Inspectioneering Journal

ASSET INTEGRITY INTELLIGENCE

Featured Article

Using Data Science to Enhance Reliability: Four Real-World Applications

Dr. Andrew Waters, *Director, Data Science at Pinnacle*
Fred Addington, *Principal, Corrosion Technology at Pinnacle*

VOLUME 28, ISSUE 5
SEPTEMBER | OCTOBER 2022

Using Data Science to Enhance Reliability: Four Real-World Applications

Dr. Andrew Waters, *Director, Data Science at Pinnacle*
Fred Addington, *Principal, Corrosion Technology at Pinnacle*

Introduction

The rise in computational power over the last decade has begged the question of if and to what extent quantitative methods such as data science have in improving reliability programs. While data science has the power to revolutionize the reliability industry, it will only be able to do so with strong guidance and review from subject matter experts (SMEs).

The combination of SME and data science enables facilities to develop solutions to a variety of reliability challenges based on each method's unique strengths. SMEs provide a wealth of knowledge, much needed context, and experience that has proved instrumental in making facilities safer and more reliable. Data science, combined with machine learning (ML) techniques, has revolutionized how facilities sift through a tremendous volume of data and find insights in near real-time.

The ability to make better decisions by leveraging data continues to be a theme across the industry and will help decision-makers make more informed strategic decisions at a faster pace. This article will highlight the efficacy of a combined SME and data science approach by showing four example applications:

1. Using equipment data and associated corrosion rates across multiple reformer units to show how predictive models using data science compare to traditional industry templates and expertise-driven models.
2. Leveraging Bayesian statistics to introduce uncertainty into remaining life calculations and probability of failure, empowering the expert to define variables better to identify and reduce uncertainty, improve equipment remaining life estimations, and reduce overall risk.
3. Leveraging data science to quantify the confidence of damage detection, including driving benefit to cost for taking readings on or omitting particular condition monitoring locations (CMLs).
4. Leveraging natural language processing on CMMS and IDMS data to identify anomalies for equipment that should have been flagged for positive material identification but were not.

In each of these applications, we will discuss the challenge and how bringing various data science methodologies into the solutions approach allowed experts to make quicker and more strategic decisions yielding enhanced outcomes.

Modeling Corrosion Rates with Facility Data

Challenge: Can we leverage historical inspection data to create the next generation of accurate corrosion models?

Corrosion estimates are typically determined by SMEs using a variety of methods. First, SMEs make heavy use of industry standard tools, such as API specifications that map facility conditions to corrosion rates. Additionally, SMEs typically review historical inspection data to get an idea of how corrosion has manifested in the past. They will often rely on the wealth of experience they have accumulated to predict how corrosion may manifest in the future.

There are several limitations with current methods. For example, the theoretical corrosion rates provided by industry standard tools may differ significantly from how corrosion manifests at a given facility. Each facility is unique and will experience its own unique corrosion profiles due to different environmental conditions, maintenance and operation practices, and other factors. Further, while historical data provides an important touchpoint for corrosion rate analysis, the sheer volume of available data can be overwhelming for a human SME to analyze adequately. Finally, while an SME's experience can provide valuable insight, in some cases, SMEs can be subjective and may not necessarily serve as an accurate predictor of corrosion.

ML techniques provide an alternate method for estimating corrosion rates on equipment. Rather than rely solely on data science alone to develop a solution, it is important to incorporate SME input directly into each stage of the analysis.

Example

For example, we recently completed a study where a data science model was able to successfully predict corrosion rates with a higher degree of accuracy than standard industry tools alone [1]. The study started with a dataset of CMLs from 20 facilities across seven different major refineries. Each CML data point contained the actual measured corrosion rate taken from the historical inspection data along with other operating and process variables for the CML, including the operating temperature, operating pressure, metallurgy, and stream constituency information. The data science model then learned how the various operating and process variables influenced corrosion rates across each facility. Each of these relationships was validated by SMEs to ensure that the general trends observed by the model were meaningful and not simply the result of overfitting to the dataset. Once the data science model is trained, it can provide corrosion estimates on new CMLs that it hasn't observed previously. The efficacy of the model was further validated by comparing the model predictions directly with industry standard tools, and the results of this comparison are shown in **Table 1**.

The result of the study found that the data science model outperformed the human SME on 8 of the 11 cases under consideration and had a significant error improvement.

Asset ID	Comp ID	Meas. Rate (mpy)	Industry Rate (mpy)	Industry % Error	ML Rate (mpy)	ML % Error
1	A	11.4	1	91%	7.5	34%
2	A	0.8	10	1150%	4.2	425%
3	A	18.9	6	68%	3.2	83%
4	A	6.9	4	42%	4.9	29%
5	A	1.9	6	216%	3.1	63%
5	B	3.3	6	82%	4.6	39%
6	C	3.8	6	58%	2.6	32%
6	C	2.1	6	186%	2.6	24%
7	A	3.1	6	94%	3.3	6%
8	A	8.2	6	27%	3.9	52%
8	B	3.3	3	9%	2.9	12%

Metric	Industry	Model
Mean Abs Error	5 mpy	**3.1 mpy**
Mean % Error	184%	**73%**

pinnaclereliability.com

Table 1. Corrosion Rates Predicted by ML Model vs. Industry Rate.

Modeling Equipment End of Life

Challenge: How can we combine historical inspection data and subject matter expertise to estimate equipment end-of-life more accurately?

A combined approach can be applied to modeling equipment end of life. Data science applications can help SMEs better define variables to reduce uncertainty, associated remaining life, and overall risk. By being able to predict the end of an asset's life more accurately, facilities will be more equipped to minimize equipment failures that can result in loss of profit for the facility.

In the case of thinning, end-of-life estimation for an SME comes down to interpreting two different sources of data: the SME-estimated corrosion rate and the historical inspection data. SMEs have their limitations when estimating corrosion rates, and while measured inspection data can provide information regarding the actual rate of degradation, its utility is often marred by measurement error. This measurement error can skew the perception of the degradation rate and cause facilities to make inaccurate predictions regarding end of life.

Example

A better approach to estimating equipment end of life is to combine the SME-estimated corrosion rate directly with the inspection data. One way to accomplish this is through the use of a data science model called the lifetime variability curve (LVC). The LVC model begins with an end-of-life prediction based solely on the SME-estimated corrosion rate and refines its estimates as more inspection data becomes available. A graphical example of this model is shown in **Figure 1**. The LVC model on the left only contains the nominal thickness and projects thickness over time as a function of the SME rate alone. This model produces an estimate of when thickness will reach a critical value and can be translated into a probability of failure (POF) curve that informs us how likely the asset is to have failed on or before a target date. As more inspection data is made available, the LVC refines its estimates regarding thickness over time and the POF. In the second plot, the model includes two inspection points that seem to tell a consistent story—degradation is occurring at a relatively constant rate that is less than the SME estimated rate. With only two data points, the LVC model balances the SME estimate and the data estimate and ends up with considerable uncertainty regarding the likely failure data, as shown by the large blue band of uncertainty. In the right panel, two more inspection points are added to the model. Now, the overall corrosion rate strongly agrees with the SME rate, causing the estimates regarding potential failure dates to converge considerably, as shown by the narrowing band of uncertainty.

In general, the LVC model will place more emphasis on the SME-estimated rate when there is little data or when there is little trust in the data. As the amount of measurement data increases and the data begins to tell a consistent story, less emphasis will be placed on the SME rate in favor of the data rate.

The LVC model relies solely on a corrosion rate provided by the SME and measured inspection data to estimate remaining useful life. The model is adaptive in that it takes into account any observed changes in degradation and adjusts its estimation appropriately. Further, it would be possible for the LVC model to take in any measured changes in process conditions and leverage that information to improve its overall estimation. Note that while this example discusses thinning, these techniques can be applied to a wide range of damage mechanisms, including vibration fatigue or cracking.

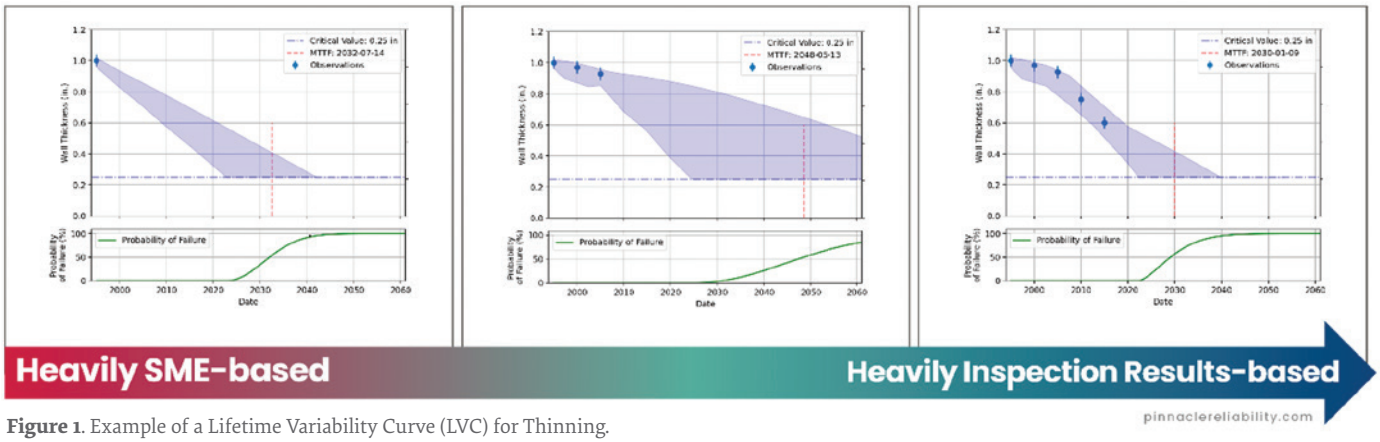

Figure 1. Example of a Lifetime Variability Curve (LVC) for Thinning.

Inspection Prioritization

Challenge: Can we leverage data science to augment a subject matter expert's ability to create effective inspection plans?

SMEs can leverage data science to quantify the confidence of damage detection to better drive benefit-to-cost for taking readings on or omitting particular CMLs.

Example

Imagine a scenario with LVC models for multiple CMLs on an asset. Asset failure will occur when any one of those CMLs fails. Now, say that there's an opportunity to inspect the asset. We can certainly inspect every CML and update the individual LVC models, but this is often a waste of resources since many of the CMLs being inspected pose no risk to the overall lifetime of the system. Instead, it would be a better use of resources to focus on the CMLs that present a significant risk of driving failure.

Recall that the LVC model provides an estimate of POF over time. Given a consequence of failure (COF), we can compute risk over time as:

Risk = POF x Consequence

Assume now that we are planning inspections at some point in time. If we have an estimate of the risk posed by each CML at this point in time, we can simply choose the CMLs that exceed a predefined risk threshold. Consider the scenario in **Figure 2**, where we plot the risk profile for four CMLs over time. The inspection date is specified as the vertical blue line, and the risk threshold of $100,000 is shown as a horizontal black line. Here we see that CMLs 0, 1, and 2 exceed our risk threshold and must therefore be inspected. CML 3, on the other hand, has not exceeded the risk threshold and does not require inspection at this time. This methodology can be applied to larger sets of assets and see potentially larger gains.

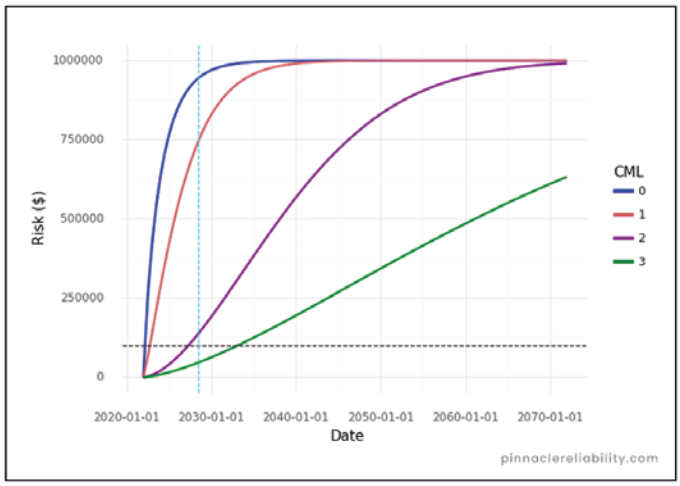

Figure 2. Risk Profile of Four CMLs.

Example

In another example, a recent case study examined six assets and found that only 50% of CMLs in the overall population required inspection. The study focused on three units which included six piping systems and more than 300 CMLs. These assets had extremely high corrosion rates due to sulfidation and erosion. The total inspection cost for the facility's previous inspection approach was estimated at $738,000 and covered two turnarounds in the future. By leveraging a combination of SME and data science, the facility was able to identify the CMLs that actually required inspection. Reducing the number of CMLs that needed to be inspected not only provided significant cost savings in terms of the overall inspection cost (as illustrated in **Table 2**). but more importantly, enabled the facility to focus its resources where they matter most rather than trying to over-inspect.

Additionally, by leveraging data science, the facility is able to optimize an inspection plan while ensuring that the overall risk

exposure of the facility is well-controlled. SME input was also present throughout this process when guiding the LVC model and reviewing the final results to ensure they were reasonable for the facility under consideration.

	Total Cost	# CMLs	Risk
Traditional Inspection Approach	$738,000	382	$97,163
Risk-Based Inspection Approach	$358,000	190	$99,860
Total risk with no inspection is approximately $100MM			

Table 2. Traditional Inspection Approach vs. Risk-Based Inspection Approach.

Leveraging Natural Language Processing to Create Efficiencies and Improvement in Quality Control

Challenge: How can SMEs leverage data science to improve quality control practices?

Lastly, a combined approach can be applied with natural language processing (NLP). NLP combines linguistics, data science, and artificial intelligence (AI) that leverages computer programming to process and analyze substantial amounts of natural language data. NLP can apply machine learning to various types of document-related tasks, such as mining U1 forms, and can evolve tasks, such as inspection grading by eliminating the need for people to read reports and make subjective judgments about inspection quality [2]. In addition to being time-intensive for humans, these tasks can also be prone to error. Leveraging data science techniques to get a computer to do these things automatically is a huge win on costs and accuracy [3].

Example

For example, a refinery was having a problem with its positive material identification (PMI) program. When new equipment arrived at the facility, a technician would log the equipment into a database with a text field describing what the equipment was for and where it would be used in the facility. After entering the text information, the technician would check a box indicating whether PMI was required. In reviewing their records, it became clear that there were a number of entries in their database for equipment that, based on the text field, should have undergone PMI but were ultimately not put through that process. The database consisted of nearly 500,000 entries, and it would be infeasible for a human to check all these entries manually for discrepancies.

A data science method was developed to comb through the database entries and search for anomalies. This method operates by learning which words have historically been identified with PMI along with the words that have not historically been identified with PMI. **Table 3** shows a selection of keywords that were extracted by the data science method, which indicated whether or not PMI was needed. These word sets were reviewed by SMEs to ensure that the machine learning method was learning accurate relationships and not failing to consider relevant terms.

PMI Positive Terms	PMI Negative Terms
residual	cs / carbon
347ss	pvc
5cr	spiral
monel	gasket
hastelloy	copper

Table 3. Keywords Extracted by the Data Science Method.

From these associations, the method was then able to analyze every text field in the database and assign not only a PMI status but also a level of confidence associated with the estimated status. **Table 4** shows an example of these classification labels. Again, SME review was critical to ensure that the model functioned as intended. During the SME review, it was noted that all entries with a confidence score over 55% were found to be very accurate. As confidence dropped below 55%, indicating that the algorithm is very uncertain of how to label an entry, it was noted by the SMEs that the text entries themselves were too ambiguous for a human to label with any real degree of confidence.

Entry	PMI Required?	Confidence?
sockolet, 2" x 3/4" 6000# lre ** low res pipe fittings and flanges - carbon steel	True	99%
bolt, cap,heavy hex head 5/8" x 2",321ss stud bolts and nuts	True	80%
valve, gate, 3/4" 800# thd, t:12, nace, valves forged steel and parts	False	87%

Table 4. Examples of PMI Outputs Labeled with Associated Confidence.

Ultimately, this method was able to analyze all 500,000 entries from the facility and produce a list of around 1,000 items that were very likely to require PMI and were not labeled as requiring PMI by the facility. This list created an immediate action list for the facility to help close potentially high-risk issues in their facility.

Conclusion

Equipping SMEs with data science methodologies holds significant promise for improving the reliability of facilities. While each technique has its strengths, the four applications discussed above highlight examples where neither the SME nor data science techniques alone could have achieved the same level of success. Ultimately, combining SME knowledge and data science methods will provide a more comprehensive, powerful, and efficient solution for facilities to improve their reliability. What problem could data science help you solve? ■

For more information on this subject or the author, please email us at inquiries@inspectioneering.com.

REFERENCES

1. Pinnacle, 2020, "Data-Driven Reliability - Using Big Data to Model Degradation in Reformer Units," https://pinnaclereliability.com/learn/white-papers/data-driven-reliability-using-big-data-to-model-degradation-in-reformer-units/
2. Pinnacle, 2022, "How Machine Learning is Changing the Data Organization Game," Inspectioneering, https://inspectioneering.com/blog/2022-06-13/10152/how-machine-learning-is-changing-the-data-organization-game.
3. Pinnacle, "Data Science Modeling in Reliability," Pinnacle Reliability, https://pinnaclereliability.com/learn/topics/data-science-modeling/.

The Next Evolution of Mechanical Integrity is here.

Quantitative Reliability Optimization (QRO) empowers leaders at industrial facilities to make better reliability decisions. QRO is an evolution in modeling that combines the best traditional reliability methods, data science, and subject matter expertise into one hybrid model.

Learn more at pinnaclereliability.com

1950's:
Reactive Maintencance

1970's:
Time-Based Inspection

1990's:
OSHA
Occupational Safety and Health Association
PSM Regulated

2000's:
Risk-Based Inspection

Today:
Quantitative Reliability Optimization

PINNACLE
DATA-DRIVEN RELIABILITY

CONTRIBUTING AUTHORS

Fred Addington

With more than 27 years in the industry, Fred Addington serves in the role of Principal of Corrosion Technology at Pinnacle. As Pinnacle's foremost corrosion subject matter expert, Fred is responsible for training, advising, and disseminating technical knowledge to Pinnacle's project teams and developing efficient management processes so team members can deliver quality services and solutions to customers. Fred's expertise includes his understanding of various processes and equipment in the area of corrosion control and material selection in the upstream and downstream oil and gas, water, and petrochemical industries. He has published papers on topics such as high-temperature sulfur corrosion, amine unit corrosion on stainless steel, and the application of hydrogen permeation in corrosion monitoring. Specifically, his knowledge includes corrosion and metallurgical analysis, corrosion control and monitoring, material selection, hydrogen permeation technology, and mechanical integrity. Fred earned his Bachelor of Science in Metallurgical Engineering from the University of Texas at El Paso after serving in the United States Navy.

Dr. Andrew Waters

Dr. Andrew Waters is Chief Data Scientist at Pinnacle, focusing on developing data-driven algorithms to enhance a variety of reliability and maintenance applications. Dr. Waters also specializes in utilizing machine learning methods to improve and augment human decision making. He has utilized these skills across a diverse set of industries including finance, communication systems, engineering, signal processing, optimizing student learning outcomes, and hiring and recruitment programs. Dr. Waters holds a doctorate in Electrical and Computer Engineering from Rice University and is the author of over 20 publications in the areas of signal processing, machine learning, and Bayesian statistical methods. His research interests include sparse signal recovery, natural language processing, convex optimization, and non-parametric statistics.