Adaptation of bacteria and its implications in environmental application

Article talks about the additional knowledge for a practicing civil engineer, apart from the curriculum taught

Models are tools that are developed from theories and experiments to predict the behaviour or outcomes of processes. For example, consider the following cases:

Case 01: Market prediction

Stock market gives significant income for many, while it has severe risks fraught with it. It is therefore required for users to make predictions with confidence, what could be expected with the stocks price trends in the future. Mathematical prediction models may then be required to predict the stock prices based on some characteristics identified and included in the model.

Case 02: Strength of concrete

A design engineer would come up with the design specifications that primarily include the required compressive strength of concrete for particular structural elements. The strength of the concrete primarily depends on the constituents, the ratio at which they are mixed and the mixing protocols. Failure to predict the compressive strength of a concrete at the mixing stage may inevitably require making multiple samples in probably multiple attempts to devise a procedure of concrete casting that would suffice the requirements of the design engineer. A mathematical model based on the attributes that define compressive strength may then come in handy, which could be used to devise the mixing protocol and strategy prior to casting concrete.

Case 03: Prediction of the spread of pandemic

We have experienced first hand the plights of a pandemic and the adverse consequences. We have simultaneously observed people predicting the spread of the disease across regions and countries, effectiveness of anti-viral treatment procedures and perhaps the effectiveness of several vaccines developed by different pharma companies. Most of these predictions were based on different attributes that defined the spread and the treatment efficacy, through models developed using different tools.

Case 04: Plan a return trip to the moon

It may sound quite too familiar to many to learn that man has set foot on the surface of the moon in the middle of the twentieth century. A travel along a road and a travel to the moon and back may not draw parallels as the latter would be leaving earth, where earth itself is on another trajectory. When we have left the Earth to set foot on the moon and on our return journey, the Earth may have shifted to another location, and may also have spun around its own axis, moving the location where the rocket left from the return trajectory. It is therefore required to predict the location in relation to the trajectory of the rocket to safely land back on earth. This would require specific computations based on theories, and the computations are termed models.

Case 05: Performance prediction of an incinerator

Reactions are common in chemical and process engineering, where the contribution of fields of study is vital in perhaps all forms of engineering. For example, we may talk about an incinerator, that is designed to incinerate materials for production of energy. Several attributes may have to be predicted inside the incinerator at the design stage, for example, the temperature gradient and the combustion process. Should the fabrication of the reactor be designed without models, several models may have to be fabricated and tested for performance in order to select the final reactor model. It is imperative to have a model that could predict the performance of reactors based on the design elements at the design stage so as to reduce the cost of time and money. Mathematical models of performance from design attributes of the reactor is therefore crucial in this exercise.

It is apprehensive from the 5 cases from different applications that prediction of performance of an entity or design is vital while it solely depends on mathematical models. An article that appeared on American Scientist (volume 111, page 44) classifies models into two classes, (a) predictive models and (b) descriptive models. In the article it is discussed that once the descriptive models are validated using real-world data, it can bs used as predictive models. Therefore, in my definition, descriptive models are also predictive models and that classification is disregarded in further discussions in this article. In addition, the models are referred to by different names in different applications. For example, a model defining the force of an object that is accelerating is called Newton's law instead of Newtons model while the consumers at the stock market may be using an application on the mobile, where in the background a model does the work which is not known to the user. The term 'model' is therefore has wide application across fields and specifi-

Models that predict!

A model without validation is nothing more than the absence of a model. How can we improve confidence in model predictions?

cations. In general, the models can be either theoretical, conceptual, empirical or even a fusion of multiple avenues of modeling techniques. In addition, the models can be either mathematical, statistical or even quality-based, where the outcomes defines the type of model required. All models in brief are by design, approximations of real-world phenomena (we call it approximations, as we would rarely get exposed to all the variables of a case analysed). In addition, we have limited information on variables, and that the models we develop are only approximations, which requires assumptions be defined to define the applicability of the model. This All models in general have some elements in common,

1. *Data acquisition (the knowledge that is already available), that can even be theories (resulting in theoretical models)*
2. *Development of model*
3. *Defining the uncertainties in the variables and the propagation of uncertainties through the processes in the models. This stage is also known as the training of models, should the model be based on training of models.*
4. *Develop the prediction boundaries and confidence of models.*
5. *Calibration and validation models.*

Most exercises carried out in model development and prediction of performance often include data acquisition and model development. However, the uncertainty propagation analysis and validation of models are seldom carried out, making the models invalid for implementation or perhaps, implemented without context. Whether a model works without an uncertainty anaylsis or validation is a mere coincidence with a probability, and one can never be sure of the outcome until it has been observed experimentally. In this article I will discuss the basic elements of model development. Before we go into details of modeling exercise, let me quote George Box, '*all models are wrong, but some are useful*'. We will discuss the useful models and how the usefulness is defined and enhanced in models.

PRELIMINARY WORK

Prior to any work, it is critical that we define the scope of the study or analysis. In a study that involves development of a model, it is vital to clearly define the objectives and the boundaries of the model that is aspired to be developed. It is blatant that some basic understanding on the exercise is required prior to data acquisition. What is expected of a model is critical in identification of variables (both independent variables, also known as design variables and dependent variables, also known as performance variables). Design variables (independent variables) define the performance of a model or analysis, and that needs asseveration on two things,

1. What factors in total, affect the performance of the entity
2. What's the scope of the design variables considered for the exercise

In a quest to develop a general model that could be used to predict a parameter in all possible circumstances, we may have to emphatically identify all the factors that would affect the prediction. In many cases, such as predicting the weather, we may never be able to develop an ultimate model (like the theory of everything, Einstein dreamt of developing), which is mainly due to either impossibility of finding the complete set of variables or the impossibility in understanding a mathematical relationship of the collective impact all factors would impart on the dependent model. However, in most engineering cases, scoping down the applicability of the model aspired to be developed is often found to be the objective. For example, developing a model to predict the compressive strength of concrete in any place in the world may need to include geographical parameters to be included in the prediction model, where restricting the predictability to a particular zone may exclude such factors in the model. This model would therefore be inapplicable in other zones without calibration and/or validation of the model, which will be discussed later in the article.

MODEL DEVELOPMENT

Let us begin the discussion with a simple example of the mathematical model to predict the attraction force between two masses, Newton's model.

where, F is the force of attraction between two masses (M1 and M2) parted by a distance of r and G is a constant defined through multiple experiments. The model is based on the understanding that the force was observed to be directly proportional to the product of the masses and inversely proportional to the squared of the distance between them. To develop the concept into a mathematical model that would predict attraction force, a constant G has to be introduced.

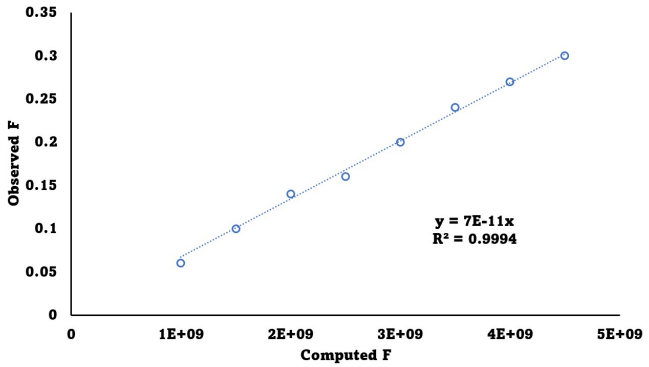
Defining the constant G

Several experiments varying M_1, M_2 and r with multiple replicates were conducted while observing the corresponding values for the force F. A graph between the computed F (= M_1M_2/r^2, without G) and the observed F was drawn. Should the relationship between the computed value and observed value be linear, and the numerical values are exactly the same, the con-

Models that predict!

A model without validation is nothing more than the absence of a model. How can we improve confidence in model predictions?

Figure 01: Computed Vs Observed F — Linear model



stant G would essentially assume a numerical value of 1. However, if the relationship is linear while the values (computed and observed) are different, the constant G would have to assume a numerical value other than 1, as observed in this example.

Let us delve a little deeper into the experiments that defined the numerical value for G. Each time an experiment was conducted, measurements were taken where each measurement would inherently include random uncertainty. In this example, the measurements taken were M_1, M_2, r (independent variables) and F (observed variable). The computed value of F obtained from the equation would have a collective impact of the measurement uncertainties, propagated through the mathematical model. The computed values of F would therefore have a probability distribution of values around the mean, often depicted as normal or log-normal distribution.

Simultaneously, the observed value of the dependent variable F would also have an uncertainty around the mean, normally distributed in general, where each value would have a corresponding probability value for occurrence. Defining these uncertainties would require multiple replicates of observation for all measured dependent and independent variables. It would be meticulous and ambitious to expect the mean values and the standard deviation (typically the PDF of uncertainty) of both computed and observed values for F to be identical.

The next step in the exercise would be to develop a relationship between the computed and observed F. In order to attain this relationship, the independent variables need to be varied numerically across a stipulated range. Similar exercise of defining uncertainty for computed and observed F would have to be replicated. Assuming a linear correlation between observed and computed F, we could expect a mathematical representation in the form of y = mx + c where y would be the dependent variable (observed F, in this example) and x would be the independent variable (computed F, in this example), shown in Figure 01. Ideally, according to Newton's model, the constant 'c' would have to be zero (an observed zero of F should correspond to a computed F of zero). In addition, the constant m in the mathematical model for linear relationship would be equal to the constant G from the Newton's model, given the constant c in the linear model amounts to zero. Best fitting curve corresponds to the lowest RMSE, which is the aggregated error of each point from the linear model.

Implementation of this model to make predictions would need further assurance on the accuracy of this constant G. For example, the figure shows high accuracy, indicated by R^2 value close to 1. However, a value less than 1 indicates presence of uncertainty in the model which would be reflected in computed G. We have discussed the simple case for a linear model, based on the theoretical understanding. What happens, when the trend between variables is unknown.

CHOICE OF MODEL

We will discuss development of a model, when the pattern of relationship between variables are not theoretically known. For example, let us say that the compressive strength is positively and linearly correlated to A/C ratio and compaction (considering the impact of other variable in this exercise as constants). It could then be approximated to a mathematical model based on linear correlations similar to that is shown below.

C = a.AC + b.Com + c

where, C is the compressive strength, AC is the aggregate to cement ratio (A/C), Com is the compaction energy supplied and a, b and c are arbitrary constants. With a regression analysis, we could determine the constants with uncertainty bounds.

Impact of other factors were assumed constant (or the impact is insignificant) when defining the scope of this study. Maintaining variables constant does not imply, they are absent, but maintained at a constant numerical value, for example, the water to cement ratio was kept at 0.45 throughout the experiments in this study. Suppose, we do another set of experiments at a constant water to cement ratio (eg. at 0.5) instead, the second set of experiments where only A/C ratio and compaction were varied with W/C ratio at 0.5 and all other variables similar to the first exercise, may yield a similar relationship (compressive strength being linearly and positively correlated to A/C ratio and compaction). We need to consider few cases hereafter,

Case 1. The relationship abides by the

Models that predict!

A model without validation is nothing more than the absence of a model. How can we improve confidence in model predictions?

same mathematical representation and same values for constants

This would be the simplest of the cases where mathematical representation , including values of constants wouldn't change with the change in W/C. This while further reinstates the mathematical relationship between the dependent and independent variables, it also suggests that the W/C ratio has insignificant impact between 0.45 and 0.5. However, we wouldn't be able to generalize the impact of W/C ratio being insignificant in affecting compressive strength as yet, as this experiment is not designed to test (we will discuss on validity of models later).

Case 2. The relation abides by the same mathematical equations, with change in constants

This is quite similar to case 1, where the impact of A/C ratio and compaction on compressive strength followed the same linear mathematical representation. However, one or more constants of a, b and c have changed when obtaining the best fitting curve (optimizing RMSE). This implies that the impact of W/C ratio affects the constant that changed with W/C ratio. Possibly, another study varying the W/C ratio and monitoring the compressive strength would result in obtaining a mathematical representation for the relationship between W/C ratio and compressive strength, which can then be substituted to develop a composite model to predict the dependent variable compressive strength from 3 independent variables (A/C ratio, compaction and W/C ratio).

Incorporating the impact of W/C ratio in the model could be done in two different approaches.

Approach 01: nesting the model

We may conduct a different experiment varying A/C ratio and compaction, while keeping the W/C ratio constant in a set of experiments and repeat it for another W/C ratio. This could be repeated for several W/C ratio, while keeping the model affixed to predict compressive strength from A/C ratio and compaction and tabulate the values for a, b and c for across experiments with varying W/C ratio. The constants that do not change across the set of experiments would then be affirmed not to have been affected by the W/C ratio, while the constants that changed with W/C ratio would be considered otherwise. Considering each constant that now becomes a variable with W/C ratio will now have to undergo regression analysis, and a mathematical representation could be obtained which may by linear or nonlinear. The new mathematical model may then be nested in to the model developed earlier, by substituting the new mathematical model for the constant(s) in the earlier model.

Approach 02: composite model

Suppose we identified that W/C ratio affects one or more constants in the mathematical model that predict compressive strength from A/C ratio and compaction, we may redesign experiments to include W/C ratio as the third independent variable (instead of repeating the same set of experiments varying A/C ratio and compaction for different W/C ratio). In this case, all three independent variables will have to be varied simultaneously, with meticulous design of optimum number of experiments, so as to facilitate statistical significance in subsequent modeling. Methods such as curve fitting (regression analysis) or neural networks can then be used to develop a model to predict compressive strength from all three independent variables (A/C ratio, compaction and W/C ratio). The model developed in this approach may be more robust compared to the model developed in approach 01, when considering the propagation of error.

Case 3. The relationship changes

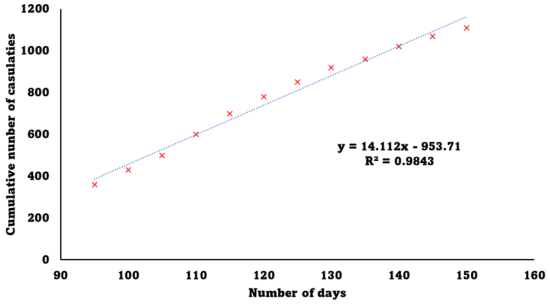
In the previous two cases we observed the relationship between the dependent and independent variables (A/C ratio and compaction) being consistent across W/C ratios. However, it need not be the case. Perhaps, we could observe a completely different relationship between compressive strength and A/C ratio and compaction instead of linear at higher W/C ratios, (0.5 in the second exercise). This for example could be represented by a mathematical model shown below;

$$C = a.\exp(d.com) + b.AC + c$$

where another arbitrary constant, 'd' has been added. This is a non-linear model, with an exponential function. It is clear that the W/C ratio affects how A/C ratio and compaction affect the compressive strength. Unlike case 2 where we had two different approaches, in this case, the first approach would not be suitable as the mathematical representation itself changes with W/C ratio, and not just the constants. Therefore, only approach 02 is applicable, and comparatively more experimental data would be required in this approach for the development of an accurate mathematical representation.

Now that we have spoken approaches and methods that may be adopted, yet another

Figure 02: Model of Researcher A



Models that predict!

A model without validation is nothing more than the absence of a model. How can we improve confidence in model predictions?

aspect is also quite vital in generalizing the models across wide range of variable scales. A model should not only represent the statistical fitting, but also logically represent the data it encapsulates.

Let us discuss that through the example of the cumulative number of people died in Sri Lanka who had contracted Covid-19. Two different researchers have studied the trend based on the data they obtained, as described below.

Researcher A: obtained data from 01.04.2020 till 31.05.2020

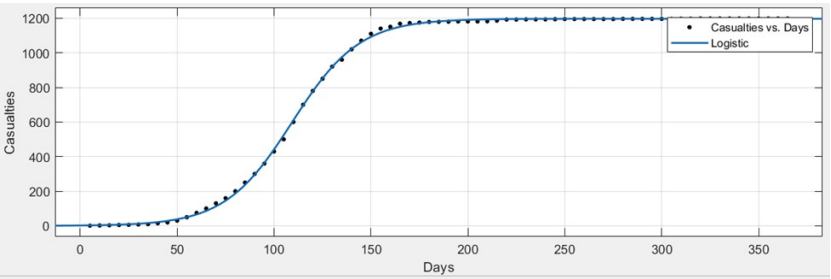
Researcher B: obtained data from 01.01.2020 till 31.12.2020

The objective of both is to develop a mathematical model to represent the casualties on 01.01.2021 based on the trend observed in the spread of the virus among the Sri Lankan population.

Candidate A plotted the data obtained and found a linear trend and developed a linear model to make the prediction as shown in Figure 02. According to this model, the prediction would be 4267 casualties by 01.01.2021.

Researcher B developed the model based on the data acquired, develop the model shown in Figure 03, where the model is a logistic curve (saturation curve). The prediction candidate B made was 1192 casualties by 01.01.2021. Both the model have proven to have achieved very high accuracy according to the data used to develop the model, while the predictions have swayed by almost three folds.

Figure 03: Model of Researcher B



In Sri Lanka, the population would be finite (the number of people would be a finite number). The model according to candidate A is an ever increasing model, where the casualties would continue to rise with number of days, and at some point would predict more deaths than the total number of population in the country. This delineates the model illogical, while the model developed by B saturates with time and would be seen more logical. Both these models however, may not be accurately predict the number of casualties by the end of 2021, should there be a second wave of contagion occurring. Does this mean the model developed by A a wrong choice?

The answer to that question would be, that it is subjective. Suppose, a prediction needed to be made on the casualties on127th day of 2020, both models would be able to make an accurate prediction, and that model A would fail only when a prediction is made beyond the 150th day. Given the studies defined the scope of their analyses and the model peripheries and that the predictions conform to the scope of the model, the conclusions hence derived would be acceptable in scientific studies.

INTRAPOLATION AND EXTRAPOLATION OF MODELS

I would like to now discuss a case that has been discussed in an article that appeared on American Scientist (Jan/Feb, 2023, page 42), to elucidate the case that I would discuss later on. Galileo develop the model $d=1/2gt^2$ to predict the distance travelled by a ball dropped form a height, and given the time of travel is known. This model had a theoretical concept, he conducted experiments to compute the value of g (gravitational acceleration), which he defined using umpteen amount of experiment. A point to note is that all his experiments were conducted on earth, from buildings that had different heights. However, later in a publication, he declared the time that would be taken by an object to reach earth when dropped form the moon to be 3 hours, 22 minutes and 4 seconds. It had some error in computing the distance between earth and moon, which is a minor issue. What was a major mistake was that, he assumed the g to be constant throughout the travel from moon to the earth. Now we know, that at the surface of moon, the gravitational acceleration would be towards the moon, and the gravitation pull towards the earth would be insignificant, compared to the gravitational pull towards the moon (the gravitational field of the earth on the surface of the moon would be 3600 time less compared to that of the moon). In addition, he wouldn't have known if the g was a constant at all? With all the experiments conducted near the earth, the distance his experiments entailed were insignificant compared to the distance between earth and moon. What Galileo did in estimating the time of travel from the moon to earth is called extrapolation of his model (he applied the model he developed on a case beyond the scope of the experiments that were used to develop the model) and failed to assess if the assumptions made in developing the model would still be applicable. The author claims that physicists to this day, tend to extrapolate the model, thinking they develop universal truths rather than paying more attention to the assumption and limitations of models they develop. In my opinion, engineers, com-

Models that predict!

A model without validation is nothing more than the absence of a model. How can we improve confidence in model predictions?

ing from physical science background, tend to reciprocate the same trend of extrapolating models, beyond the context and scope of the model.

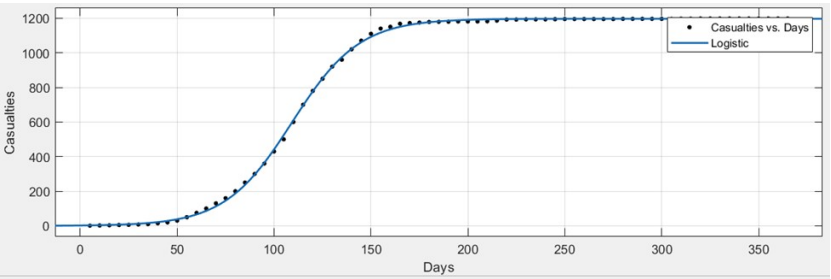
The article further discusses the improvement Newton brought to Galileo's model in computing the gravitational pull. The gravitational attraction force between two masses is given by the model , developed by Newton. This model could be used to compute the time needed for an object to travel from the moon to the earth, more accurately compared to Galileo's model. The model of Newton again became a universal model, that was widely applied beyond the scope of the experiments with which it was developed, yet, the extent to which we explore in the universe, we are well within the scope. It could be argued, with the model of Einstein, space-time model, proves that the Newton's model is only an approximation and that the theoretical concept of the model is not accurate. However, one could argue that with the approximation, we made a trip to the moon and back, and if the uncertainty in the model is worrisome. Only when we tried to use the model to compute our travel to galaxies, may the uncertainty be magnified enough to delineate the model unacceptable.

Now let us get back to the discussion of intrapolation and extrapolation. Suppose we are to predict the compressive strength of concrete based on the A/C ratio, given all other variables are kept constant. In this experiment, assume we have varied A/C ratio between 2.5 and 7.5, and we developed a linear model that would predict the compressive strength. If we are to predict the compressive strength of a concrete sample between 2.5 and 7.5, it would be intrapolation of the model. However, if we are to predict the compressive strength of a concrete with A/C ratio less than 2.5 or more than 7.5, we would then have to extrapolate the model. The mathematical model that was developed may or may not be correct beyond the experimental range, making the prediction vulnerable and untrustworthy. This was the case observed in the cases of Galileo's and Newtons model. However, I wish to discuss on another aspect of the mathematical model that would need caution when extrapolating the model, in relation to uncertainty.

Speaking of uncertainty, let me discuss another case that was discussed in the article mentioned above. NBC developed a model to predict who would win the presidential election in US in the year 2016, a week before the election. The prediction was that Clinton would garner 51% of the popular vote while Trump obtained 44% with a margin of error of 1%. In this case, 1% is the uncertainty defined by the model. However, in the elections, Clinton received 48.2% while Trump got 46%, which means the model emphatically failed to make a decent prediction (the results were beyond the uncertainty margin that was defined). The question asked by the author of the article was, what was the meaning of 1%? Theoretically the 1% meant that the actual vote Clinton was expected to receive lied between 50 – 52 with a 95% confidence. The uncertainty arose, due to many factors, the most important of that being how well the sample used to develop the model represented the actual population. A sample to represent the population emphatically must be inclusive of all characteristics or attributes of the population, which is hard to attain. And hence arises uncertainty. When defining uncertainty, the representation of error must be defined categorically and completely, so as to get an accurate estimate of uncertainty. In this example, it could be concluded that the uncertainty was much larger than what was reported, or in fact, blame it on the probability for missed prediction.

Now let us get back to the concrete compressive strength prediction example. When experiments were conducted with A/C ratios between 2.5 7.5 and the observations on compressive strength were made the scope of the experiments are defined accordingly. Thereafter, regression analysis was carried out to fit a linear model (y = mx + c, where y is the compressive strength, x is the A/C ratio and m and c are arbitrary constants). As the uncertainties from y and x mean values propagate through the function while another form of uncertainty arises on fitting the mean values on to the linear model, the constants m and c would have uncertainty boundaries. This phenomenon would lead to a confidence envelope as explained earlier. What I would like to draw the attention to is the shape of the uncertainty envelope of the model. The uncertainty envelope is the thinnest in the middle of the experimental range and continue to expand non-linearly towards the edge of the model on both sides. Especially beyond the experimental range, the confidence envelope increases in area, indicating a higher uncertainty in prediction beyond the range. Therefore, even if the

Figure 03: Model of Researcher B



Models that predict!

A model without validation is nothing more than the absence of a model. How can we improve confidence in model predictions?

model is valid beyond the range of the experiments, the prediction would entail higher uncertainty beyond the range. For optimum application of the model, and for better prediction, it would be prudent to apply near the middle of the range of the experiments.

Authored by:

D N Subramaniam
Professor in Civil Engineering
University of Jaffna