Inspectioneering Journal

ASSET INTEGRITY INTELLIGENCE

FEATURED ARTICLE

Don't Throw Out Your IDMS Data: Use Statistics to Rebuild Trust

Vyacheslav Nadvoretskiy, Ph.D., *Principal Data Scientist at Pinnacle*
Andrew Waters, Ph.D., *Director, Data Science at Pinnacle*

VOLUME 29, ISSUE 5
SEPTEMBER | OCTOBER 2023

Don't Throw Out Your IDMS Data: Use Statistics to Rebuild Trust

Vyacheslav Nadvoretskiy, Ph.D., *Principal Data Scientist at Pinnacle*
Andrew Waters, Ph.D., *Director, Data Science at Pinnacle*

Introduction

As an industry, we spend a tremendous amount of money, time, effort, and resources collecting data. Despite this, how often do we mistrust the data we have collected and worry about using our data to make reliability decisions? We recognize that quality data is a critical component of a successful mechanical integrity or reliability program. However, erroneous data coupled with a lack of trust prevents us from extracting its full value, possibly limiting the quality of our reliability programs.

In many cases, the result of not trusting data is, paradoxically, collecting more data through inspections, which can further exacerbate reliability challenges. In addition to adding to the volume of data to sift through, over-inspection costs valuable time and money that could be spent on more valuable activities that mitigate asset risk and increase facility uptime.

Suspect data comes in a variety of forms. A typical example is inspection data for thinning over time. A facility might not trust data from a specific time frame or may question the process that was used to collect that data. Even when the data is mostly trustworthy, it is easy to find anomalous data points that don't make sense (e.g., readings above nominal, readings below critical, etc.). Another common example of suspect data is when asset operating parameters such as temperature, pressure, and metallurgy are missing, which limits the effectiveness of a subject matter expert (SME) to accurately estimate potential corrosion mechanisms and rates for fixed equipment.

In a recent study across 15 global refineries from six different operators, we identified four common industry data integrity challenges:

1. Outlier data readings
2. Growth in thickness readings
3. Potential repairs/replacements
4. Missing data

In this article, we discuss these common challenges and dive into how facilities can leverage data science and statistical techniques to quickly identify and potentially correct or quarantine suspicious data to create more stable data analytics that can be utilized to drive more confident decisions and value from the data.

Outlier Data Readings

The first common data integrity problem is outlier data readings. Outlier data readings are data points that deviate from other data points within the same data set. If outlier data points are removed, then there would be a consistent trend in data points over time.

For example, consider the plot of thickness data for a given condition monitoring location (CML) over time in **Figure 1a**. The thickness measurements are mostly within expected parameters and show a relatively consistent corrosion rate over time. A regression line in this plot shows the average corrosion rate over time and the confidence interval of the regression estimate, shown by the shaded gray region around the regression line. A full treatment of confidence intervals is beyond the scope of this article, but an excellent introductory treatment can be found in Reference 1. The confidence interval shows the amount of uncertainty associated with the corrosion rate. Any line that can be drawn in the shaded region is the potential true corrosion rate of the given CML. Most practitioners would be fairly happy with this data and would feel comfortable making predictions about an asset's health from this data.

However, now consider the data in **Figure 1b**. Here, the data is largely consistent but with one exception—a single outlier thickness reading (outlined with a red circle), which dramatically increases the uncertainty of the analysis.

It is straightforward for software to compute confidence intervals on TML data and flag any data points falling outside the confidence interval. Points that are flagged as being potential outliers can then be reviewed by an SME or can be corroborated by collecting additional inspection data. After detecting and removing the outlier data point from the regression model (**Figure 1c**), the size of the confidence interval is dramatically reduced, and the overall trend of the data becomes very clear.

In our analysis, we found anomalous data points account for less than 2% of all inspection data. Despite this being a small number relative to the total population of measurements, this 2% can dramatically impact estimated corrosion rates, so isolating them from analysis is crucial.

Growth in Thickness

Another common occurrence in thickness data is to see CMLs that show growth over time, as shown in **Figure 2a**. An increase in thickness measurements often leads to mistrust in data since, taken at face value, the data shows that the thickness of the asset increases over time. To offset this mistrust, facilities often remove growth measurements from datasets in order to create "sensible" CML data. Removing data, however, typically leads to increased uncertainty as well as missed opportunities for understanding the underlying mechanisms of asset degradation.

Data science tools provide a better approach to managing growth corrosion rates, and doing this requires us to consider the confidence interval of our data again. When we consider the

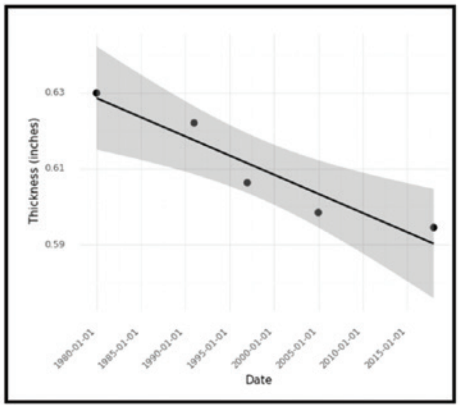


Figure 1a. A typical set of CML data points along with a best-fit regression line (solid black) and confidence interval (shaded gray). Overall, this data is well-behaved, and the overall trend shows consistent degradation over time.

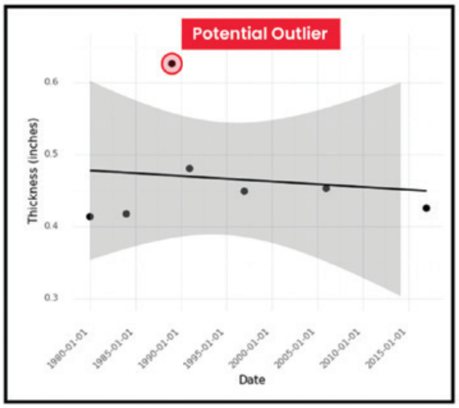


Figure 1b. A single outlier point (red circle) dramatically increases the size of the confidence interval.

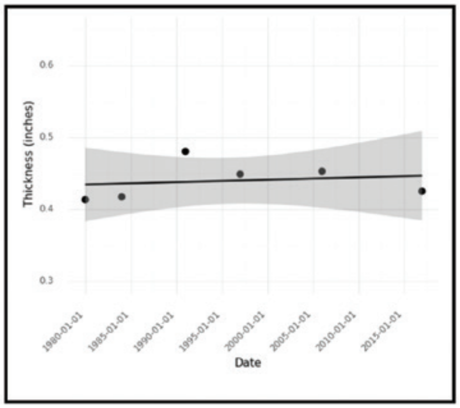


Figure 1c. After removing the outlier, the confidence interval shrinks significantly, and the overall data trend becomes clear.

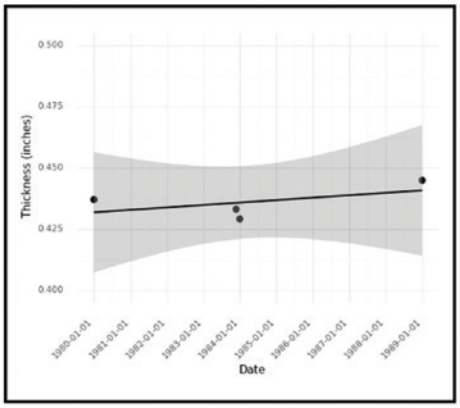


Figure 2a. A typical CML dataset that shows a growth in thickness over time.

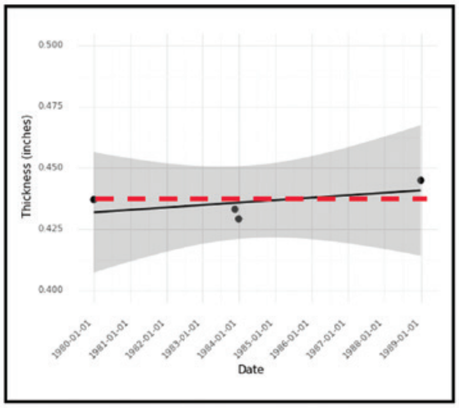


Figure 2b. For the data of **Figure 2a**, we find that the confidence interval is wide enough that a zero-corrosion rate scenario is entirely plausible. This is the case for most growth data flagged by practitioners.

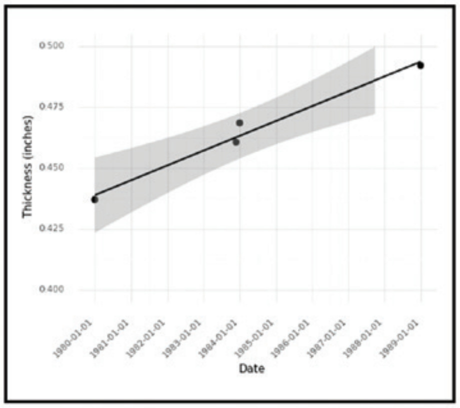


Figure 2c. An example of a growth CML that cannot be ruled out as zero corrosion with some statistical noise. Examples like **Figure 1c** account for a small portion of typical industry data.

confidence interval of the data, we find that most growth CMLs typically indicate a scenario where corrosion is essentially zero (see **Figure 2b**). In this case, we can easily ignore the growth corrosion rate since it is statistically equivalent to the scenario in which the corrosion is virtually non-existent. Additional SME input or data collection may, however, be required when the confidence interval does not allow us to make this decision (see **Figure 2c** for an example).

Our analysis of data from multiple facilities reveals that the case of growth corrosion rates is typically attributed to statistical insignificance. While roughly 20% of CMLs in our dataset are flagged as growth CMLs, only about 3% of CMLs show growths that are statistically significant (see **Figure 2c** for an example). These subsets of CMLs are typically attributed to bad data readings that throw off the analysis and require some human attention or further data collection to resolve.

Potential Repairs/Replacements

In the previous section, we discussed growth CMLs and how most CMLs that get flagged as growths are cases of extremely low corrosion with some statistical noise. A common cause for the remaining growth CMLs is often undisclosed repairs and replacements (see **Figure 3a**, **3b**), which can have a significant impact on the reliability and availability of industrial assets. Our comprehensive analysis of facility data shows that these occur, on average, across 10% of components. While it's relatively straightforward for a human to identify a potential repair or replacement in a small dataset, the challenge arises when scaling to larger datasets, where manual inspection becomes impractical.

Data science offers a solution to this scaling challenge. Advanced analytics algorithms can flag potential repairs or replacements, allowing for SME confirmation. A simple way to do this is by looking for large jumps in CML data where we move from a thickness value close to critical to values closer to nominal thickness. In the case of a single CML (**Figure 3a**), it is quite possible that we may falsely identify a repair that has not occurred or miss a repair that has actually occurred. Our accuracy improves significantly when we consider all CMLs on the component or piping circuit jointly

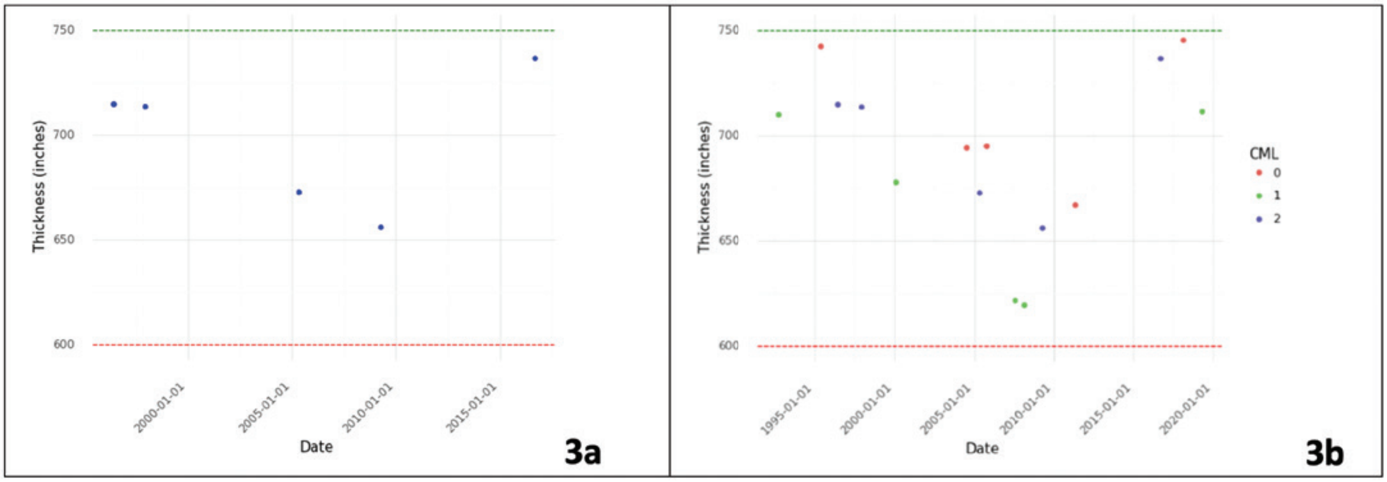
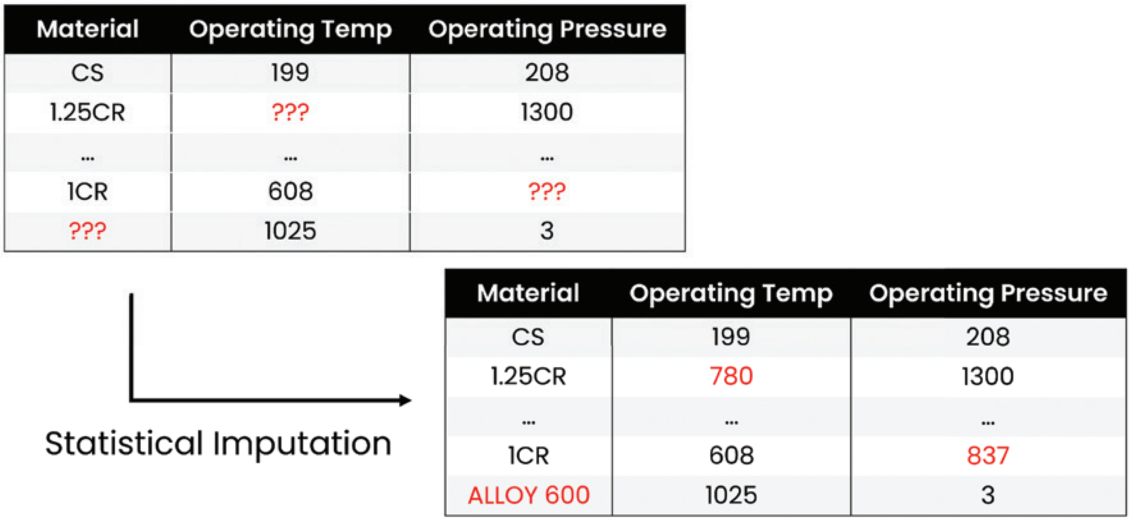


Figure 3a. A single CML suggesting a repair/replacement occurred between 2010 and 2015. While it may be difficult to identify the repair with only a single CML data point, it becomes very evident when looking at multiple CMLs together (**Figure 3b**).



Material	Operating Temp	Operating Pressure
CS	199	208
1.25CR	???	1300
...	...	...
1CR	608	???
???	1025	3

Material	Operating Temp	Operating Pressure
CS	199	208
1.25CR	780	1300
...	...	...
1CR	608	837
ALLOY 600	1025	3

Figure 4. Missing value imputation. Statistical methods can be used to estimate missing data values with incredibly high accuracy by leveraging the observed data for similar assets in a dataset.

(**Figure 3b**). This not only saves time but also enhances the accuracy of our predictive models. Ignoring these potential repairs or replacements can lead to inaccurate corrosion rate predictions, affecting the integrity and reliability of assets.

Missing Data

Our focus thus far has been on noisy or anomalous data. A final major problem faced by the industry is missing data. This is often the case for asset-critical data such as materials of construction, operating temperatures, and stream information. This data is crucial for damage mechanisms reviews and RBI as it is used in various risk-based calculations. However, it is often incredibly time-consuming and expensive to dig through documentation to find a particular missing data point needed for a given RBI calculation. With data science, facilities can better-focus on the data that has a strong impact on mitigating risk.

For example, consider a missing operating temperature for an asset in a refinery. For simplicity in this example, we assume that we have operating temperatures for every other asset in the refinery. The operating temperature will vary wildly across the refinery, say from 50 degrees to over 1000 degrees. Given this extremely wide range of temperatures, it may seem impossible to say anything meaningful about our missing temperature of interest. However, when we start looking at the other data for our asset, we often find that we can reduce our uncertainty dramatically. Knowing, for example, that our asset is a pressure vessel constructed from carbon steel with an operating pressure of 25 PSI may reduce our temperature uncertainty dramatically. Further knowing the chemical stream properties of the material contained in the pressure vessel will reduce our uncertainty further.

This process of using statistics to estimate missing data is known as *statistical imputation* or *missing value imputation* (see **Figure 4**) and can have extremely impressive accuracy [2,3]. In analyzing facility data across multiple operators and sites, we found that a variety of common variables could accurately be recovered. For operating temperature, our imputation methods are generally within about 20 degrees Fahrenheit of the correct temperature. For material specification, our methods could impute the correct material in 95% of all cases. The imputation methodology could also recover critical limits (T_{min} values) to 0.02 inches of accuracy, on average. Furthermore, even in cases where large uncertainty for a given parameter is still present, we may find that the uncertainty has little impact on risk calculations. Methods like this can then be used to prioritize further data collection by focusing on the data that can have a strong impact on risk and focusing less on data with a smaller impact on risk.

Conclusion

Quality data is a critical component of a successful mechanical integrity or reliability program. Data science, machine learning, and statistics can help identify, interpret, and, at times, correct bad data without a tremendous level of human effort. Doing this ultimately enables facilities and reliability practitioners to obtain greater value from their data, ultimately leading to a better understanding of asset risk and how best to mitigate that risk.

For more information on making your data more usable, watch Pinnacle and Inspectioneering's August 2023 webinar, Using Data Science to Eliminate Your Fear of Bad Data for Predictive Asset Management. ■

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

REFERENCES

1. Illowsky, B. and Dean, S., 2013, Statistics, OpenStax, https://openstax.org/details/books/statistics.

2. Gelman, A. and Hill, J., 2006, Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press, Chap.25.

3. Zhang, Z., 2016, "Multiple imputation with multivariate imputation by chained equation (MICE) package," Annals of Translational Medicine, 4(2).

CONTRIBUTING AUTHORS

Vyacheslav Nadvoretskiy, PhD

Dr. Vyacheslav Nadvoretskiy is a Principal Data Scientist at Pinnacle, where he develops innovative solutions to generate data-driven insights into the reliability of industrial assets. He specializes in developing algorithms and software tools to process customer data and assess and minimize risks, improving asset reliability and efficiency. Dr. Nadvoretskiy holds a PhD in Math and Physics from the Moscow Technical University of Electronics and Computer Science (Moscow, Russia), where he also studied Biomedical Engineering. He has over 20 publications in the areas of magnetic nanofluids, ultrasound and photoacoustic tomography, signal and image processing, and machine learning applications. His research interests include image reconstruction and processing algorithms, medical ultrasound, and industrial applications of artificial intelligence.

Andrew Waters, PhD

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.

Unlock the full potential of your data.

When it comes to data quality, we are all familiar with the term "garbage in, garbage out." But with the plethora of data available today, distinguishing good data from bad can be challenging.

ON DEMAND WEBINAR

3 ways data science can help you rebuild trust in your data:

1. Prioritize human focus to enhance predictive asset management strategies
2. Eliminate manual efforts
3. Ensure effective resource allocation

For more information, watch the recent webinar "Using Data Science to Eliminate Your Fear of Bad Data for Predictive Asset Management" at pinnaclereliabililty.com

PINNACLE