how do you gauge on seasonality? So, seasonality means there should be repetitions. So, if you see that the air quality is kind of increasing, then dipping again, then again increasing up to same levels, then again dipping, then again increasing up to same levels and then dipping.

And all the peaks, if you observe, are at constant level. So, seasonality is very much present here. You don't see an overall trend per se, but again, one can see mild trend in batches. So, for example, you can see a trend here or you can see a trend here or you can see a downward trend here, etc. Okay, so you don't have an overall trend here, but seasonality is present because you see repetitions which are kind of constant over all these years.

And one can actually observe some mild trends in batches. The next example is quarterly sugarcane prices in India. So again, a very simple example. So, on the Y axis, you have sugarcane prices in India. And on the X axis, you have all these quarters.

So, the first quarter, the second quarter, up to the 140th quarter. And again, one can actually immediately guage some behaviors of the data sets. And let us say you might have a very, very mild trend here. Of course, the trend is stronger here in this period. And again, you have a downward trend towards the end.

Now I'll talk about one important point. So, what exactly is the difference between this application or this example and let's say the previous one. So, the previous example, the data set was monthly, right, over all these years, of course, but here the data set is quarterly. So, immediately one can sense that depending on the problem statement or depending on the application area, one can actually change the timeframe. So, should you go with monthly or should you go with yearly or should you go with daily or should you go with quarterly and so on.

So, for example, sales data. So, sales data can be talked about either in daily or quarterly or semi-annually, etc. Or let us say temperature data. So, temperature data can be talked about again at a daily level or weekly level, monthly level, etc. So, understanding the time frame is really important. So, it in fact points to this slide as to how should one choose the time scales or how should one pick the correct time scales.

So, these days what has happened is due to modern technology, one can actually record data on much more frequent time scales, right? So, let's say a few examples could be stock data are available at ticker level. So, ticker means whenever an order is placed. So, let's say either a buy order or a sell order, a tick happens on that chart. So, we'll say that stock data is available at a ticker level.

So very, very minute time frame. Or let us say online and in-store purchases are recorded in real time. So, as and when a purchase happens, the data entry is done. So, due to modern technology, I can actually record data on a very frequent time scale also. Hence, when it comes to picking the correct time scale, one must consider the scale of the required forecasts.

This is important. So, what do you want to achieve? What is the goal? Or how do you want to see the forecast to happen? So, do you want a daily forecast or do you want a monthly forecast, right?

Or do you want a weekly forecast or do you want an annual forecast? So, depending on the scale of the required forecasts, I should correctly pick the time scale. And one should actually observe the level of noise in the data. So, noise is again a very particular terminology when it comes to time series, okay? And probably I'll explain right now as to what exactly do you mean by noise in a minute or so.

But again, just to summarize that if you want to pick the correct time scale, you should consider two things. So, scale of the required forecasts and the level of noise in the data. So, an open question to all of you. So, think maybe for half a minute or something like that. And then trying to answer yourselves that if you want to forecast the next day sales at a grocery store,

would you use minute by minute sales data or daily aggregates? So, what do you think? The question is to forecast the next day sales at a grocery store, right? So, would you pick a minute-by-minute sales data or daily aggregates? So, I'll tell you the answer now.

So, the correct answer is daily aggregates, of course. And why exactly? Because if you spend your time in collecting minute by minute sales data, it would contain a lot of noise in the data. Firstly, we'll explain as to what exactly you mean by noise.

But the other thing is, since you are focusing the next day sales at a grocery store, why would you even concern about minute-by-minute sales data? Isn't it? So, using the daily sales data or daily aggregates from, let's say, the past month or past half a year or past several months, it would be very apt to kind of forecast the next day sales at that grocery store. Now I will kind of tell you exactly as to what do you mean by noise in the data set. So, a noise in the data set I will give you a very simple example, means that if you have lot of random repetitions in the data.

So, let us say something like this, right. So, let us say if you have a time series observation where you have lot of fluctuations in the data set then we will say that you have lot of noise in the data. So, noise means, where you don't see any pattern, it's completely random and you have lot of fluctuations. So, let's say the entity is either going up or going down very rapidly in a very short time frame. So, let's say these might be days.

So, let's say day 1, day 2, day 3 and day 4. So, over the course of 4 days itself, you have lots and lots of fluctuations. So, we'll say that you have a lot of noise in the data. So, again try to picture this idea of noise in the data with minute-by-minute sales data. So, don't you think that if you have minute by minute sales data it would contain lot of noise in the data.

So, answer is yes. Of course. Right. Because the minute-by-minute sales data might look again, something similar to this, where you have a lot of fluctuations. So, let's say the next minute you have a couple of sales. Then again, in the next minute, you don't have any sales. So, it dips down. Right. So, it keeps on fluctuating and which is kind of unwanted for us. Right. If the ultimate goal is to forecast the next day's sales at a grocery store.

So, looking at minute by minute sales data won't be useful. Now, the next idea that we'll discuss is two kinds of time series. So, the first one is called as long time series and the other one is called as short time series. So, idea is very simple if you try to understand. So, what exactly you mean by long time series?

So, the underlying time series contains lot of observations. So, I will give you some examples. So, let us say weekly interest rates over 5 years. So, let me complete this statement over 5 years. So, let's say if you take 52 weeks in a year, so roughly around 250 observations, which are a lot of observations or let's say daily closing stock prices again, let's say over five years.

So, lots of daily closing stock prices over the next 5 years. The third one could be electrical activity of the heart measured at every millisecond interval. So, again lots and lots of observations even here. So, small definition which is again a simple one is long time series. So, any time series where you have lots of observations.

And as opposed to that what do you think a short time series means? So, short time series means, the time series does not contain a lot of observations. So, for example, census data for India starting in 1990. So, roughly around 30 observations, right? If you talk about, let us say, yearly data set, if you talk about, let us say, 10 yearly collections, then even lesser.

So, only 3 per se, right? But even if you consider yearly data, then again, it is not more than 33, right? Or let us say annual student enrollments at a college since 2000. So, again annual say you are talking about yearly values and starting in 2000. So, somewhere close to 20 observations only.

Or let us say daily stock price for a stock over a month. So, again since you have lot of holidays in a month. So, we have roughly around 22 to 23 trading days in a month. So, these many observations in a month. So, distinction between long time series and short time series is a simple one.

So, if you have lots of observations in the data set, it's called as long time series. If you don't have those many observations, if you just have a handful, then it's called as a short time series. So, now we'll spend some time on a particular example here, which is the Indian population between 1955 to 2024. So, just to understand the fact that how can you play around with some actual data? Now again as we discussed in the first lecture there will be some practical component in between also where we will deal with some actual data sets and then try to implement them in R.

So, for example here if you talk about the population of India which is X_t and then the first year is 1955 then second time frame is 1960. So, the data collection happens in every 5-year time intervals. And then the last time frame is 2024. So, I think after 2020, so let us say somewhere here in between you have 2020. And then you have the data collection which has happened this year.

So, the first population value was that much. Then the second population value rose. And now currently we are sitting here. And this is exactly how the data set behaves, right? So, population in India between 1955 to 2024, the plot of the data set along with all the values.

So, a small note, if you see here, that trend is present obviously. So, you have a trend in the data and the data set is called as non-stationary time series, right? So, I think in the next lectures, we'll spend some time on describing as to what do you mean by stationary time series? What do you mean by non-stationary time series? But again, if you remember, generally speaking, 90% of the applications are non-stationary.

For example, here, what makes this data non-stationary is the trend aspect. So, since you have a trend here, So, trend is making the data set to be non-stationary per se. Or even if you have some repetitions, let's say seasonality. So, seasonality can also make a data set to be non-stationary.

So, I think we elaborate the difference between stationarity and non-stationarity in subsequent lectures. Now on the other hand, one more very interesting example is international airline data. So, here, you see a plot of the data. And what it gives you is it gives you the monthly totals of international passengers between January of 1949 to December of 1960. So, over all these years and all these months, you have the monthly totals of international passengers.

And right up front, if you notice a couple of things that you have a trend which is present, right? So, the data is generally upward sloping. So, you have an upward trend and you also have seasonality, right? So, you see those peaks and troughs. And one can actually assume that since it's an international airline data, so one might encounter more passengers in summer or winters when you have vacation times and not so many passengers in between.

So, data set is kind of seasonal in that sense. However, there is one more aspect which is very important in this example is if you look closely here in the third point which is variability increases with time. So, if you notice, what is happening is initially the peaks are kind of not that tall but as you go down the line the peaks are kind of taller as you progress. This means that if you hypothetically draw some bands, let's say here and there, then one can actually see that the variability is kind of increasing. I'll give you one more example just to appreciate this point that what do you mean by variability increasing over time?

So, let's say if you have a data set which behaves like that, right? So, now can you see that initially you have lower peaks and then as you go down the line, the variance in the data, or the deviations are kind of increasing and then the same thing is being observed here also. So, in a sense this example is kind of a complete or classic example of a non-stationary data set where you have an upward trend, you have seasonality and the variability also changes. So, just to summarize as to where we sit, that the main difference between time series and any other statistical sample is that the time series observations are dependent, the time series observations are ordered as we discussed even in the first lecture. And the second aspect is what should be the goal of any time series analysis.

So, the goal of any time series analysis is twofold. So, explanatory and predictive. So, what do you mean by explanatory? So, explanatory means understanding the stochastic behavior. Stochastic means something that changes as per time or something that behaves as per time.

So, we call that as stochastic process. So, time series is a stochastic process. So, understanding or modeling the stochastic mechanism that gives rise to an observed series. This is the first fold. And then second fold is obviously forecasting. So, forecasting the future values of a series based on some history of the series or based on historical data.

So, again just to summarize very quickly the goals of time series are twofold, explanatory and predictive. Explanatory means understanding the hidden model or the applied model on the time series and predictive means forecasting something in the future. So, now I think the next lecture we will talk more about stationarity. We'll talk more about some very basic models of time series. We'll talk more about non-stationarity, the difference between stationarity and non-stationarity and so on through some examples and of course using some practical applications in R. Thank you.