Corrosion, Environmental Degradation and Surface Engineering

Prof. Harish Hirani
Department of Mechanical Engineering
Indian Institute of Technology, Delhi

Lecture – 25
Digitization of Established Curves & Data Driven Approach

Hello and welcome to the 22nd lecture, of course, on corrosion environmental degradation and surface engineering. So, the topic of this lecture is the digitisation of established curves and a data-driven approach. Now, this topic is very recently, may be say 3, 4 years ago, people have started using this kind of work, and we do not have many established relations among the variables. We often use sophisticated methods, but the question comes whether those methods are going to be very useful even in operating time or will not be useful. Our experience indicates that the very sophisticated finite element method and boundary element method are very good at the beginning, or may be the design stage, but when it comes to the maintenance or operating condition time, the data-driven approach will be providing better results compared to those methods. So, how do we really go ahead, as I mentioned, is more like, you know, the recent approach, the current approach.

So, we want to cover this in one complete lecture, and may be next lecture we will be giving a little more details on this. So, the first question arises: what is the meaning of digitization? The design and digitisation of an established curve, and what is the meaning of an established curve? If you go through the engineering books, mostly the data are being presented in a graphical form in curves, and then there are well-studied established curves that we have been utilising over more than 30 years, 40 years in the engineering discipline. So, what we see in engineering research and economics, in economics also we present the results in a graphical manner.

So, this graphical representation of data is very common. Now when we want to utilise these data for optimum maintenance of the system, we really require the digitisation of that. We really require the data, and in the previous lecture we covered a better objective method where we need to really give the numbers and then give weights to the different variables. That cannot happen without digitisation and without really numerical values. So, that is why we say digitisation is a process of converting analogue or graphical representation to digital one or digital form, and then it can be done either by scanning or it can pick up the data from a graph and directly type 1 by 1 those data. Either we can scan or we need to plot a complete graph or curve on graph paper, and then we find x and y points, and then we tabulate those and then give to the software or the input to the software. So, this is very important for the digital platforms or algorithms, particularly when we are talking about a MATLAB algorithm. We will be developing, and we will be using for those algorithms we really require this kind of data. Now, coming to the data driven approach, we see that optimisation or maintenance strategies can be decided based on the data.

The first step is understanding where the data comes from, particularly from a digitization standpoint. Even if we have a large dataset, say 10,000 entries, it’s unlikely that we will use all of them. Instead, we need a data-driven approach to filter and narrow down this information, ensuring we focus on the most relevant data for

specific scenarios, whether it’s situation 1, 2, or 3. The decisions we make will be based on the data available, but it’s possible that additional data might be needed, in which case further experiments may be required.

In essence, a data-driven approach helps us make informed decisions and gain insights. Sometimes, we seek to enhance our understanding through data analysis. Often, we are unaware of existing relationships until data analysis reveals them, which is why uncovering hidden relationships and trends is crucial. This approach has long been utilized in tools like Excel for trend analysis over the years. However, refining these methods—such as using something like the Kauffman method to develop relationships—is necessary. But we must ask ourselves: will the Kauffman method be suitable in every situation, or will modifications be required? This topic will be explored further in the present lecture.

In summary, when dealing with large datasets, uncovering hidden relationships or trends requires robust statistical methods. The Kauffman method is one such statistical approach, and we'll discuss how algorithms can effectively be employed in this process.

However, we are trying to emphasise machine learning, which really is a more current approach, and then we get many advantages when we go with a machine learning approach. So, based on this data-driven approach, we will be able to really extract the valuable information, and we will be able to develop some predictive models. Modelling can be done with a number of other analysis tools, but when we have realistic data available, those data have been tabulated or may be graphically presented better. We utilise those data and continuously improve, and the more and more better optimisation, the better product will be given to the society.

In summary, by integrating digitization with data-driven methodologies, we can significantly enhance decision-making and automate many operations. For example, automation can be achieved by combining these two tools, enabling us to develop predictive models that can be verified with only a few experiments.

If I want to develop a new model, typically, I will need to conduct numerous experiments. However, by leveraging existing data or graphs, digitizing them, and applying data-driven methodologies, I can extract meaningful insights and develop a predictive model. This approach reduces the need for extensive experimentation to verify the model's accuracy.

So, there is an insight already given to us or will be coming from those data, but to really determine whether everything is as per reality or everything is going to be good from that point of view, we need to conduct the few experiments. So, instead of conducting 10,000 experiments, we may conduct 100 experiments, or maybe 50 experiments is quite possible; we may get better results in that situation. So, by examining data, finding bottlenecks and covering trends and anomalies, we can think about how the digitisation curve will be really when the bringing some sort of optimisation, and that is what we really required. Of course, if this also involves a statistical analysis, we really require some statistical analysis to be performed.

We will explore a few examples to see how statistical analysis can help identify potential dangers, estimate failure probabilities, and assess related risks. By doing so, we can mitigate these risks and make better-informed decisions.

This highlights the importance of combining digitization with data-driven methodologies, which represents the future of comprehensive strategies. For instance, in the context of corrosion, minimizing losses is achievable through such an approach, enabling us to develop effective maintenance strategies.

Consider the example we discussed in lecture 6, where we looked at wear maps. These established curves, particularly for steel, are well-documented in the literature. One key aspect is the ultra-mild wear domain, which is crucial for ensuring a long service life for products. However, achieving this requires careful attention. There are also curves for mild oxidation wear, plasticity-dominated wear, and severe oxidation wear. It's impractical to manually compare these with established curves every time.

If we can digitize the data, analyze it, and apply a data-driven approach, we can focus on the most relevant and dominant factors for our work. For instance, using techniques like Principal Component Analysis (PCA) allows us to identify and prioritize the most critical factors among many. This is why a data-driven approach is essential—it helps us filter and utilize the most meaningful data.

Take the example of steel. Numerous wear maps are available for different materials, and depending on the material we choose, we can select the appropriate wear map. With so many wear maps documented in the literature, it's important to first leverage existing research, applying the right methodologies to analyze it and identify any gaps. This helps us determine how many additional experiments are necessary and what kind of maintenance strategy would be most effective, considering the experiences and findings of various researchers. By doing this, we can move forward more effectively.

So, it will continuously improve our domain knowledge, and it will give very fruitful results. When we do a totally new kind of experiment, it's quite possible in a 90 percent experiment, which gives only some sort of understandable viewers. However, if we analyse existing literature and much more meaningful in a way quite possible, we can really accumulate good knowledge and then start our work. So, that is the aim of this lecture to provide that kind of approach. So, where you can really think about existing literature in a more meaningful way, not only some research has been done, some work has been done, that is ok. Now, we need to know really what is missing and how we can really improve our work, and existing literature will really give much more insight into that. What was happening even though there was some still behaviour? One author is presenting some results, and another author is presenting quite different results, but when we go with the data-driven approach and digitisation or combination on this, early on we will be able to know whether the data were noisy, or maybe the bad sensor, and the data are not appropriate, or it was not well within a range of the sensor, or maybe some other typography errors are there.

These issues can be diagnosed more effectively by combining digitization with a data-driven approach. For example, in lecture 12 on fractography, we discussed the concept of a digital twin. A digital twin is essentially a digital replica of a physical object or system. When materials come into contact, they release wear debris particles. If we can track the surface where this debris originates and match the debris size and dimensions, we can physically verify and construct an accurate model.

The data on debris sources can come from sensors, IoT devices, or simulations. When these data sources are combined, they help create a time-deteriorated system model that tracks deterioration at a localized level, not

just in general terms. Monitoring deterioration at such a detailed level leads to a better understanding of the system, resulting in fewer failures, replacements, and breakdowns, ultimately saving costs.

This is why the digital twin concept is gaining prominence—it represents a significant advancement in system management. Sometimes referred to as a cyber-physical system, this concept is often misunderstood as being solely computer related. However, it's much broader. It involves understanding the real-time health of a system, like how we monitor the health of the human body. Just as different people develop different health issues over time, every system experience unique forms of deterioration, and a digital twin helps us understand and manage these variations effectively.

So, if we have a digital twin and then we have so many systems, it is quite possible, even if one gear pair faces one kind of problem and the and the other gear pair faces another kind of problem. So, naturally, the remedies also need to be different; if you have a stomach problem, you still take a headache medicine. So, the same thing will happen in this situation we are trying to really replicate and then the deterioration system using the wear debris analysis. And if it is done, then it will really give very good results. If we are able to reduce deterioration, we are able to reduce the replacement of the product. Naturally, the damage to the environment will come down, and significant savings in our products or component subsystems will happen. So, this cyberphysical system does not mean it is related to computer science; it is very much related to us, and then here it gives a real-time insight into the state of the physical asset, which is a product or subsystem, and then what are the chances of the failure? It can predict the failure well in advance. In addition, it really gives some sort of help to think about maintenance. How frequent maintenance is really required? A few systems may not require maintenance at all, whether there is no initiation of the crack at all.

We know that maintenance becomes important if we have a sensor suit attached, which is really giving data, and we are using the data-driven approach along with those data, then it will give very good results to us. Now, we say that TT is used to get an image of the thermal image, and then that will have a lesser signal-to-noise ratio that needs to be improved, and then a good kind of filtration system is really required. In this case we have shown that there are 16 to 18 defects out of 30, and then if your thickness is increased, the number of defects that can be detected or lesser as a further increase in the length, or then the thickness number of defects will be further reduced. So, thickness-related variation was shown. Now, this system itself will have many variables; how do we diagnose each variable appropriately? In those cases, they use a principle component analysis, which will be covered in our next lecture.

We have heard about the Eigen vector. The Eigen value gives many the principle dimension where the maximum impact is happening. If we know the main component, we can focus on those components; it is not necessary to have 100 variables and start thinking about 100 variables. So, this is a technique to filter out. So, that is why the goal of the PCA is to condense many dependent variables into manageable subsets. So, if I have 100 variables and an impact of almost 90 variables, Only 5 percent I need to consider those, or shall I go with a much more robust system, because those variables 95 will have a lot of randomness or maybe some sort of variation that I will not be able to control.

So, it will not be manageable. That is why the PCA really will be helping to reduce dependent variables into manageable subsets, and this is very useful wherever there are many variables. When we are going for the defect analysis or defect capturing, it will have many dimensions or many types of variables in that, and then

we need to really filter and come up with the right results. And in this paper and that paper that we covered, only PCA was used to get more meaningful results, and then there are many variables that should be addressed properly, and then we need to really filter out, what are the really more significant variables that have a greater impact, and which we need to control on those variables to get much better results. Now that is why we are leveraging on the data-driven approaches, and we will be giving some weight to the maintenance, and we will be covering there the more predictive maintenance.

For predictive maintenance, I prefer to use a data-driven approach. Even though there are already numerous algorithms and sensors available, it's always beneficial to consult the literature, identify trends, develop a model, and then apply it. This model should be continuously refined based on the data collected from sensors.

Data-driven and predictive techniques rely heavily on historical data, which is why it's important to incorporate information from literature, sensor measurements, and advanced analytics like PCA. These tools help us focus on machine performance and make predictions about potential issues that may arise in the next five or six months, allowing us to take proactive measures.

Even when we do this kind of data analysis, we can discover patterns or indicators that really indicate when the failures are going to happen. What kind of possible failures will happen, and that is why it will really help the maintenance team to take preventive actions or measures before any breakdown occurs. If we go at this kind of data-driven approach, breakdown may not occur at all; it may occur only when we do not have sufficient knowledge about the system or something badly is going to happen, which is not under our control. So, we will be covering condition monitoring preventive maintenance, that is, condition monitoring is critical for predictive maintenance, and then it allows real-time monitoring of the equipment conditions. Now sensor data, which will be gathered and analysed, really can provide some sort of deviation from normal operating conditions. So, again, what are the most normal operating conditions? We need to know from literature from a historical point, and then once we have those data, we can say these are the deviations from normal operating conditions. So, historical data are important for us, and in a previous lecture, I mentioned a weighted objective technique. We say that it is a way to quantify the state of an asset's health, or maybe system health, by combining the multiple indicators.

We mentioned that we can really choose a many variable manufacturers, many indicators, and then attributes, and then we can really come up with which technique or scheme will be the best, and then digitise strategies and combine them with a model that can be trained. Now here we are using a word: we develop a model, and that improves the model based on the data that we have collected. So, we will be really utilising the collected data, the new sensor data, to train that model; however, if we have good historical data available, we can make a model initially, train those based on those models, and then we refine based on experimental results, which we are going to cover or maybe we will be obtaining. So, determining the basically what will be the optimum strategy again optimum strategy which we get from weighted objective method may not be full proof. Reason being that we are taking those decisions based on available data or maybe historical data, and as we start working and then maybe IOT, based on we are collecting more and more data, this again strategy can be modified in situ also.

So, if we have an automated algorithm available when we do good coding, there is a possibility that even in a in a dynamic situation, the weighted objective optimisation method will continuously be improved and will be

changed also. Now, we say the predictive maintenance model can be refined over time with input from maintenance activities. So, whatever activities in the last 6 months have been done now that those can be utilised to improve the predictive maintenance model, and then go ahead with a better strategy, and we say new data collected after maintenance measures should be utilized. Optimisation of maintenance schedule prioritising component replacement, and then whatever the recommendation we want to give will be useful in this case. So, we want to give some sort of recommendation that maintenance should be done immediately or maybe components should be replaced; those can be given.

The maintenance personnel operating the system will have access to this information. Now, how do we proceed with digitization? There are several methods available. In our course, we'll be using MATLAB, which will be provided by the company for the course duration from July to October, and this software can be utilized effectively.

When considering digitization, even graphs can be treated as images, allowing for image digitization, which converts graphical representations into digital formats, as previously mentioned. We will be using Origin software for this purpose, but you can use any other suitable software.

To operate the software, first launch it and open the file you wish to digitize. A manual will appear upon launching the software. You can start a new project, name it, and then import the file you want to convert to a digital format. After that, it's important to identify reference points to calibrate the graph, whether it's from 0 to 5 mm, 0 to 100 mm, or 0 to 10 meters. The software includes a measurement tool to assist with this calibration.

So, that can be given as a reference object, the reference point, and then all the values can be calculated based on that. Once that is done, then we need to really think about the software that needs to be given instructions to track the image. So, that the digitisation process can be started, and once that is done, all the data can be collected or extracted, and the software will be able to give data, and then there is something like the option available, and we can collect all the data to assign. Now they have a number of formats available. Next model or next software, whichever we want to utilise, either Excel, MATLAB, or some other software, you can collect the data in that format. So, whole, whatever the graph available can be converted to digital form.

Either you go have at the points that are available in the graph or you go with a line, or maybe you want 100 points, 500 points, or 5 points dependent, you can select that. So, now, here the data extraction we see once the characteristics have been traced, convert them into new real data depending on the complexity of data technology, something like digitisation or data extraction can be utilised, and then we will be able to collect the data and then store the data. Now, in this case we are going to concentrate only on x and y coordinates, but there is a possibility you go with a three-coordinate x, y, and z also. For this course, we want to just stick to the two-coordinate system x and y and then maybe y and z or maybe z and x, but with only the two-coordinate system, we will not be worrying about the three-dimensional. This can be done on a surface with the more advanced courses, but for present purposes we want to go ahead with the only 2D graphs, and then we convert and then digitise those graphs, and we come up with meaningful results.

Now, suppose we have collected the data. As I mentioned, if you do not have origin, you can collect the data manually also. If you know the 5 points, so ideally for 5 points or 10 points we can plot on the graph paper,

and then we can directly get those data also and we can feed those data. Now, how do we go ahead with the next step? One method is to go ahead with a curve fitting. Curve fitting is also an established method; curve fitting also happens in MATLAB, and curve fitting can also happen in an Excel sheet. So, depending on what we want, we can directly come to the MATLAB or we can utilise the Excel sheet for completeness and explaining in this slide on the Excel sheet first.

So, you see that we need to because we have Excel in our system, Excel 365, that can need to be launched, and we need to really open a new spreadsheet to organise the collected data. Whatever the data we have, either manually collected or from Origin software, those can be utilised there. Then we need to in Excel sheet we can have a x column as a one column and we are assuming the x in this case is independent and then it can be varied and it will affect the value of y. So, y is dependent on this case. So, for x and y, x is independent, y is dependent, and now we do have the option we can go ahead with a scatter plot or we can think about the trend line also. So, if I have 5 points or 10 points, I can go with a scatter plot, but if you want some sort of trend analysis, we want some sort of mathematical formula to be utilised later for some other purpose, not only the scattered data.

So, that can be. This is also possible when you think about the trend line Excel sheet, which has a number of options. Now, we know that exponential is one option, linear is another, and logarithmic, polynomial, power, moving, and some sort of trend lines are available in an Excel sheet. That is why I am trying to utilize them. And it is more common, like almost every system has; these days, MS Office has a given to the students. So, you can utilise those kinds of things also. Now, in this case another one is a trend line. When we talk about the trend line, it has an equation as well in the r squared value.

What is the r squared value? We will discuss it later, and then we can add it directly; it will come as an equation form on a curve itself. We will cover a few examples on that, and then it will also display the r squared value. So, one is a trend line equation and r squared value; those can be really obtained using an Excel sheet or maybe, say, the Excel software based on giving as an input. Whatever the curve in the data that we have collected. After that we can think about plotting the those data may not be really required, but the completeness we want to see whatever we did a digitization when I am plotting. I am getting the same thing or I am getting something different. Now, why are we talking? We may collect the data from one source, another source, a third source, a fourth source. When we use MATLAB we can compare those we can find out what are the differences in this, and then we can use a data driven approach to really analyse why the difference is coming in those data that what is really different.

So, when we do a literature review, we do not really look at and one go all those things once this data have been plotted and then the converted to the graph or plots. We have a number of plots that may be enough from various resources, and then we find there is a huge variation in those values, then we try to diagnose, and then we can figure out whatever gaps between one author who has used temperature, like 30 other authors who have used the operator temperature 70, and then they did not really use those 30 degrees and 70 degrees in their model, and now because of that there is a difference in a value. Now, if we want to add after that the temperature as one of the variables, that is also possible. So, that will give more and more realistic value, more and more creativity, and more and more possibility of innovation. In this example, we've utilized one of the many plotting functions available in MATLAB, specifically the `fplot` function. A simple code has been written to demonstrate this with a basic function where `f(x) = $x^2$`, with `x` being the variable.

In this case, `f(x)` essentially represents `y`. So, when we refer to `x` and `y`, `f(x)` corresponds to `y`, with `x` being the independent variable. For now, we're setting `x` in the range of -5 to 5, but it could easily be -1 to 1, -10 to 10, or any other range like -1 to 5 or -5 to 1, depending on the requirements. After defining the range, you can add a title, axis labels, or a legend, depending on what's needed. You can also customize the plot by using different line colors or styles, such as a red line, a blue line, or various dashed or thick lines, especially when working with data from multiple sources.

In this MATLAB code example, we've kept it simple: `f` is defined as the function `$x^2$`, followed by setting the range.

We've set the range for `x` from -5 to +5 as the independent variable. After defining this range, we can plot the function within that range. The software will automatically determine the increments or divisions for the plot. Once the plot is generated, you can label the `x` and `y` axes. In this case, `f(x)` is treated as `y`, which is noted in the figure.

Let's consider an example we covered in a previous lecture, where we examined the corrosion of aluminum 7075 with a nano-composite in an H_2SO_4 acidic environment. We also tested in an HCl environment. The corrosion rate in H_2SO_4 started at around 0.15, whereas in HCl, it was almost 50% lower, making HCl less harmful. Additionally, NaCl had an even lesser effect compared to these acidic environments.

When digitizing the data, we can identify trends—such as multiple trend lines—using an Excel sheet to visualize the results.

We used Origin software to collect data points, and based on those points, we created the plot for this figure, with the second figure to be covered next. When fitting the data, we provided the R-squared (R^2) value and the polynomial equation, which is a second-order polynomial. It's clearly indicated that all the curves are polynomial, featuring an `$x^2$` term, an `x` term, and some constant value.

As I mentioned, both the first and second figures were taken from the literature and have already been covered in a previous lecture. Now, let's discuss the significance of the R-squared (R^2) value. This value indicates how well the regression model fits the data. For example, in this case, the model is represented by the equation $Y = 5 \times 10^{-6}X^2 - 0.0017X + 0.1584$. The R^2 value of 0.9999 suggests that the model fits the data extremely well.

This high R^2 value shows that the regression model accurately represents the data points we provided. This is a statistical measure, and as mentioned earlier, when using a data-driven approach, it's crucial to rely on statistical methods and measures.

Now, it depends on how much variance occurs between the dependent and independent variables. The R-squared (R^2) value typically ranges from 0 to 1, not from -1 to 1, because it is a squared value. A value of 0 means that the dependent variable cannot be explained by the independent variable at all.

So, these two variables do not have a correlation, 0 correlation; they are not connected; they are not interrelated

at all. Well, coming to the value of 1, it says the independent variable has a complete explanation, or maybe the way has very strong dependence, and the way we have given this as a model is really matching well.

Everywhere you can see 0.99. So, these are the models very suitable for this model. Now, this is what we see when using the Crawford method of an Excel sheet. This method provides a valuable understanding of the subject matter. What kind of relationship is there between the dependent variable and the independent variable? This gives a relation between the two, maybe say one independent variable and another dependent variable. Similarly, we may have a two-dependent variable or a three-independent variable, and we can do digitisation of those things also.

Moving on to the second case, which involves HCl, we see a comparison between this and the H_2SO_4 scenario. The second slide presents results for H_2SO_4, followed by the HCl data. In the MATLAB code, which is straightforward, we define several functions—function 1, function 2, function 3, and function 4—corresponding to 2%, 4%, and 0% concentrations, as well as hybrid nano composites. The results demonstrate a strong correlation, with the green line indicating a perfect match. This indicates that the model performs very well and can be effectively used for initial analysis and application in our system.

Although this model has been adapted from existing literature, it shows a strong correlation with at least one variable, which is time. However, to enhance the model, it would be beneficial to also establish a relationship involving the percentage of nanocomposite, a variable not covered in the literature. The current literature addresses variations at specific nanocomposite percentages (0%, 4%, 6%) and hybrid cases, but only in relation to time. By digitizing these data points and incorporating both time and nanocomposite percentage as variables, we can develop a more comprehensive model. This will allow us to create a model with three variables: y_1, y_2, and an independent variable x.

So, in that case we can really get good results, or maybe say y as a one-function that is a corrosion rate, and then I can think about x 1 as a time and x 2 as a percentage. So, depending on how we have it and what kind of results I want, we can really do good data manipulation and then come up with a better understanding and insight into what will be the impact, what will be the impact of the percentage individually, time individually, and when they are operating together, what will be the overall correlation, the overall relation for that. So, this is H_2S0_4. Now we are coming to the second one, which is a HCL.

As you see, HCL is less harmful. Similarly, we are getting a good correlation for that also (0.9954), 0.99 is the almost very good correlation, and we can reproduce the same thing in MATLAB. So, whatever we are getting from literature, similar relation or similar kind of curve we can reproduce here.

This does not imply any intention to cheat the system. Rather, the goal is to understand the current mechanisms and evaluate whether my comprehension of the existing system and established literature is accurate. With this understanding, I can explore innovative applications for my own industry and consider how this knowledge could benefit society.

In the analysis, we have plotted the data for various values such as $4 \times 10^{-6}, 3 \times 10^{-6}$ and so on. The results are visualized with different lines in red, green, blue, and violet, demonstrating a strong correlation and accurate comparisons.

Additionally, we examined the impact of corrosion over time using eddy current testing methods. The results indicated that corrosion spreads less after 3 months, increases after 6 months, and further increases after 10 months.

The authors provided three values from their study, comparing quoted and unquoted examples. Their findings showed minimal variation between the two conditions. This curve illustrates how passivation affects corrosion height over time, highlighting that increased exposure leads to higher porosity and greater failure risk.

This slide already covers these points. To digitize this data, we can use the current figure as an example, noting that it contains a relatively small number of data points (1, 2, 3, 4).

So, the 4 data points we have can be utilised for this purpose; in other cases, we have only 3 data points. So now this also indicates that if we have a lesser number of data points, the curve fit option is not really giving very good results. Therefore, a larger number of data points will yield better results for us. Now here you can see that in the regression earlier case it was almost 0.99 something, while in this case instead of going for the polynomial, we have tried to fit the logarithmic relation, which is getting better results compared to the polynomial. That is why here the log plot has been established.

You can see here that in one case it is 0.9987, while in the other case it is a much lesser 0.93. However, we keep a limit; if any weird correlation or R square values are less than 0.9, we should reject it. We say no, this is not really good data, and we will not analyse this.
In this analysis, we observe that the damage on the coated surface is slightly less compared to the uncoated surface, with a correlation value of 0.93. We have translated the developed relationship into MATLAB code, which includes equations such as , $0.0148 \times \log(x)$ and $0.0182 \times \log(x)$ with these constants applied similarly in different cases.

The plotted figures are as follows: Figure F1 is shown in red, and Figure F2 is depicted in blue. These plots demonstrate a good correlation, though the red line shows a slightly lower correlation constant. While the correlation constant alone isn't a definitive measure, it provides insight into the strength of the correlation. Ideally, a correlation coefficient around 0.9 suggests a robust model. If the coefficient drops below 0.9, it may indicate the need for an alternative model, repeated experiments, or further investigation.

Additionally, in the other curves presented, we see three data points plotted in red and four points in blue. The correlation in one case is 0.91, but deviations are evident.

Now, correlation is we are not getting very good. Despite creating this function in MATLAB and plotting it, we observe a deviation. The F1 plot, colored in red, exhibits a strong correlation, while the F2 plot, colored in blue, displays a significant gap. So, this I do not. I am not able to see much separation while we think about MATLAB coding. We are able to see much separation. So, as far as possible, we should go for the high correlation constant, or maybe if it is not the both are not following a logarithmic, then quite possibly I need to change this relation, but that will not justify. If I change a relation, it will not justify; maybe in a coding case it is really following a logarithmic variability, and while coming to the other unquoted one, it is not really

following the same thing. So, there will be, or maybe if it is, then we need to understand why it is really happening.

So, it will provide some more understanding. Now with this approach, we are trying to understand the existing spectrum, how the dependent variable is really affecting our independent variable, is affecting the dependent variable, and what is really the measure of that. So, that has been done. Now in the last one, only one data point has been given, and then we are also getting a good correlation. So, in 3 logarithmic cases, 2 cases we are getting good relations, and 1 we are not getting very good relations. So, either that experiment needs to be repeated or just an indication that that experiment needs to be repeated sometime. This kind of approach also helps us to understand maybe when we were performing some experiment some data, which we got either the signal problem, some sort of problem, or maybe variable in a variation in some sort of humidity, or maybe temperature, and those needs to be changed.

So, in this case also, we are getting perfect matching because of the high correlation. That is why I give a lot of weight to the high correlation, and then the correlation coefficient on the higher side we can believe in this. So, if the if the correlation coefficient is not very high, then there is a doubt, and then, if possible, we should reveal the experiments. Now last one here in this sequence I am taking again ultrasonic on one testing. We say that this ultrasonic testing can be utilised to diagnose defects. However, a closer look at the diagram reveals a significant amount of overlap in the data. So, there is not a clear-cut difference, and if you look at this one also, the data are very scattered.

This scattering of the data creates a lot of confusion, and then this indicates that we really require refinement in a result; we are required to utilise, maybe, more sophisticated techniques for the betterment of the results. Now look at this data too; we are getting some high value here, then low value here, then high value here, then again low value, then high value. So, a lot of randomness, even though this is a kind of mass decrease rate reduction in a mass that should follow some sort of uniform train. So, this again is coming when we are trying to utilise a digitisation and data-driven approach. We are able to look from that point of view; this is something when I am trying to digitise and I do not find that kind of regular pattern.

And then when we try to look at this one question, whether these data are sufficient, do we require a repetition of data, or I think that the curve-fit method may not be suitable at all. So, let's see what the curve fit results are in this case. So, in this case, we have been plotting energy versus the corrosion period, which is a time. You can see here the corrosion and correlation.

We tried three things. One is that the that the polynomial correlation coefficient is 0.90 while we go ahead with a linear curve; it is still 0.86 when we go with the exponential 0.778. We can go through all six kinds of relationships, but this is just for understanding purposes. So, now that energy versus electric corrosion period is plotted, we find that the polynomial is giving the best fit of 0.9015, but still, I believe that needs to be repeated, and then we do not find that kind of result. Whatever the function we get in an Excel sheet, we are trying to plot and then compare and see the results. And as you are able to see clearly here, the data are very scattered, and these are the realistic cases. Many times we get very scattered data. If I do a curve fitting correlation coefficient, that is one possibility. Another possibility is that we go for the better techniques, and the better technique in this case is generally the PCA. Go ahead with the machine learning algorithm and try to modify not only this, but whatever we do, curve fit, we limit.

If I want to go with the next experiment again, the curvefit method may not be giving the good results for that. Then I have to go for the recurve fitting, which may be in the refitting of that. So, instead of that, we go for the better approach that will give us good results. Now look at that was on versus side. This was where I mentioned that already there is a lot of randomness as such, mass decrease rate sometime increasing then decreasing, and then this is basically from a distance. Now and then there is some sort of attenuation of the sensor signal, and also there is a change in the crack width over, even though there is a crack, and then what we are able to predict if we are able to predict a constant value makes a lot of sense.

Because of that constant value, the crack is not changing, and whatever the value, even what I am measuring from a 7.5 centimetre away, the crack length is a constant, and it is not having any effect.

We observed that polynomial curve fitting suggests a relationship of the form $0.001 \times x^2$ with a regression coefficient of 0.66. For the blue line, no significant correlation is evident. Ideally, the constant value should be consistent across the range, but there is noticeable variation. This indicates that as the distance increases, the crack dimension may change accordingly, which suggests that the polynomial fit might not be entirely accurate.

Additionally, the results for mass loss or mass decrease appear to be quite random.

To improve the model, we might consider using higher-order polynomials. Instead of x^2, we could try x^3 or x^4 or even higher-order terms to see if these provide better results.

There is a possibility, and then we need to explore that, which is what we have done. You can see here that even the ninth degree polynomial and tenth degree polynomial have been fitted when this is a wear mass and this is a crack width dimension that has been shown, and then we are seeing there is a good correlation happening (0.90, 0.991). However, a closer look at the equation reveals that we frequently use the term "overfitting." Now if I go ahead and come up with a new one of the new data value again, I will find that it is not matching with this. So, instead of going for this kind of higher series, it will be preferred to change the method and then go for the kind of machine learning, or PCA, and then we try to figure out which is really impacting more severely and then which is not really significant, and that can be done when we do PCA, and that will be covered in the next lecture. Thank you for your attention.