

Probabilistic Digital Twins for Transmission Pipelines

Francois Ayello
DNV GL
5777 Frantz RD
Dublin, OH 43017
United States

Guanlan Liu
DNV GL
5777 Frantz RD
Dublin, OH 43017
United States

Yonghe Yang
West Pipeline Company
CNPC
Lanzhou, Gansu
China

Yuchong Zhang
West Pipeline Company
CNPC
Lanzhou, Gansu
China

Long Li
West Pipeline Company
CNPC
Lanzhou, Gansu
China

Shuhui Zhang
West Pipeline Company
CNPC
Lanzhou, Gansu
China

ABSTRACT

Digitalization in the oil and gas industry has led to the formation of digital twins. Digital twins bring closer the physical and virtual world as data is transmitted seamlessly between real time sensors, databases and models. The strength of the digital twin concept is the interconnectivity of data and models. Any model can use any combination of inputs (e.g. operator owned data sets and sensors, third-party databases such as soil composition or weather data, results from other models such as flow assurance, threat modelling or risk modelling). Consequently, the result of one model may become the input of another. This strength is also a weakness, as uncertain (or missing data) will lead to a great source of uncertainty and may lead to wrong results. Worst case scenarios have been used to solve this issue without success. This paper presents a new concept: probabilistic digital twins for pipelines. Probabilistic digital twins do not lose uncertainty as results pass from one model to another, thus providing greater confidence in the final results. This publication reviews the probabilistic digital twin concept and demonstrates how it can be implemented using gas pipeline data from West Pipeline Company, CNPC.

Key words: Bayesian networks, Corrosion, Digital Twins, Erosion, Microbiologically Influenced Corrosion, Probabilistic Digital Twins, Risk Assessment, Uncertain Data, Uncertain knowledge

INTRODUCTION

One of the major current technology trends is “digital twinning”, identified by the journal Nature as a revolution in industry [1]. Stated briefly, a digital twin is a digital copy of a physical asset. Digital twin prototypes may be developed before the physical asset is constructed to evaluate performance, safety etc. In operation, the digital twin can be connected to sensors on the physical asset and updated based

on actual observations, to mirror the physical asset, and can be used for diagnostics and prognostics to support decisions in operations. Digital twins of a retired physical asset may be retained in a 'digital graveyard', to provide learning for other assets and operations.

In practice, a digital twin is a repository of knowledge about an asset. Some knowledge is encoded in models and some knowledge is contained in asset specifications (*i.e.* data) about the asset obtained during its lifetime. The novelty is not the existence of models and data about an asset, but the way in which these elements are connected and maintained over the asset's lifetime. The implementation of digital twins is gaining momentum with the rapid evolution of the internet of things (IoT) and deep learning technologies.

Knowledge of the asset, both about the conditions (*i.e.* data) and threats (*i.e.* subject matter expertise), will enable better planning of inspections and maintenance. The ability to simulate a system forward in time based on current conditions may also allow better forecasts of performance and safety.

Digital Twin Definition

The idea of digital twins is often attributed to Grieves, who presented the concept in 2002 in the context of Product Life-cycle Management [2]. The idea was that when designing and manufacturing a physical system, a virtual copy should be produced and live alongside it, mirroring the physical system throughout the lifecycle. Grieves and Vickers proposed the following definition of a digital twin [2]: "The Digital Twin is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level. At its optimum, any information that could be obtained from inspecting a physical manufactured product can be obtained from its Digital Twin."

Cabos and Rostock [3] propose that a digital twin is constituted by three basic components, which can be described in general terms as follows:

- A digital representation of the object, which relates to the system structure. Such representations could have a visual user interface (*e.g.* 3D model).
- Behavioral models, which describe the dynamic response of the system (*i.e.* its state and its outputs). Such models may include the operation of the system (*i.e.* for simulating how the system performs its intended functions), the structural response to the loads it is exposed to, as well as degradation and damage processes.
- The configuration or condition of the system, its data records of the past and current state of, or changes to, the system, captured through sensors, inspections etc.

Uses of Digital Twins

Digital twins may be used in all aspect of the asset life-cycle:

1. In the design phase, digital twins can be used to explore un-built assets in virtual reality. They can also be used for virtual system testing, *i.e.* to explore relevant scenarios and responses prior to building or deploying a physical asset, or when tests or experiments on the real asset would be costly, dangerous or time consuming to perform. Digital twins can also be used for testing and integration of software systems, *e.g.* by letting the control system control a digital twin instead of the real asset.
2. During the manufacturing phase, production processes can be simulated, and the product can be monitored and compared to the virtual design version. As Grieves and Vickers say [2], "We can collect the data from the machines that perform operations on the physical part to

understand exactly what operations, at what speeds and forces, were applied. For example, we can collect the torque readings of every bolt that attaches a fuel pump to an engine in order to ensure that each engine/fuel pump attachment is successfully performed.”

3. In operation, a digital twin may be continuously updated to reflect the current condition of the asset. These updates are based on sensor data, test and inspection results, conducted maintenance, modifications etc. Thus, the digital twin can serve as a repository of information which is easily retrievable during operations. Remote workers may be able to move around the digital twin in virtual reality, easing collaboration and communication with personnel on site. Similarly, workers on site may be able to retrieve information from the digital twin as they move about on the real asset wearing augmented reality technology. Beyond information retrieval, digital twins can also be used to predict future performance. They do this by both allowing simulation of the asset conditional on its current state and forecasts of its environment, and by extracting insights from the gathered experience data. Digital twins may allow complex and dangerous operations to be simulated before they are carried out in real life, thereby reducing the risk of complications or accidents.
4. In the decommissioning phase, after the physical asset is retired, the digital twin can be retained in a “digital graveyard”, providing learning for future projects. Accumulated data from digital twins of retired assets may also help in the development of improved designs and better predictive models for new or similar assets.

Digital Twin Limitations

The digital twin concept allows any model to use any combination of inputs. Consequently, the results of one model may become the input of another. This may lead to unpredictable uncertainty propagation depending on what data is used and the order in which the models are exercised. Data uncertainty (both inputs and predicted results) must be quantified and tracked as it is transformed from one model to another. In addition, a model’s uncertainty should not be ignored as no model is perfect and any model may introduce another source of uncertainty.

This is especially true for pipeline engineers, who must deal with uncertain data (including missing data) and incompletely understood threat mechanisms. Moreover, pipeline engineers’ decisions often have long term consequences extending many years into the future. These decisions may lead to different outcomes based on other seemingly unrelated parameters. This is a serious issue as digital twins use deterministic data and deterministic models. A simple solution would be to use worst case scenarios. However, worst-case scenarios are difficult to define. The worst-case scenario for one threat mechanism may be the best-case scenario for another mechanism. For example, a high flow rate may be considered as the worst-case scenario by an erosion specialist but as best-case scenario by a corrosion specialist (*i.e.* low probability of water accumulation). Moreover, worst case scenarios often lead to a lack of differentiation in risk results. Therefore, the digital twin’s deterministic approach is convenient but lacks the ability to deal with uncertain data and the possibility of multiple possible futures.

To overcome the limitation, we must address two needs: (1) probabilistic database where data is represented by probability density functions and (2) probabilistic models that consider uncertainty in the behavioral models (*i.e.* lack of knowledge).

TOWARD PROBABILISTIC DIGITAL TWINS

Bayesian networks have been used to address the two challenges presented above [4-8]. Bayesian networks are graphical models that encode probabilistic relationships among variables of interest. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes dependencies among all variables, it readily handles situations where some data entries are missing. Two, a Bayesian network can be used to learn causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention. Three, the model has both causal and probabilistic semantics; it is an ideal representation for combining prior knowledge (which often comes in causal form) and data [9]. These characteristics make Bayesian networks perfectly suited to create a probabilistic digital twin.

Moreover, Bayesian networks operate according to Bayes' theorem that states that the probability of an event can be estimated after observation of a correlated event (cause or consequence) according to the equation:

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{\sum_{j=1}^n P(B|A_j) * P(A_j)} \quad \text{Equation 1}$$

Where $P(A_i|B)$ is the posterior probability (*i.e.* what we want) of an Event A_i given the observation B . $P(A_i)$ is the prior probability of the event A_i before the observation B , and $P(B|A_j)$ is called the likelihood function and is the probability of the observation B given that event A_i occurred (*i.e.* knowledge). As shown by the equation, Bayesian networks only exchange probabilities. Therefore, it is easy to connect multiple models and models will exchange probabilities according to the equation above. It should be noted that information (*i.e.* probabilities) flows in any direction: from cause to consequence (*i.e.* deduction) or consequence to cause (*i.e.* inference).

Finally, Bayesian networks' probabilistic semantics addresses the two main issues discussed above: the need for probabilistic data and probabilistic models.

Probabilistic Data

Bayesian networks use probabilistic data. Probabilistic data deals with data uncertainty using probability density functions. The probability density function describes the relative likelihoods a variable might take and considers every possible state. The probability density function might be continuous or discrete as shown in this example of pipeline flaw depth:

flaw depth	0 - 10% ¹	=	11.0% probability
flaw depth	10 - 20% ¹	=	57.0% probability
flaw depth	20 - 30% ¹	=	22.0% probability
flaw depth	30 - 40% ¹	=	8.0% probability
flaw depth	40 - 50% ¹	=	1.9% probability
flaw depth	50 - 60% ¹	=	0.1% probability
flaw depth	60 - 70% ¹	=	0.0% probability
flaw depth	70 - 80% ¹	=	0.0% probability
flaw depth	80 - 90% ¹	=	0.0% probability
flaw depth	90 - 100% ¹	=	0.0% probability

¹ percent of total wall thickness

Even if the state is unlikely, it is considered (e.g. flaw depth 50-60% percent of total wall thickness = 0.1 %). Probabilistic inputs must consider the amount of certainty (or uncertainty) in the data. And probabilistic outputs must consider both the uncertainty of the data and the uncertainty of the models.

The probabilistic model must handle various type of data uncertainty (known data, unknown data or uncertain data). Figure 1 uses pH value to illustrate the various degrees of data uncertainty.

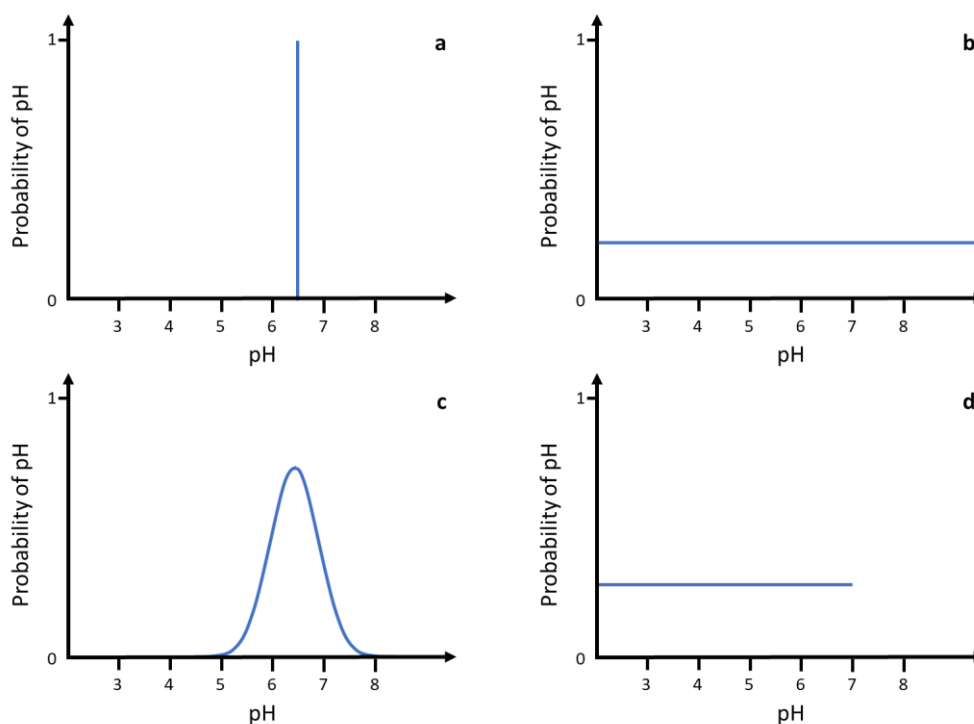


Figure 1: Examples of the data with various data of uncertainty

- Figure 1.a represents the data input distribution of known data without notable uncertainty. In this scenario, only one single value is available to be populated into the model. In Figure 1.a, the pH value is 6.5, and the probability is 1.
- Figure 1.b is an example of unknown data, which means there is no available information regarding the pH value, in this case, a uniform distribution represents the data probability within acceptable boundary conditions. It should be noted that boundary conditions are set using subject matter expertise, some experts might use the 0-14 pH scale while others might consider that it is unlikely to have a pH below 3 and higher than 8 (in the presence of a weak acid like of CO₂).
- Figure 1.c shows an example of uncertain data approximated using a normal function. Data might be uncertain because it changes in time or because measurement are not easy to obtain. Understanding of how the data was obtain and a statistical analysis of the data will help decide what probabilistic function is most appropriate.
- Figure 1.d shows an example of uncertain data where measurements are not available. For example, if a report reads “the pH of the water phase in the pipeline is acidic” this information must be used. In this case, a uniform distribution using boundary conditions (similar to Figure 1.b) is the best option.