Why Use Phenomenological Models Over Empirical Options

R. Russell Rhinehart
www.r3eda.com
2025-01-20

Question: Why use phenomenological models over data-based modeling options?

Russ' Responses: I would broadly classify models as empirical or phenomenological. I'll contrast the two by examples then discuss pros and cons of each approach.

Empirical models are developed by matching flexible mathematical functions to data. The functions might be neural networks, or classic statistical power series for regression. These approaches are used in Big Data, Machine Learning, and Artificial Intelligence. They might be termed data-based, or regression, or even model-free models. Many folks are reporting benefits from such data-based, empirical models.

As an example of a flexible modeling function, an auto-regressive moving-average (ARMA) model might predict the next $\dot{Q}$-value (a process flow rate response) from past $\dot{Q}$ and $\Delta P_{friction}$ and x-values (two influencing forces such as friction losses and valve position). In Equation (1), i is a time counter, and coefficients a, b, c, d, e, and f are adjusted to best fit the model to data.

$$\dot{Q}_{i+1} = a\dot{Q}_i + b\dot{Q}_{i-1} + c\Delta P_i + d\Delta P_{i-1} + ex_i + fx_{i-1} \qquad (1)$$

By contrast, in phenomenological models, the mathematical functions are grounded in a mechanistic relation. These models could be termed rigorous, first-principles, mechanistic, or cause-and-effect. First-principles models would be the same sort as those used in undergraduate instruction in process units, revealing basic principles, but not attempting to be rigorous about details of secondary importance.

Again, using the example of fluid flow inside a pipe, the fluid has momentum, and the driving force from pump and the retarding forces from friction losses and valve cause the fluid to accelerate or decelerate. Using Newton's law "$F = ma$" as a mechanistic starting point. The mass of fluid is acted on by several forces $\sum F = ma/g_c$, where $\sum F$ represents the sum of all the forces (where $F = A \cdot \Delta P$). Acceleration is the rate of change of velocity, $a = \frac{dv}{dt}$. With a simple numerical approximation to the calculus derivative, $\frac{dv}{dt} = \frac{v_{i+1} - v_i}{\Delta t}$, the $F = ma$ mechanism can be rearranged (by algebra, not calculus!) to predict the fluid velocity at the next time increment.

$$v_{i+1} = v_i + \Delta t \frac{g_c A\left(\Delta P_{pump,i} + \Delta P_{Head,i} + \Delta P_{friction,i} + \Delta P_{valve,i}\right)}{m} \qquad (2)$$

This represents an overall momentum balance. An overall balance of either thermal energy, component, or momentum is the first step in phenomenological, first-principles modeling.

The second step is to include constitutive relations in the model. In this example, these relate mass to pipe and fluid properties $m = \rho L \pi r^2$, volumetric flow to fluid velocity and cross sectional area $\dot{Q} = v\pi r^2$, friction losses to pipe friction factor and fluid flow rate $\Delta P_{friction} = f\frac{L_{equ}}{d}\frac{1}{2}\frac{\rho}{g_c}\left(\frac{\dot{Q}}{\pi r^2}\right)^2$, and the valve equation relating flow rate to pressure drop across the valve and valve stem position, $\dot{Q} = C_v x\sqrt{\frac{\Delta P_{valve}}{G\xi}}$ (this is a model for a valve with a linear characteristic). Doing so results in a first-principles type of phenomenological model.

$$\dot{Q}_{i+1} = \dot{Q}_i + \Delta t\frac{\pi r^2 g_c}{\rho L}\left(\Delta P_{pump,i} + \Delta P_{Head,i} + f\frac{L_{equ}}{d}\frac{1}{2}\frac{\rho}{g_c}\left(\frac{\dot{Q}_i}{\pi r^2}\right)^2 + G\xi\left(\frac{\dot{Q}_i}{C_v x_i}\right)^2\right) \quad (3)$$

I did not include the constitutive relation between pump head and speed and flow rate. And one might prefer a K-factor approach to the friction losses, or need to use an equal-percentage valve relation.

As a note, a first-principles model does not include phenomena that would be of secondary relevance. Such might include the thermal influences on density, friction factor dependence on Reynolds number, the volume in a pump volute or in an expander-contractor, or the 1.85 exponent value that better match pressure loss phenomena than the squared exponent of the Bernoulli ideal flow relations. One could progressively complexify the relation seeking to develop a rigorous model, and violate the K.I.S.S. principle. I don't recommend that, until you know that it is necessary.

In Equation (3) there are actually four influences on the flow rate, in Equation (1) only two. Equation (1) could be modified to model all 4 influences by adding 4 more terms which means a total of 10 coefficients that need to have values determined by best fitting the equation to data.

In contrast, all coefficients should be known in Equation (3). Perhaps the $f\frac{L_{equ}}{d}$ term may not be well characterized due to fouling or debris on screens, and it might be expedient to determine its value from operating data. This means that the first-principles model would have only one model coefficient that would need to be adjusted to best fit the model to data. That is a 10:1 reduction in complexity and data required to match the model to the process.

Now for a discussion of the pros and cons associated with empirical and first-principles models.

Pros for empirical model

- Simplicity: Empirical models are composed of repeating units of the same structure. In the Autoregressive-Moving-Average approach, it is the addition of linear terms multiplied by a coefficient. In a power series approach it is the variable raised to an integer power and multiplied by a coefficient. In neural network models it is the addition of exponential terms (neurons) multiplied by weighting factors. The simple elements and structure make for easy understanding and coding.
- Versatility: The same model structure is used for every application, regardless of the process. Even diverse processes associated with health care or education would use the same model structure that might be used for a chemical process.
- Effort: The model structure is defined. Greater engineering effort and skill are needed to derive first-principles models.

- Guidance: If a process is not mechanistically understood, then empirical modeling can identify possible input/output relationships. But use caution. 1) Correlation is not causation. Grey hair does not cause facial wrinkles. If the correlation is taken as causation, then dying one's hair might be hypothesized as a cure for facial wrinkles. And 2) a mechanism might not be visible within the historical data. Dennis Williams sent to me a link to https://www.tylervigen.com/spurious-correlations, about statistically significant correlations, and too-funny-to-believe explanations that GenAI gave for the connection.

Pros for first-principles models.

- Nonlinearity: The model naturally represents the actual phenomena. Although the number of units in empirical models can be increased to better fit model to data, it might take many. And each rise in complexity has an associated increase in the number of model coefficients that need to be fit to data. This nonlinearity includes steady-state sensitivity changes as well as non-stationary behavior (how transport delays and mixing lags change with operating conditions). Seek K.I.S.S.
- Training: First-principles models preserve mechanistic understanding. This helps operational staff understand fundamental phenomena, and discard folklore from long ago. Mechanistic understanding is essential for rational process analysis, abnormal situation detection, process improvement, and such activities.
- One-Model: The first-principles model is useful for design, operational optimization, control, training, maintenance forecasting, and other digital twin applications. Having "one model to rule them all" is efficient.
- No diversion of staff: Staff are not diverted to learn irrelevant mathematical techniques such as back propagation, or parsimony (choosing the number of empirical model units to best fit data with minimum complexity), or time-matching input and response variables, or the concepts of Laplace or FIR transforms. The models use engineers' mathematics.
- Extrapolation: Empirical models are best fit to a data set, but since they do not capture the mechanistic phenomena, extrapolation might be doubtful. First-principles models should extrapolate to conditions for which they were designed (such as within turbulent flow).
- Data acquisition: None, or very little, process upsets are required to generate data for first-principles model calibration or validation. By contrast a significant amount of data is needed to develop empirical models. There may be much historical data that could be used for training empirical models. But because of inconsistencies (such as data from fouled or clean service, or prior to process upgrades) or incomplete data, or operating at new flow rates, new data would be needed to generate empirical models. The data generation could be costly, time consuming, and risky. And, in my experience, such trials are often truncated by production overrides.
- Insensitivity to data abnormalities: These include noise, or other data aberrations.
- On-line monitoring: Process attributes, such as fouling, efficiency, or reactivity change in time. By observing how an adjustable coefficient changes in time, one can have a useful forecast for maintenance scheduling, and ever-changing constraint conditions.
- Simplicity of model calibration: There are only a few adjustable model coefficients, which means minimal process experimentation is required to calibrate and validate a first-principles model.

I prefer the use of first-principles models. But expediency might justify empirical modeling.

In subsequent Pods, I'll introduce how to create your own first-principles models, how to simulate environmental vagaries, how to calibrate and validate models, and how to use the models to evaluate the various economic indicators of transient events. I hope to visit with you later. Meanwhile, visit my web site www.r3eda.com to access information about modeling, control, optimization, and statistical analysis.