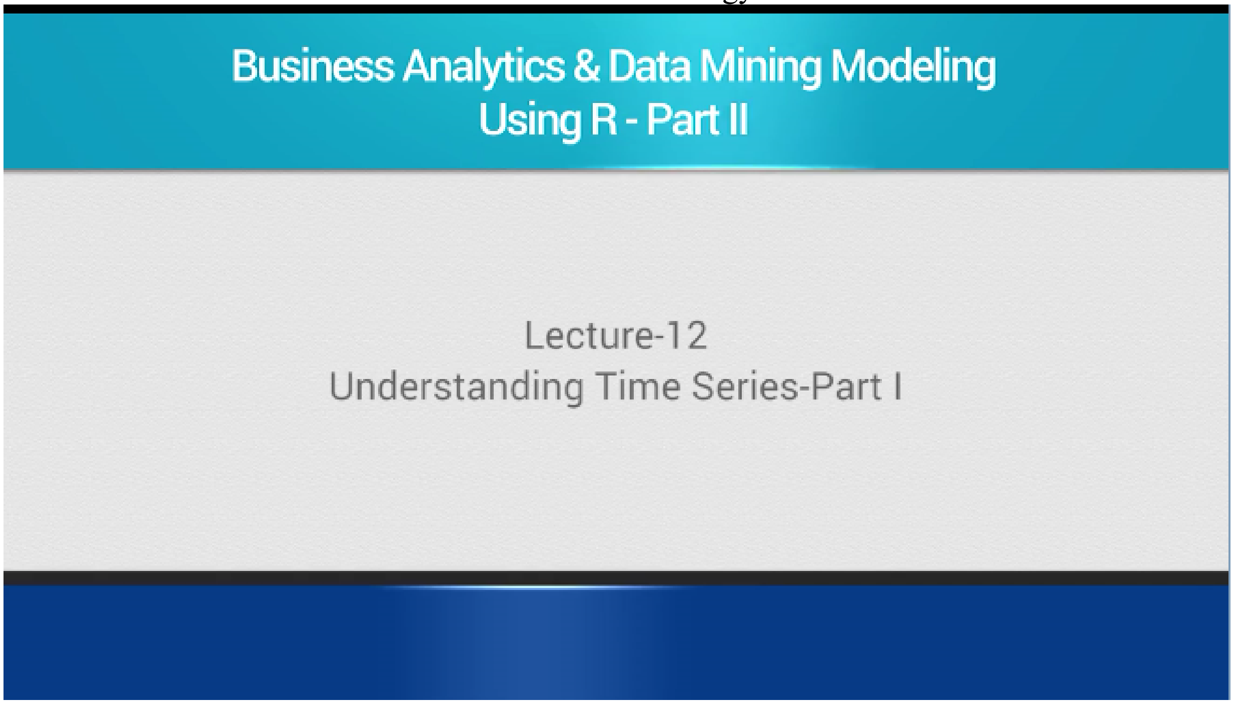
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
NPTEL
NPTEL ONLINE CERTIFICATION COURSE
Business Analytics & Data Mining Modeling
Using R – Part II
Lecture-12
Understanding Time Series– Part I
With
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology Roorkee



Business Analytics & Data Mining Modeling
Using R - Part II

Lecture-12
Understanding Time Series-Part I

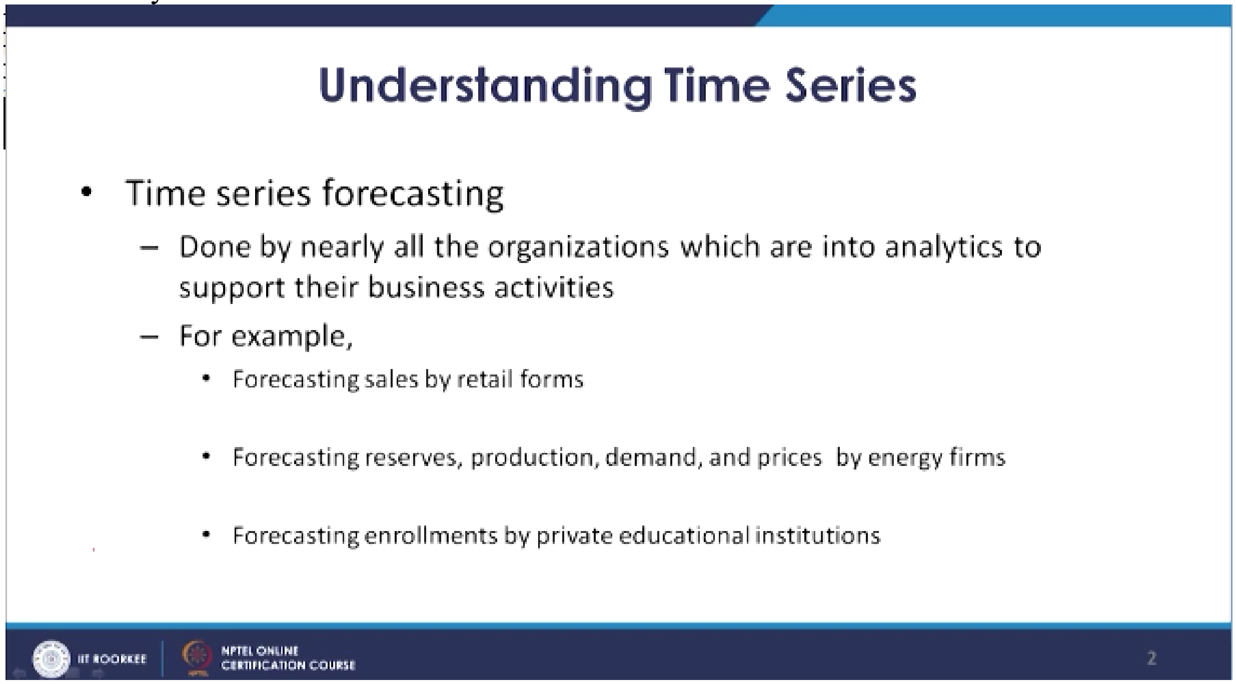
With

Dr. Gaurav Dixit

Department of Management Studies

Indian Institute of Technology Roorkee

Welcome to the course Business Analytics and Data Mining Modeling Using R – Part 2, so in previous few lectures we have been able to complete our previous module that is unsupervised learning methods, so where we were able to cover two techniques, association rules mining and cluster analysis as well.



Understanding Time Series

- Time series forecasting
  - Done by nearly all the organizations which are into analytics to support their business activities
  - For example,
    - Forecasting sales by retail forms
    - Forecasting reserves, production, demand, and prices by energy firms
    - Forecasting enrollments by private educational institutions

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 2

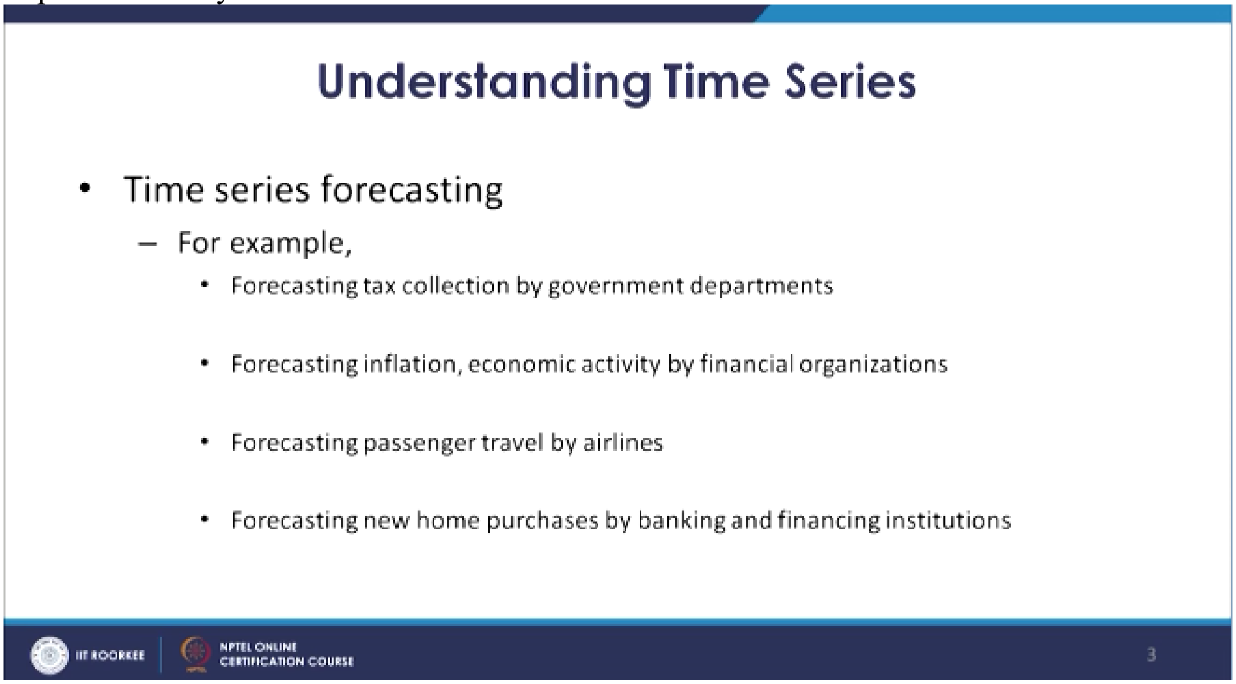
Now we are going to start our second module and the last module as well which is time series forecasting, so therein first we'll start with you know first topic is going to be on time series understanding time series and different components of time series, so this is named as understanding time series, so we'll start with this one, so the main in this particular module, time series forecasting we are typically going to focus on time series forecasting as such, and some other aspects of time series are not going to be covered in great detail in this particular module, so most of the discussion is going to revolve around how you know time series forecasting is performed, so let's start our discussion on this, so time series forecasting typically this is done by nearly all the organizations which are into analytics to support their business activities, so how we can say this, so let's have a look at few examples, so some of the activities is specifically the analytics you know task that these organizations have to perform or like this, for example forecasting sales by retail firms, so as we know that most of the firms which are into manufacturing something or producing something, you know sometimes they also have detail outlets so they would be doing this business activity as well which is about sales, selling the manufacture products, right, or units so in that process if they are also into the analytics so they would also, we helped in this particular business activity, if they are able to forecast sales, right, because they would be able to you know find out how many units they need to manufacture or produce or purchase, you know purchasing of raw materials and other things would also depend on their forecast of how many units they would be required to manufacture or assemble or whatever format they do it, so in that sense forecasting sales would be an important analytics task for them to be able to you know complete their business activity in a more efficient manner, so this is one.

And with respect to forecasting sales itself as I said other materials, raw materials other you know items or material which would be part of that producing or manufacturing production or manufacturing process would also you know there you know planning would also depend on this particular sales, so because the market itself, the demand and supply aspect of market that itself you know keeps on changing, so depending on the kind of demand that is going to be there in the next year and the kind of sales that a particular company has in that market and whether they would be able to increase that sale, so if they think that they would be able to increase their sale they would like to manufacture bit more you know to fulfill that particular demand, so overall looking at the market demand economic you know, economic environment as such manufacturing production environment as such, they would have to look at the kind of you know markets share, the kind of percentage of the you know market demand they can fulfill and therefore accordingly they will have to plan their operations, for that they need to find out how many units they need to manufacture or you know produce, and there planning's for other raw materials and other items would also depend on this, so all those you know all this things that I just talked about they might in a way depend on how you know, what is going to be the forecast for next year, sales forecast for next year.

So you can see this simple analytics task of forecasting sales by retail firms is actually have you know multifold impact on other operations of those firms as well as their supply chain partners and other partners.

Similarly if you look at the second example that is about forecasting reserves, production, demand and prices by energy firms, so in the energy sectors if you look at this because the countries and the firms which are into producing you know petroleum products, so they influence they control that supply of petroleum products and other energy you know products, so therefore for energy funds to be able to maintain their market share or even increase it they need to, they should be able to forecast few things, for example reserves that they have, the production you know their own production that they can plan out, the demand in the market and the prices that they can you know, that they can charge for their products, so all these things you know they are so much dependent on market factors and other internal, external factors, that if they are able to forecast their planning and execution is going to be much better, they would be able to you know maybe, there is a good chance that they would be able to you know increase their market share as well on and more profits or increase their revenues, so this simple analytics task, if a company is into analytics and into energy domain so their ability to perform this kind of analytic task would go in a you know gateway in terms of helping their core business activities.

Similarly if we look at the third example that is forecasting enrollments, private educational institutions, so as we know in our country typically in government educational institutions the seats are fixed, the number of seats are fixed, however you know private educational institutions they can always plan for this, and for them it’s going to be a challenging task, in terms of you know forecasting enrollments, the students that would be enrolling for the institute because there is competition within the private you know this educational sectors, and the number of enrollments fluctuate year on year depending on the you know various factors, performance of the students belonging to those educational institutions, their infrastructure, their ratings, and you know other things, right, so because of this their planning, their staff arrangement, infrastructure arrangement all that, will is going to be dependent on this particular analytics task, so if they are able to forecast this number they would be able to plan out their other activities, other supporting activities in much better fashion, so for them also this is an important activity.



Understanding Time Series

- Time series forecasting
  - For example,
    - Forecasting tax collection by government departments
    - Forecasting inflation, economic activity by financial organizations
    - Forecasting passenger travel by airlines
    - Forecasting new home purchases by banking and financing institutions

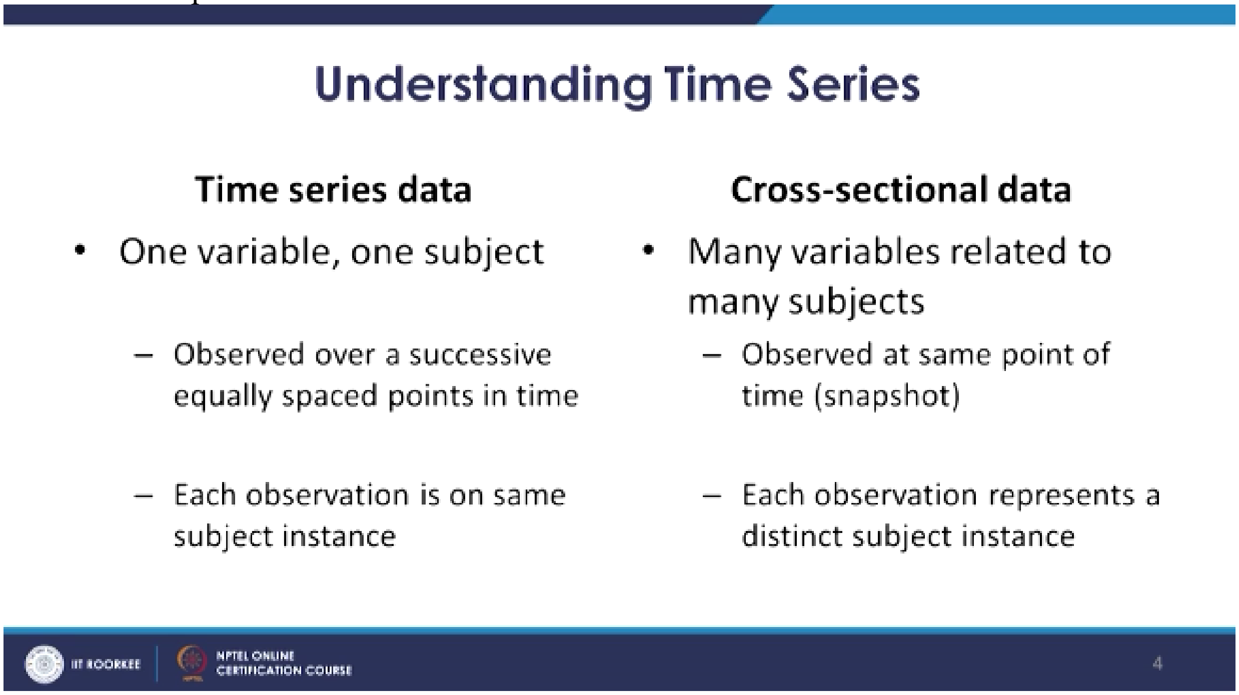
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 3

So similarly we can talk about other examples as well, for example the forecasting tax collection by government departments, so the next year what is going to be the increase in tax receipts, tax collection that is important for government departments, because they also have to be ready with that amount of you know paper work or even if it is in digitized format, that amount of processing would still be required, so for that all that preparation and planning and execution if they are able to forecast tax collection you know then this is going to help their activities in a much improved manner.

Similarly we can look at other examples, for example forecasting inflation economic activity by financial organizations, so there are various organizations which keep on forecasting different macro level economic factors about you know developing countries, emerging countries, developed word and overall globally, and there is always hype around these numbers in ever space of our society, so this is also an important activity because this gives an indication to you know, to the government and other organizations how the economy is doing, so inflation economic activity forecasting this becomes an important you know element for this organizations.

Similarly if we look at you know the transport sector, so now this example forecasting passenger travelled by airlines, so as we understand that airlines the kind of you know traffic that they observe you know keeps on changing year on year and also within a year there are so you know, there are seasonal cycles, there are times where that you know peak them, they were a times of peak demand then you know, there are times where you know there were off seasons also, so if they are able to forecast the passenger travel then of course in that fashion they would be able to plan their other activities accordingly.

Similarly if we look at the last example in the slide, the forecasting new home purchased by banking and financing institutions, so if we look at this banks and other financing institutions are able to forecast you know the, you know the amount of home purchases, new home purchased that are going to take place in the next year, so accordingly they can do their budget planning and other core activities, planning of other core activities and execution, so if we look at all these example that we have till now talked about, you know they vary across different industrial domains and a firm ability to be able to forecast these numbers with higher accuracy and reliability would in go in a great way to help in terms of planning and executing their core and other supporting business activities, so that is the importance of forecasting time series which is the topic of our discussion in this module.



Understanding Time Series

Time series data

- One variable, one subject
  - Observed over a successive equally spaced points in time
  - Each observation is on same subject instance

Cross-sectional data

- Many variables related to many subjects
  - Observed at same point of time (snapshot)
  - Each observation represents a distinct subject instance

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 4

So let’s move forward, so now we are going to discuss few other important aspect about time series forecasting, so let’s start with the kind of data set that is used in time series forecasting, so here we are going to start with a comparison between time series data and cross sectional data, because this is going to give us better perspective in terms of understanding how time series forecasting is done and how it is different from what we have learned in the supervise learning methods and unsupervised learning method where we were typically dealing with the cross sectional data set, so in previous modules one module that was covered in the you know previous course that is supervised learning methods and the other module that is unsupervised learning method that was covered in this course, and both this modules typically the kind of dataset that we were using was you know cross sectional data.

So now let's understand the difference between time series data and cross sectional data, in our previous course also we have discussed few aspect of different kind of dataset, so you can refer to that particular video lecture as well, so let's discuss this, so in time series data typically the data is about one variable and one subject, so this you know this variable is measured for one specific you know subject and this measurement or this observations are recorded over a successive equally spaced points in time, so and each observation is on same subject instance, so if we talk about a particular index that is being measured, so that index is going to be measured in successive equally space points in time and if we have data you know, if we have both measurements for a number of you know, for a number of points then that would form a time series.

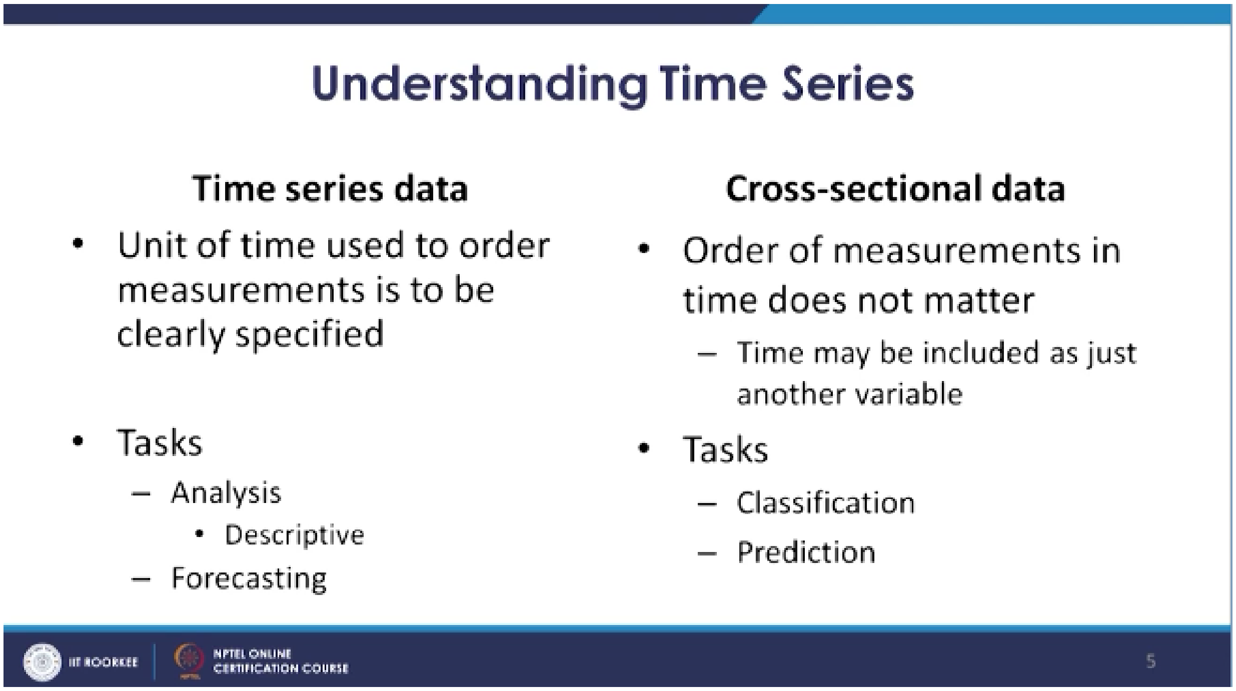
Now if we look at the cross sectional data, so data is on many variables and which could be related to many subjects, so like in previous modules that we have discussed, in all the data sets that we had, the data was captured, data was measured you know all the observations had many variables, so typically the tabular format we had, matrix format, you know dataset format that we have used, and there the rows typically represented the you know observations and columns represented the variables, so from that format itself we can understand that there were many variables and these variables were not necessarily you know measuring something about a particular subject, particular type of subject but it could be related to you know many other things as well, for example if one variable, V1 could be about you know a particular customer, the second variable V2 could be about the organization of that particular customer, the third variable V3 could be about the you know family, about that particular you know customer, so or you know some other activity that you know particular activity which the, you know that particular customer and user might be associated with, so in that sense we can say that many variables are involved, many variables are measured in cross sectional data, and they could be related to different subjects, however when we talk about time series data it is typically one variable and it is about one subject, and the observation in time series data as I said that is successive equally space point in time, what does that mean is so one observation is recorded on day 1, the second observation is recorded on day 2, and third observation is recorded on day 3, or it could be one observation recorded from month one, another observation is recorded from month two, month three, or it could be one observation recorded for year 1, the second observation is recorded for year 2, third observation is recorded for year 3, and all these observation are on the same subject instance.

So the index you know it could be one index or it could be inflation or it could be a you know GDP number, so for example if we talk about the GDP number so if we are recording it year on year you know we could be recording it quarter by quarter, so if we look at these, when we say year on year, so those are equally spaced points in time, when we say quarter by quarter, those are equally spaced points in time, and when we say each observation is on same subject instance, so GDP is going to be of a particular country, so if it is the GDP if we are talking about you know measuring India GDP so at each observation is going to you know measure India's GDP and at different you know, you know equally space points in time, year on year or quarter on quarter, so that is what we you know mean in time series data.

But if we look at the cross sectional data as I talked about, tabular data format or matrix data format where there are going to be so many variables on columns and they could be related to different subjects, in terms of observing these variables in cross sectional data so these are observed at some point of time, so if we are looking at first observation and this first observation is related to a particular customer then customers age at a particular point in time and at the same point of time it, you know income of the customer, at the same point of time that particular you know gender of the customer same, at the same instant of time the organization of that particular customer at the same instance of time, you know any other you know observed variable or perceptual variable that is to be measured for that you know particular customer, you know so all those things are recorded at the same point of time, so all the variables you know that we have in columns they are going to be recorded at same point of time.

And the second observation that we had would be on the you know second customer and it would also be recorded at the same point of time, approximately or hopefully at the same point of time, for example if we are going for you know survey kind of data collection so we planned to finish it within a limited time frame, so the data on all the customer can be the free set has been recorded at same point of time, so we don't make you know such a, you know a difference in terms of you know time, so each observation represents a distinct subject instance, so as I said when I talked about the second observation in the cross sectional data I said the second customer, but when we talk about the time series data it's not the you know second countries GDP, it is the same countries GDP that we are talking about, so it is India's GDP which is recorded for year 1, and then year 2, and year 3, but when we talk about the cross sectional data that is the customer, once you know data on different variables related to different subject then customer whose data on different variables related to the same subjects, so the subject distance is changing, so each observation is representing a distinct subject instance, however in time series data it is the same you know subject which is India, and India's you know variable on India is that is GDP which is being recorded at equally spaced points in time, so successive equally spaced points in time, so I think this would have given you a better picture in terms of understanding the time series data and how it is different from cross sectional data.

Let's discuss a few more points about the same, if we look at the time series data, unit of time used to order measurements is to be clearly specified, so when we say particular time series and the time series is about GDP or India's GDP then we have to, it is to be clearly specify whether these numbers are annual, whether these are yearly numbers for GDP or whether they are quarterly numbers for GDP, so that unit of time that is going to be used to measure a particular variable that is to be clearly specify, so this is one important you know characteristic of time series data.



Understanding Time Series

Time series data

- Unit of time used to order measurements is to be clearly specified
- Tasks
  - Analysis
    - Descriptive
  - Forecasting

Cross-sectional data

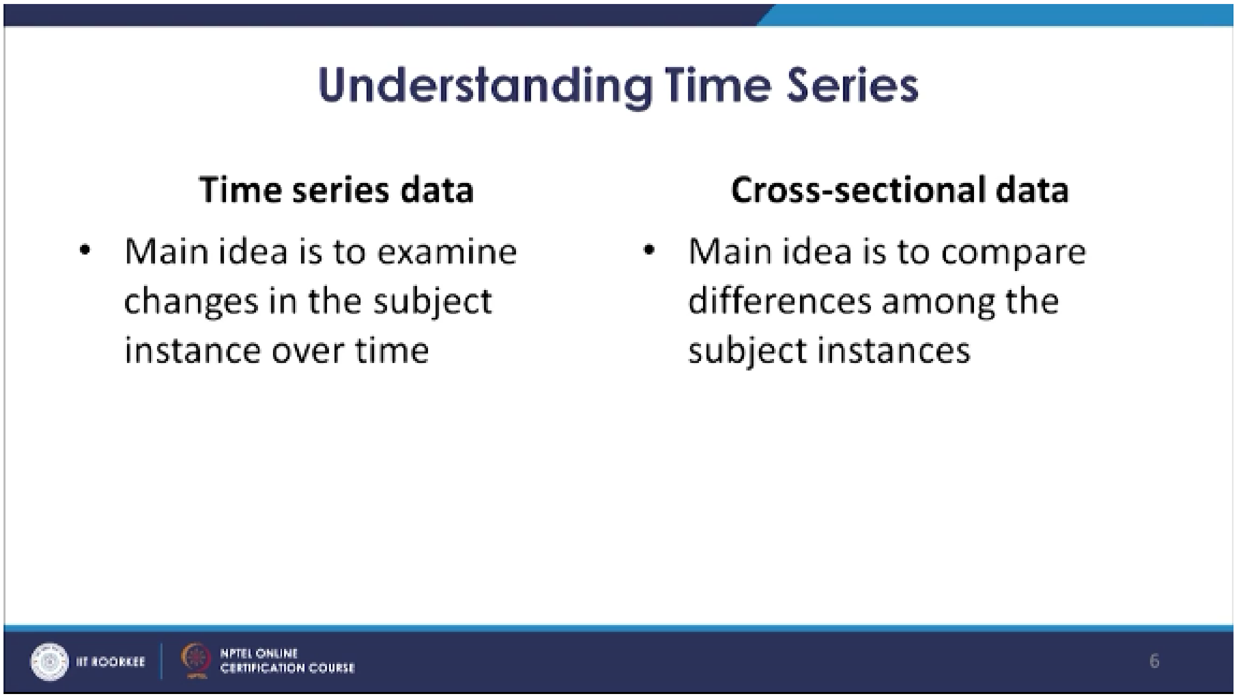
- Order of measurements in time does not matter
  - Time may be included as just another variable
- Tasks
  - Classification
  - Prediction

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 5

However, if we look at the cross sectional data, the order of measurements in time does not matter, so you know when data on a particular you know customer 1 was recorded, and when data was, data on customer 2 was recorded, so we are not looking at that you know difference in time gap, so we are not really valuing this particular difference in time. However, there might be cross sectional dataset where time might be you know being treated as another variable, so you might have cross-sectional dataset where you have you know one column having the customers status, another column having a gender, other columns having data related to other important variables, and then you might also have one column you know saying you know year or some date or something, so in cross sectional data time can sometimes be, if it is part of the analysis that variable is important for the analysis then time could be just another variable in cross sectional analysis.

Let’s move forward so the kind of analytics task that we typically perform on time series data are of two types, so one is where we analyze time series which is strictly called time series analysis, if this particular task is strictly descript you or you know explanatory in nature, then the second one is forecasting, so as I said this module is that we are discussing here is mainly we are going to focus on time series forecasting, however some aspect of you know even while our discussion on time series forecasting you know some aspect are going to be common with what we do in time series analysis.

If we compare this with the cross sectional data and the kind of analysis that is done there is two kinds of task are typically performed, classification and prediction, so as we have seen in the previous course when we discuss many supervise learning techniques, so there typically we were either doing classification or prediction, if we look at the you know, the module, first module that is covered in this course, the unsupervised learning methods so there we had the two more types of task where one was clustering, the other one was association rules you know that rules mining, these are the kind of tasks that are typically done in you know cross sectional data, however if we look at the time series data typically it is about the analyzing it, analysis task or the forecasting task.



Understanding Time Series

Time series data

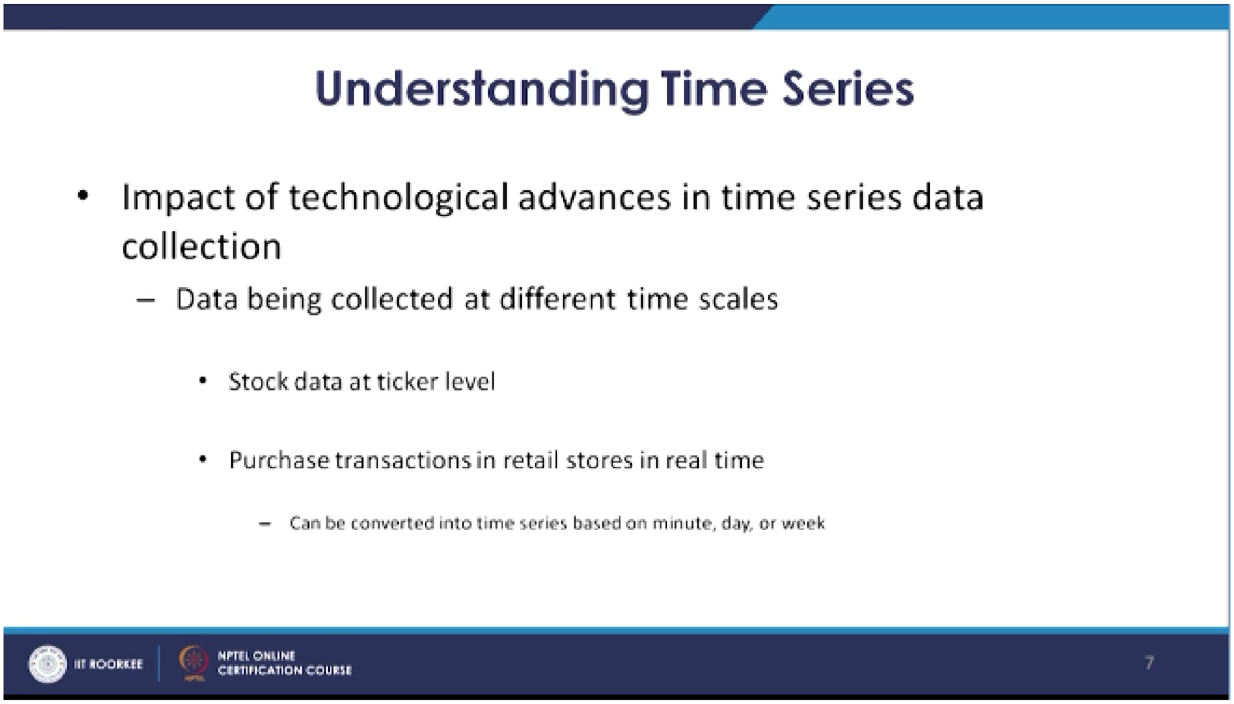
- Main idea is to examine changes in the subject instance over time

Cross-sectional data

- Main idea is to compare differences among the subject instances

Now one more important difference between time series data and the cross sectional data, so for time series data the main idea is to examine changes in the subject instance overtime, when we or you know whether we are doing time series analysis or time series forecasting, so the main idea remains this that we are looking to examine changes in the subject instance overtime that is why India's GDP you know when we look at the time series it is being recorded at you know successive equally space points in time, because we want to, you know understand or analyze the changes that are happening in this particular variable over a period of time.

If we look at the cross sectional data, the main idea is there to compare differences among the subject instances, right, so we are always looking to identify you know the relationship between the you know different variables association between different variables or you know, you know cause effect kind of the casual relationship between variable, so there typically we are comparing you know differences among the subject instances, and that is why if we look at the dataset that we typically use, use in cross sectional analysis, you know each observation is representing a distinct you know subject instance and the generalization is on that particular you know target population which is you know, each observation is coming from that target populations, so different observation is representing either different individual or you know different you know firms, or any other unit of analysis that could be there.

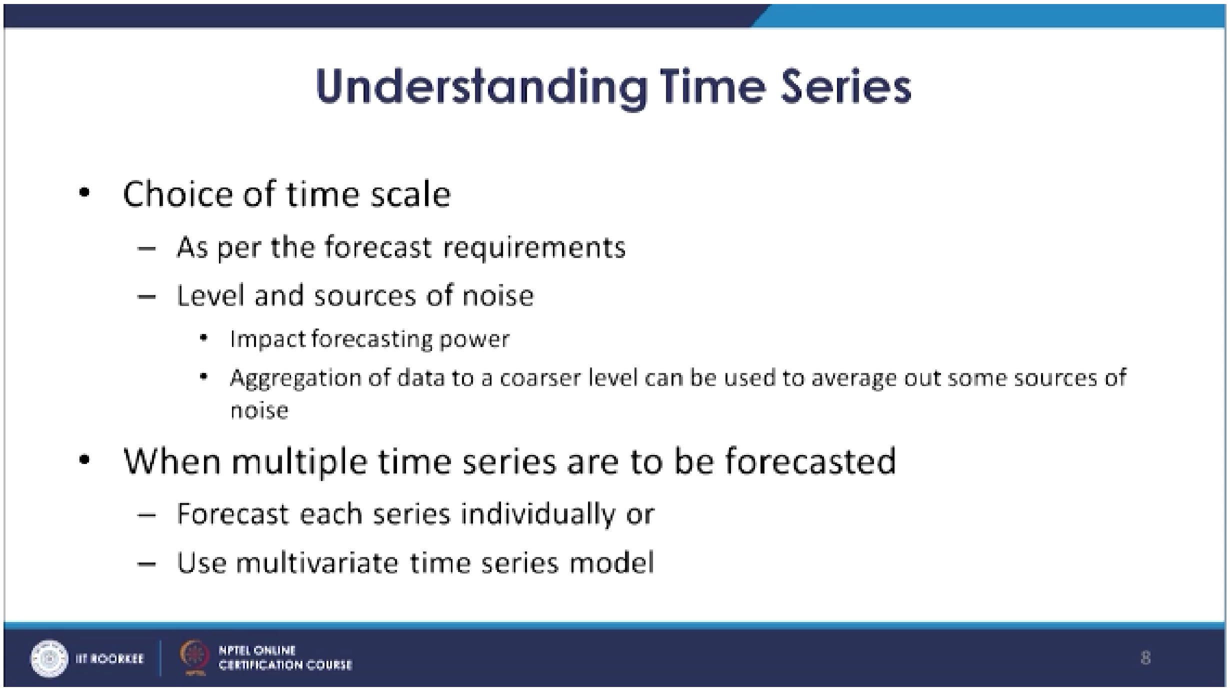


Understanding Time Series

- Impact of technological advances in time series data collection
  - Data being collected at different time scales
    - Stock data at ticker level
    - Purchase transactions in retail stores in real time
      - Can be converted into time series based on minute, day, or week

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 7

Now let's talk about the technological advances that have happened in terms of information technology and different you know advances and how they are impacting the time series data collection as such, so nowadays as we are aware the data that is being collected you know, it can be collected at different time scales, because of the computer systems and the advanced technology infrastructure that we have nowadays, so you can see stock data for example it is you know captured article level, if we look at the purchase transactions, so nowadays when we reach a retail stores so they are being recorded at real time, so now the kind of transactions that are there, so these can be converted into time series based on minute, day, or week, because that is the advantage when you are, when everything is being captured in real time, so that time stamp is always going to be there and therefore we can always convert the data into any unit of time, for any time scale, so all this is possible because of the technological advances, because of the you know bar code scanners that we have, when we visit retail stores, because of the you know automatic data capture mechanisms that are nowadays present and every walk of life, so because of most of the things that we do somewhere some kind of systems are involved, so most of the things that we do are you know digitalization has reached to that level that specially in the business context almost you know everything can be recorded in real time, and therefore it is easy to you know collect the time series data and at different units of scales as well.



Understanding Time Series

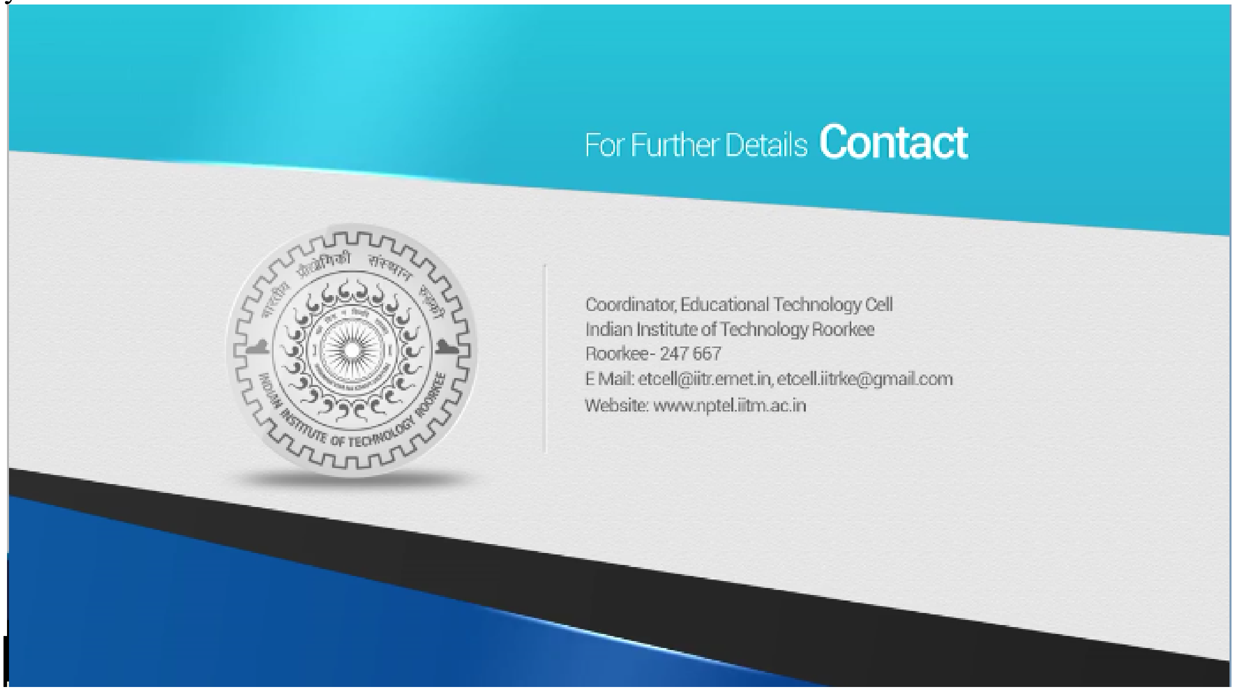
- Choice of time scale
  - As per the forecast requirements
  - Level and sources of noise
    - Impact forecasting power
    - Aggregation of data to a coarser level can be used to average out some sources of noise
- When multiple time series are to be forecasted
  - Forecast each series individually or
  - Use multivariate time series model

IIT ROORKEE
NPTEL ONLINE CERTIFICATION COURSE
8

Now the another important time as we were talking about the impact of technological advances, and the ability to capture, ability to you know collect time series data on at different you know, for different units of time scale, and now how do we decide when this ability is there, right, where we can get any you know kind of time series, you know for different you know even at different you know time scales, so how do we decide which kind of time scale is going to be more suitable for our analytics tasks, so choice of time scale is an important decision to be made before time series forecasting, so there are certain factors which are important for this kind of decision, so first one as per the forecast requirements, so if we talk about the different examples that we discussed at the start of this particular lecture, for example forecasting sales by retail firms, so we'll look at you know that particular example depending on the requirements, so typically you know the way retail firms plan for their operations, if the planning is typically done for you know year on year, the annual, at annual unit of time, annual time scale then probably the forecast requirements you know would also be year on year, so you would like to forecast sales for the next year, right, but if the planning is monthly then probably we would requires sales number for our monthly planning, and so that we can plan their operations accordingly.

So what is the requirement typically that depends on the kind of industry we are talking about, and the kind of firm which might have you know its own internal way of doing business processes, so therefore forecast requirement the kind of unit scale that is to be used, can actually come from there, then the second important point is level and sources of noise, so this particularis, this particular point is more with respect to analytics expertise, so depending on the level of, depending on the time scale that we are using, the level and sources of noise you know can actually you know cancel out or can actually increase, and this will have an impact on forecasting power, so for example another point that is mentioned, aggregation of data to acquires a level can be used to average out some sources of noise, so as we know typically when we think about time series we can always think about a time plot where you know different observations they go up and down, for example stock data if we visualize this then stock data will you know typically swing in surgical fashion up and down, up and down in that fashion, and if we are looking to analyze you know any kind of you know time series data, and if we change the scale and if we aggregate, if we take it to a course or level then we would observe that some of the noise that we see, some of those swings ups and down that we see they will cancel out, they will average out in a longer time period, so in that sense, the noise levels minimize and therefore the forecasting power in a longer time period, in that sense increases a bit, but the kind of forecasting that we require is also dependent on the kind of requirement that we have, so both factors have to be looked at the kind of requirement, forecasting requirements that one has, and then the kind of level and sources of noise that are going to be there and how they are going to impact the forecasting power of the models.

So with this we'll stop here and we'll continue our discussion on this in the next lecture. Thank you.



For Further Details Contact

Coordinator, Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee - 247 667
E Mail: etcell@iitr.emet.in, etcell.iitrke@gmail.com
Website: www.nptel.iitm.ac.in

For Further Details Contact
Coordinator Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee – 247 667
E Mail:-etcell@iitr.ernet.in, iitrke@gmail.com
Website: www.nptel.iitm.ac.in

Acknowledgement
Prof. Ajit Kumar Chaturvedi
Director, IIT Roorkee

NPTEL Coordinator
IIT Roorkee
Prof. B. K Gandhi

Subject Expert
Dr. Gaurav Dixit
Department of Management Studies
IIT Roorkee

Produced by

Mohan Raj.S
Graphics
Binoy V.P
Web Team
Dr. NibeditaBisoyi
Neetesh Kumar
Jitender Kumar
Vivek Kumar
Dharamveer Singh
Gaurav Kumar
An educational Technology cell
IIT Roorkee Production