Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 04
Lecture 02B - Motivation and Overview 4

So now we will talk about what one of you has mentioned. How to deal with this randomness is all about time series analysis of course, you know more than that analyzing and extracting information. So when a process is random, when we cannot predict the outcome accurately; the natural recourse is to list all the possible outcomes. The classical example is a rainfall, if I ask you whether it is right now or you think it is going to rain at 10 o'clock now in the morning; none of us will be able to predict accurately let us admit that, but of course, you can turn to the metrological departments forecast and then do the reverse of it; maybe that is more accurate, but none of us will be able to predict it accurately.

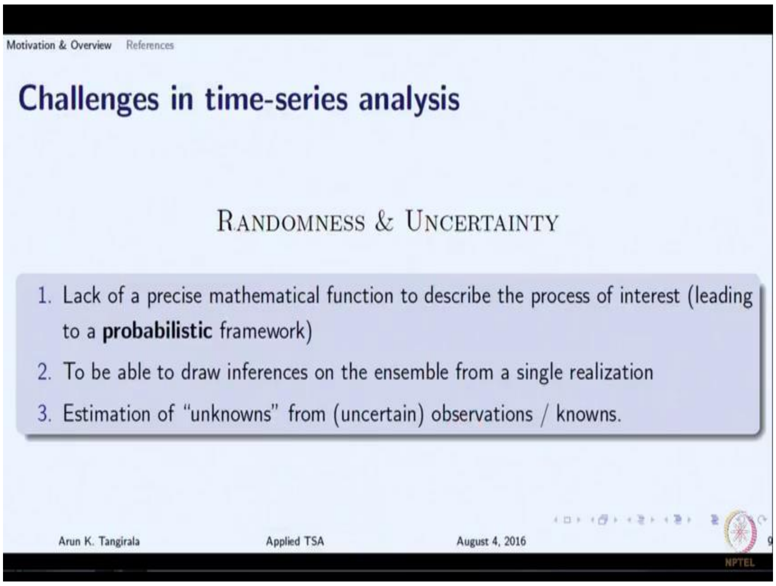
So what do we do, what is a natural way to list all the possible outcomes; you say it may rain or may not rain there is nothing like half rain and so on. So, we have listed two outcomes and then the next step is to assign some chances; you say there are 50 percent chances or 60 percent chance that it will rain or not rain and so on. This is the central idea in probability theory using probability theory for forecasting; listing the outcomes and assigning chances or probabilities to those outcomes and that is why we take asylum in the probabilistic framework, we move away from the deterministic world and move to a to a probabilistic world. And of course, many of us are scared about probability theory, but that is the basic idea then comes notions of probability distributions random variables and so on.

So, the main challenge in time series analysis is randomness which leads us or forces us to take recluse in the probabilistic framework and this so called uncertainties, we will talk about the sources of uncertainty shortly, but this entire probabilistic framework is what makes a subject very challenging, this entire uncertainty in not only in the process, but in the data everything that we observe that also makes the subject very challenging and the heat of this challenge is felt in the analysis that is when you have data with you,

you have a record of data and you want to draw some inferences of the process that is actually generating the data.

Why is the heat felt there? Because what you are observing as you will learn in the theoretical part, what you are observing is only one of the millions of possibilities that could have occurred. We call that single observation as realization; I will show you one example shortly, but to be able to draw inferences about this process which can produce many outcomes, from a single record of observations it is only one of the millions of possibilities that you have is the biggest challenge correct and you will feel that when you get into the practicing aspects of time series and then you also face challenges when you want to estimate certain unknowns.

(Refer Slide Time: 03:30)



Motivation & Overview References

Challenges in time-series analysis

Randomness & Uncertainty

1. Lack of a precise mathematical function to describe the process of interest (leading to a **probabilistic** framework)
2. To be able to draw inferences on the ensemble from a single realization
3. Estimation of "unknowns" from (uncertain) observations / knowns.

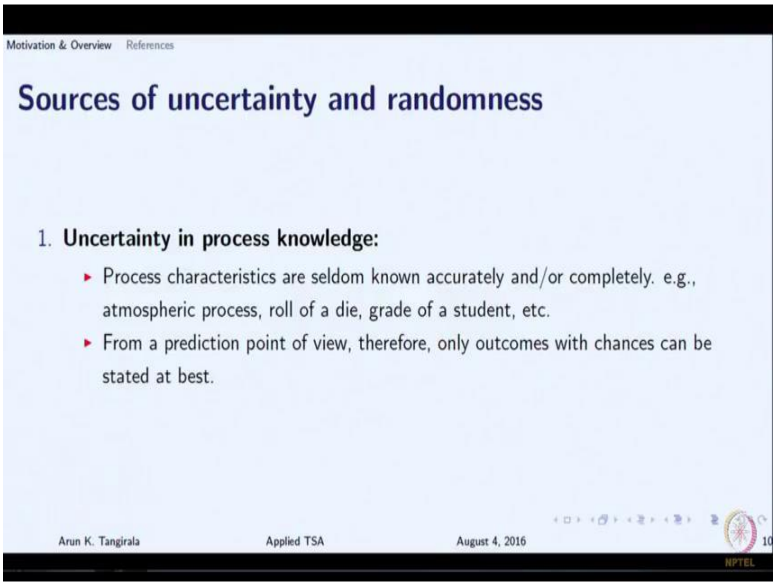
Arun K. Tangirala Applied TSA August 4, 2016 9

What we mean by unknowns here? Are for example, as you must all be knowing any random process is characterized by mean or average or variance and variance or standard deviation and so on, certain statistical properties we call them a statistical parameters or when you are building models in your fitting models to data, you want to be able to estimate the model parameters which are unknowns from knowns. What are the knowns? That you have that; you have you have essentially observations with you, you have data with you. From this data, we are supposed to be able to estimate model parameters and that becomes very challenging. See estimation is at the heart of time series analysis and therefore, in this course we will spend considerable time in learning estimation theory. I always been very vocal about this; that in none of the engineering curricula not only in the country, but across the globe today as it stands engineers are not trained in estimating parameters.

Maybe there is a course or two in electrical engineering or you know maybe somewhere selectively or accidentally in some the discipline, but by design there is nothing in the curriculum that trains the engineers to estimate parameters, amidst uncertainties; that I mean estimation itself is challenging because of uncertainty, if everything is deterministic; it is easy. As an engineer whether you follow engineering or whether you go to you know pursue an MBA or whatever line that you pursue ultimately at some point you will have to start making some estimates, you will have to estimate something. Therefore, there is no escape from estimation theory and perhaps this is the only course as it stands right now which exclusively talks about estimation theory and so on.

So, you should actually give a lot of attention to estimation theory because it is at the heart of time series analysis. Ultimately what does analysis constitute; estimating something; you want estimate patterns or you want actually infer some features of the data or you want to estimate model parameters and so on. Everywhere estimation theory comes into picture and therefore, you should be well versed with that.

(Refer Slide Time: 05:56)



Motivation & Overview References

Sources of uncertainty and randomness

1. **Uncertainty in process knowledge:**
   - Process characteristics are seldom known accurately and/or completely. e.g., atmospheric process, roll of a die, grade of a student, etc.
   - From a prediction point of view, therefore, only outcomes with chances can be stated at best.

Arun K. Tangirala Applied TSA August 4, 2016 10

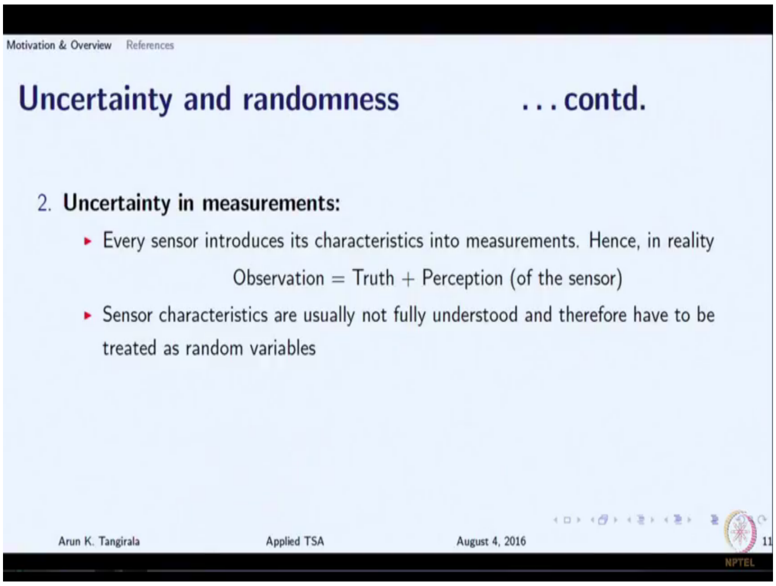
So let us move forward and ask what are the sources of uncertainties that generally prevail and then we will talk about the notion of realization. So, what are the sources of uncertainty and randomness; typically there are three sources of uncertainties that one can encounter, whether all these three apply to your process or not depends on your situation. But the top most source of uncertainty is the uncertainty in process knowledge that is in knowing how the process is behaving, how the process is actually evolving in time; maybe for a few processes I can exactly describe, maybe I can use conservation laws like laws of conservation of mass, conservation of energy or momentum and so on and write a few equations we will be able to describe what is happening, but for a large class of processes, we will always have uncertainties troubling us. In some cases we have a sound knowledge, but not a full knowledge; in some cases we have very poor knowledge.

But the reality is that you would have uncertainties in process model. So, this is the first source of uncertainty alright and of course, it actually prevails in different forms and so on, we will not go into that. From a prediction point of view what this means is that; most of the processes that we encounter we will not be able to predict accurately and that may come as a shocker for many who have been trained to think that yes I can write a model for heat exchanger, I can write a model for a reactor or I can tell you how the flow goes through a pipe or how a spring mass system behaves, how a circuit behaves; I have been learning so much.

Now somebody is coming and tell me no no no you know; that is not necessarily accurate that is what I meant yesterday by saying that we are trained to look at processes it as an exact process as I know exactly, it is not true. There is always something that will evade your understanding; question is how much, what part of the features you are able to explain with mathematics and the rest is all uncertainty. In time series analysis, we assume that there is no way I can write a first principles model, we begin by that assumption and then of course, gradually if you think that no there is a part of the process that I can, there is certain phenomena in the process which I can explain using mathematics that is your differential equations or algebraic equations and so on, yes you can do that and that comes under system identification.

But when you look at time series analysis per say, you are solely relying on data to tell you what is happening in the process and the model that you fit is not going to be necessarily anything that you would have developed from physics, it is a data driven model that you are going to build alright and that is a hallmark of time series analysis.

(Refer Slide Time: 09:06)



Motivation & Overview References

Uncertainty and randomness ...contd.

2. **Uncertainty in measurements:**
   - Every sensor introduces its characteristics into measurements. Hence, in reality

     Observation = Truth + Perception (of the sensor)
   - Sensor characteristics are usually not fully understood and therefore have to be treated as random variables

Arun K. Tangirala Applied TSA August 4, 2016 11

Then of course, you have the next source of uncertainty which is in measurements, so to top this uncertainty in process knowledge, you have uncertainty in measurements. We just now said because I have uncertainties in process knowledge and because I shall believe that I have no idea about the physics or chemistry or the biology of the process, I am going to rely on data, I am going to perform experiments, collect data. Either I or someone would be collecting data, but ultimately I rely on data. Now the bad news is that this data will come with an additional source of uncertainty, this is because of the observational error because of the error induced by the sensors and that also makes life interesting complicated nice and so on.

So, the fact is any observation is truth plus a masala always; I mean news channels are classic examples of this. When you look at news channels; obviously, you can see that there is a truth plus some masala; I mean somebody would say nation would like to know or somebody would say that I would like to know or whatever, but basically it is all the views that are given out by the journalists, apart from reporting what is happening at the site of the event and this is inescapable, it is this perceptional error that you have is not only when humans observed, but also when you have sensors that are manufactured; this hard sensors that are being used to sense.

For example, if you take the thermometer that we use in our daily lives at our homes to measure the body temperature; that also has an error in it, do not confuse error with the resolution of the instrument, but that also has an error of course, there again you can ask whether the error is more or large or less and so on, but the fact is there is going to be an error. Which means that for the same process, exactly the same process if you were to use two or more sensors; each sensor will give you a different reading, slightly different reading, even if the sensor is manufactured by the same manufacturer.

Why do you think that happens, what is the source this uncertainty that is coming up in the sensor reading, what is the source of this error and why are the errors going to be different across different senses even though they are manufactured under same conditions by the same manufacturer; what is the root cause?

Student: Operating condition might differ.

Operating conditions might differ; no I am going to have the same process, so imagine that I am going to measure the room temperature here; in this room or any other room I am going to choose a spatial location and I will use three different sensors temperature sensors alright. We all agree that those three temperature sensors will give us numerically different readings correct, they may differ in the first decimal, second decimal but the fact is they are going to be different. I am going to choose these three sensors from the same manufacturer; manufactured on the same day perhaps in the same hour and still I will get different readings; why is that.

Student: Manufacturer (Refer Time: 12:28).

Why manufacturing process considered random.

Student: (Refer Time: 12:40).

Better answer; the time varies.

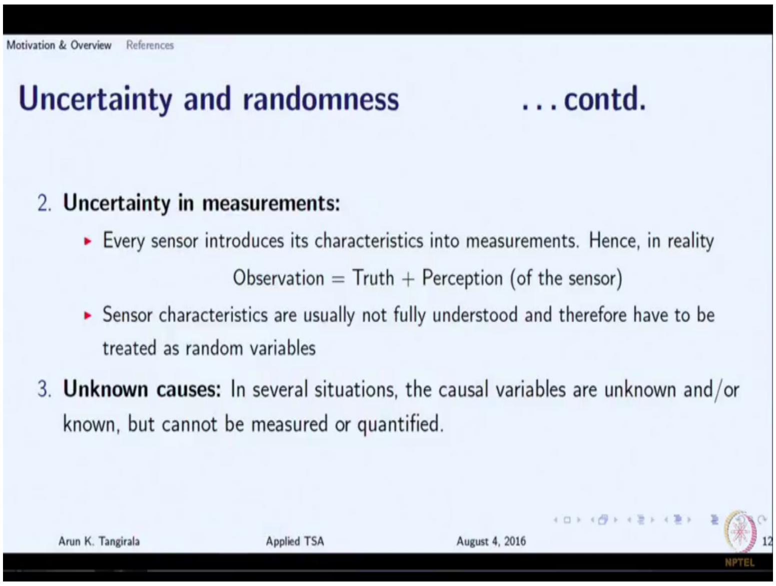
Student: (Refer Time: 12:45).

Let us say they are kind of twins, do you know twins also even though they are born two seconds or few seconds apart they are completely different; why, because whether it is twins; manufacturing of twins; I do not like the word manufacturing or manufacturing of sensors even though they are made at the same time or a few seconds apart and so on, the factors that go into this manufacturing they there are always going to be factors beyond our control.

Any manufacturer of a sensor will try to keep as many factors as possible under you know under control, but there are always going to be factors beyond his control or the manufacturers control. As a result of which there is no way we can actually say that yes this is the error that this sensor is going to give me and the next sensor is going to have this error no way. So, as somebody said yes that is the source of randomness in a manufacturing process that is now giving you the trouble. So, the randomness in the manufacturing process manifests as the random error in your sensor readings and that now causes trouble in your assignments and your projects and so on.

So, the fact is now the second source of uncertainty is your measurement error, even you may have a [FL] knowledge of a very good accurate knowledge of the process, the moment you rely on data; you want you have to respect this uncertainty in measurements.

(Refer Slide Time: 14:33)



Motivation & Overview References

Uncertainty and randomness ...contd.

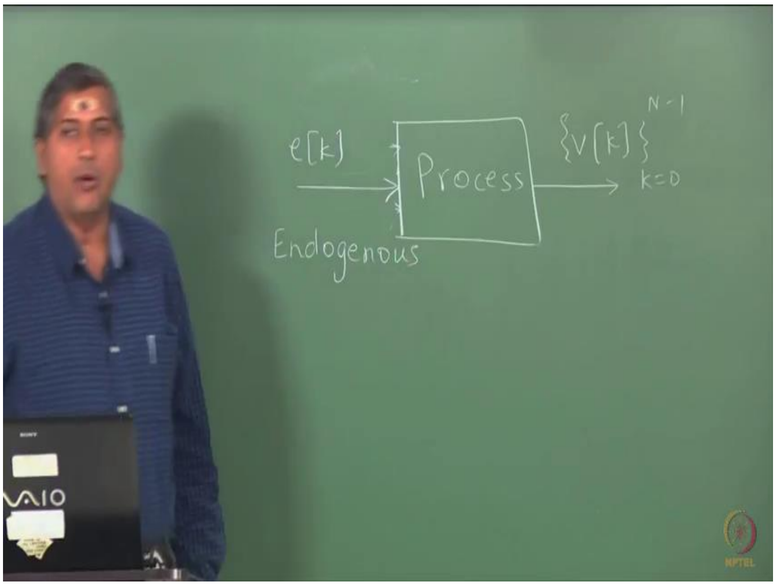
2. **Uncertainty in measurements:**
   - Every sensor introduces its characteristics into measurements. Hence, in reality

     Observation = Truth + Perception (of the sensor)
   - Sensor characteristics are usually not fully understood and therefore have to be treated as random variables
3. **Unknown causes:** In several situations, the causal variables are unknown and/or known, but cannot be measured or quantified.

Arun K. Tangirala Applied TSA August 4, 2016 12

And finally you have this, unknown causes which of course, you can say is the reason why you have uncertainties in measurements.

(Refer Slide Time: 14:45)



But in general, if I actually take any process, so this is my process; this is a typical systems engineering perspective to represent the process in a box, we do not know what is happening, but I am observing some outcome of the process; it could be temperature, it could be stock market index, could be e c g or whatever, so I have a variable here. Let us call this as V; V for variable and this is a signal that I will explain this notation later on maybe in tomorrow's class. In time series analysis essentially I am relying on these observations, this curly braces essentially mean that I am looking at a record of data and k means kth instant in time. So, this is the data record that I have, maybe k running from 0 to N minus 1, you can even consider an infinite record. Time series analysis is about actually looking at this random signal or analyzing this random signal and drawing inferences about the process that is generating it without paying any attention to the cause that is what if I were to ask what is generating V, there must be something exciting the process.

So, when you see people walking on the roads these days you will see they are talking into themselves 30 years ago, there used to be termed as mad people or you know that there is some mental disorder, but today you realize that actually we are talking over the cell phone alright. So, there is a cause something is exciting this person to speak right 30 years ago it was endogenous; that means, something is exciting that person within and thus that person is speaking, but today you cannot say that, you cannot actually pull up and person and say are you mentally deranged why are you talking to yourself, you cannot do that because there is another person; actually exciting this person to talk correct, but we do not know who that person is, all we see is that person is talking we do not want to know that is a personal matter.

But in general, if you see suppose I take rainfall is the classic example a rainfall or even temperature a lot of interest to us; suppose I take the atmospheric temperature and I want to make a prediction of what is the temperature tomorrow or maybe in the evening. So, I collect data in the morning and I start analyzing the data, the engineering approach is to find out what are all the causes you know maybe you know there is wind movement, there is you know cloud movement you know sunshine everything that is actually affecting the temperature may be the pressure everything.

But if you look at it from a practical viewpoint there is no way I can actually measure, even if I were to list all the causes for many processes. There are two things that are going to happen; one that I may be able to list all the causes, but I will not be able to measure them and even if I am able to measure them I probably am able to measure a few, but not all; that is one situation, the other situation is; I cannot even list all the causes that is producing this V. So, these causes it could be either unknown or known, but cannot be measured and you can apply this to many many processes like you take stock market index, I know we can write down a few factors that affect stock market index right what is one factor sorry economic policy or political situation or maybe two brothers fighting we do not know right we know who those two brothers are I do not want to name.

But something is happening or some kind of a natural calamity something could happen and I may be able to list a few causes, can I measure for example, how much two people have fought or the political situation or the economic policy; it is going to be very difficult say you say forget all of that I am not going to take any of those approaches, I am going to rely on the history of this signal and see if the history has anything to offer.

If history is repeating itself if not fully, but partly if it exactly repeats itself we say it is a periodic signal correct. If it partly repeats itself then we say there is some correlation and exploit that; if there is absolutely no scope for any repetition, there is nothing in the history that can help us make a prediction then we say it is an ideal random signal; it is unpredictable, but the fact is in time series analysis we ignore the causes to begin with.

At least in univariate analysis, we ignore the causes alright and that is the difference between what you have been probably looking at until now and what we will look at in time series analysis.

One is that we are actually going to build models from data and two we are not going to take into account any causes at all; to begin with, in multivariate time series analysis we do take into account causes. So, for example, I want to predict the relative humidity of the atmosphere in a certain geographical region, we know from the physics what affects relative humidity, what is the key variable, but what affects moisture; temperature. Temperature is a key factor, if you look at even the definition of relative humidity, you would have temperature coming in there.

So, if I have a record of the atmospheric temperature then I can use that perhaps to make a prediction of the relative humidity, apart from it is own past I can do that and that is a part of, you can say multivariate analysis or you can say part of system identification but by and large the stands that we take in time series analysis at least univariate time series analysis is, I will ignore all the causes simply because I do not know them or I know, but I cannot measure. The day I can measure, I will include that in my model and that is how system identification was born.

So, if you look at a history of time series analysis, it began with making predictions by ignoring the causes and then gradually including the causes which were called as exogenous variables. We will build models of the processes assuming that I do not know the cause, but later on we will actually slightly deviate from that and say that no no no there is a cause, but this cause is not external, but this causes is endogenous that is what we assume and this is a big premise in time series modeling that whatever is causing this V k, this random signal that I am observing to change is not external is from within and that concept has to be understood carefully, we will do that in due course of time. So, in the interest of time I will actually stop at this point.