Estimating ILI Tool Performance in Identifying Unique Populations for Material Verification

Ryan Stewart[1], Jason Skow[1], Ryan Okamura[2], Mohammad Al-Amin[2], Zain Al-Hasani[2], Luigi Calabretta[2]

[1]Integral Engineering
[2]TC Energy

Organized by

CLARION
TECHNICAL CONFERENCES

https://doi.org/10.52202/078572-0047

https://doi.org/10.52202/078572-0047

Abstract

49 CFR §192.607 defines unique populations for pipeline material verification. These are defined by wall thickness, pipe grade, manufacturing process, manufacturing date, and construction date. Using in-line inspection (ILI) to identify pipe attributes has the potential to significantly improve the efficiency of satisfying the material verification requirements outlined in 49 CFR §192.607 for delineating unique populations along a pipeline. However, some of the ILI technologies used in this process are relatively new, and ILI vendors do not yet provide a performance specification for their performance in identifying populations.

This analysis explores quantifying ILI tool performance in defining pipe populations by similar heuristic rules. It incorporates them into a probabilistic model specification for ILI population identification for three attributes: wall thickness, pipe seam detection, and yield strength. The models were fit using populations identified by the ILI vendor and validated using records review and field verification by an operator. Models adequate for characterizing the ILI tools, assumptions regarding pipe populations, and heuristics that can be used as proxies for the ILI vendor's interpretation of measurement data are discussed.

The methodology outlined in this study aims to quantify ILI's ability to classify transmission pipeline segments into populations with unique material properties to comply with the Pipeline and Hazardous Materials Safety Administration (PHMSA) material verification requirements. The framework uses statistical methods similar to those outlined in API 1163 for ILI measurement validation, which the industry has been using for about twenty years. These models are designed to estimate the accuracy of future population identifications and identify populations that are more likely to be misclassified, thereby informing an operator's decision-making to prioritize population validation efforts.

Introduction

49 CFR §192.607(b), Documentation of material properties and attributes, (PHMSA 2019) states:

> *Records established under this section documenting physical pipeline characteristics and attributes, including diameter, wall thickness, seam type, and grade (e.g., yield strength, ultimate tensile strength, or pressure rating for valves and flanges, etc.), must be maintained for the life of the pipeline and be traceable, verifiable, and complete.*

ILI tools have historically been used to identify pipeline attributes such as diameter, wall thickness, and presence of long seam. Newer ILI technologies show promise in identifying additional pipe properties to help determine pipe grade based on estimates of yield strength and ultimate tensile strength. Using ILI to identify pipe attributes is likely to become a central tool in material verification programs that assist in satisfying the requirements outlined in 49 CFR §192.607.

https://doi.org/10.52202/078572-0047

However, ILI technologies for material property determination are relatively new, and ILI vendors do not yet provide a performance specification for population identification. As a result, there are essential questions to address. For example, given an ILI report defining pipe populations, how can we characterize identification performance? How do we validate future ILI runs? Before answering these questions, a high-level understanding of the process used to identify populations and the tool performance for each attribute being used to discern populations must be reviewed.

Population Identification Procedure

In this paper, the population identification performance is based on our understanding of the procedure used by an ILI vendor, which follows a hierarchical procedure where the following attributes subdivide populations:

1. Diameter
2. Wall thickness
3. Seam detection
4. Yield strength and ultimate tensile strength
5. Joint length
6. Pipe segment location (e.g. facility piping, crossings, population density)

The populations are classified using a combination of ILI measurements and data interpretation by the ILI vendor. It is straightforward to subdivide populations based on substantial differences in attributes across pipe segments. For example, if a significant difference in wall thickness is measured between two adjacent pipe joints, they are likely unique populations. The same applies to changes in the presence or absence of pipe seams.

Identifying populations becomes more complex when it is based on yield strength, and typical unsupervised machine-learning algorithms are likely to fail. Populations are only sometimes contiguous, and clustering algorithms will have difficulty correctly characterizing the small intricacies involved in determining whether two segment groups should be split or aggregated. Therefore, there is inherent utility in incorporating engineering judgement when defining individual populations. ILI vendor judgement allows subtle nuances to be considered, such as the history of line pipe manufacturing, pipeline ownership, and other details, which are difficult to capture with rule-based methods or automated clustering.

ILI measurement tools and the subject matter expertise used to distinguish populations can be considered two different sources of uncertainty in population identifications: Type A and Type B (JCGM, 2008).

- Type A pertains to the uncertainty inherent in the statistical analysis of the ILI tool's series of observations (i.e., the tool's performance).

https://doi.org/10.52202/078572-0047

- Type B pertains to uncertainty by means other than statistical analysis from a series of observations (i.e., ILI vendor skill/experience in interpreting ILI data).

This paper discusses the attributes used to identify unique populations: wall thickness, seam detection, and yield strength. Each attribute is evaluated in terms of its performance when used to subdivide populations ILI data.

Performance Model

General Approach

At its simplest, population identification is a binary classification problem. Measurements from an ILI tool and data review can attempt to identify populations based on segment attributes, and the ILI vendor's calls can be correct or incorrect. There are two ways the tool can be correct (true positive and true negative) and two ways the tool can be incorrect (false positive and false negative); see Figure 1. In the case of population identification, these outcomes have the following definitions:

- True positive: an identified population is a distinct, unique population
- True negative: no distinct population is identified, and none exists
- False positive: an identified population is part of an existing population and not a unique population
- False negative: no distinct population is identified, but a distinct population exists

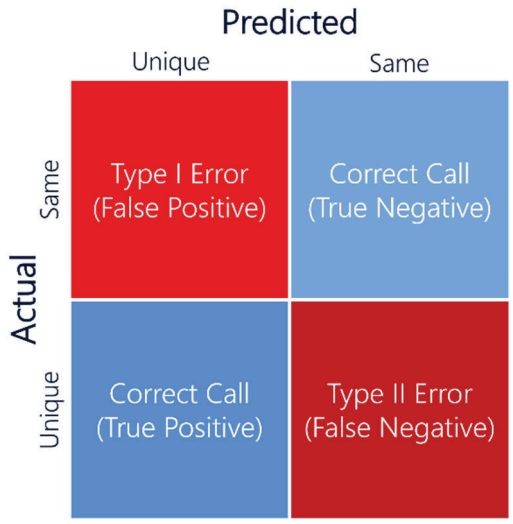

Figure 1. Possible Outcomes of a Binomial Event.

In the context of population identification, tool performance can be quantified in terms of either population precision or population recall.

Recall refers to the probability of identifying unique populations and the ability to avoid missing sub-populations within the data. Recall relies on the bottom row in Figure 1 to define the number of correctly distinguished populations out of all populations:

https://doi.org/10.52202/078572-0047

Recall = TP / (TP + FN)

Where:

TP = true positives (unique populations correctly identified within a set of ILI measurements)
FN = false negatives (unique populations missed within a set of ILI measurements)

Precision relies on the left column in Figure 1 to determine the number of correctly identified populations out of all populations the ILI attempted to identify as unique:

Precision = TP / (TP + FP)

Where:

TP = true positives (unique populations correctly identified within a set of ILI measurements)
FP = false positives (a single population incorrectly identified as multiple within a set of ILI measurements)

The goal for population identification performance is to determine methods to quantify recall and precision. For example, when dividing populations based on yield strength, what is the recall or probability of correctly identifying populations for a subset of measurements?

In many cases, binomial performance is simplified as a single ratio that does not vary with other parameters. When expressed as a ratio, the performance metric can be illustrated as a single point with an upper and lower whisker representing the confidence interval. The confidence interval is commonly calculated using the exact methodology (Clopper and Pearson, 1934) or an accepted approximate method such as Agresti and Coull (1998).

Sometimes, the performance is a function of a variable. For example, deeper anomalies are easier to detect than shallower anomalies. Therefore, detection probability as a function of anomaly depth is a reasonable model. This concept can be extended to multiple dimensions, where the recall or precision of a prediction is a function of multiple parameters. Precision or recall is generally expressed as one of the following terms.

Recall as a single ratio:

$Recall_j = \frac{TP_j}{TP_j + FN_j}$ for j in wall thickness, seam detection, and yield strength.

Recall as a function of a single variable:

$Recall_j = f_j(x_j)$ for j in wall thickness, seam detection, and yield strength.

Recall as a function of multiple variables:

$Recall_j = f_j(x_{j1}, x_{j2}, x_{j3}, \ldots x_{jn})$ for j in wall thickness, seam detection, and yield strength.

For each attribute, three checks are performed:

1. Is Precision or Recall (Type 1 or Type 2 error) the dominating factor in incorrect calls?
2. Should the Precision/Recall be modelled as a single value, a function of a single variable, or a function of multiple variables?

https://doi.org/10.52202/078572-0047

3. If it is a function, which model and variables characterizing the data should be used to fit the performance model?

Each attribute that identifies unique populations requires a model to estimate the probability of correct identification using unique measurements and heuristics.

Model Approach

Performance estimation based on each attribute is calculated by comparing the populations identified by ILI tool to those identified by reviewing historical records and in-field verifications. This method is similar to the API 1163 Level 3 method that estimates as-run tool performance from field validation data.

Two approaches are discussed in API 1163 Level 3: statistical tolerance intervals (method 1) and Bayesian inference (method 2). For this paper, Bayesian inference was chosen to evaluate population identification performance for the following reasons (McElreath, 2020):

1. There is no minimum required sample size.
2. The shape of the resulting curves considers the sample size.
3. The prediction is not a point estimate, allowing for full probabilistic analysis (engineering judgment can be used for deterministic applications).
4. All assumptions about population performance are made upfront. The proposed models deduce the assumptions' real-world implications to quantify population identification performance.

In the data used to fit models, the recall and precision of the data used indicate good performance across all populations, on average exceeding 95% for both attributes (populations are rarely missed based on the attributes used to delineate them, and incorrectly dividing a single population into to two more is infrequent). Due to the limited number of incorrect calls in the validation data, the dataset is imbalanced. This imbalance can significantly affect the performance of conventional models. However, the Bayesian approach offers a robust solution by quantifying the uncertainty arising from the imbalanced data and increasing flexibility in characterizing our models, if necessary. By leveraging Bayesian inference, the variability and confidence in predictions are better understood, even when dealing with limited and imbalanced data.

Binomial Regression

For all population identifications, the observed population identifications are a set of counts of the number of successes y_i out of n total trials. In other terms, based on a particular attribute used to distinguish populations, there is a proportion of times the identifications are correct, p_i. Here, i indexes different subsets of the data based on specific attributes or data characteristics. The appropriate method to characterize these classifications is the binomial distribution (Vincent, 2022):

$$y_i \sim \text{Binomial}(n, p_i)$$

Where:

y_i = number of successes (number of correct calls) for the i-th subset of data (e.g., the subset defined by a particular attribute or data characteristic)
n = number of trials (number of observations) in the i-th subset
p_i = probability of successes (proportion of correct calls) for the i-th subset

https://doi.org/10.52202/078572-0047

For any unique population, the goal is to estimate the probability that it is a correct call; either two populations are correctly identified as different, or one group of pipe segments is correctly identified as one population. The goal of fitting each probability of the correct identification model is to estimate p_i for a given predictor variable or set of predictor variables, x_i or $x_{i1}, x_{i2}, \ldots x_{in}$ respectively. A Bayesian model is developed for each attribute with its unique probability of success formula.

Wall Thickness Model

The difference in measured wall thickness is the first attribute used to subdivide pipe segments into unique groups. The limitation of distinguishing populations based on wall thickness is the measurement performance of the ILI tool.

Figure 2 shows the nominal wall thickness measurements from an ILI as a function of the pipeline distance. We can determine if the call was correct for every unique population identified based on the difference in wall thickness between them. For example, the ILI tool may have incorrectly identified two unique segments of pipe with a wall thickness difference of 0.031 inches even though they were, in truth, the same wall thickness, but also correctly identified all unique populations with a wall thickness difference of 0.22 inches (transitions from 0.281-inch to 0.5-inch wall thickness). Figure 3 shows the example scenario as a proportion of correct population identifications. The proportion of correct calls in Figure 3 is a function of the difference in reported wall thickness measurements on the same pipeline. The size of each data point is proportional to the number of observations by wall thickness difference. A data point is added at the origin since unique populations cannot be identified without a difference in measured wall thickness.

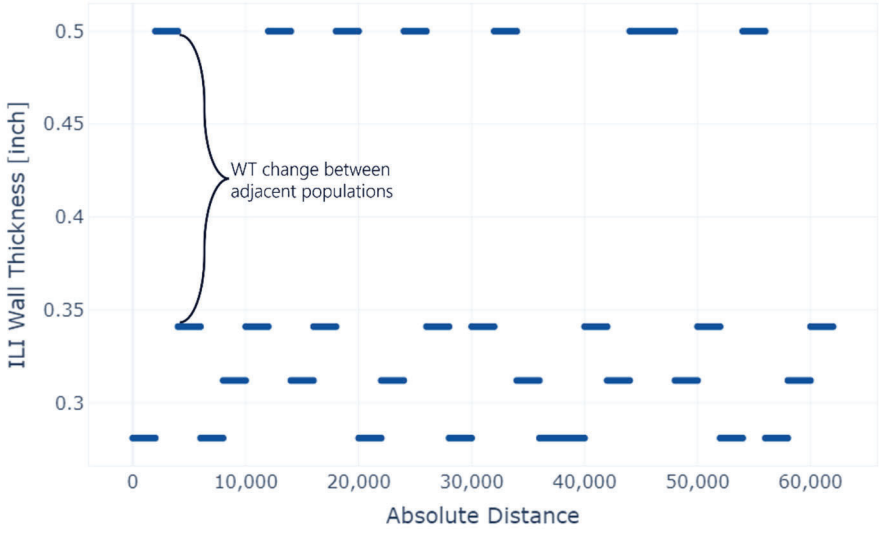

Figure 2. Wall Thickness ILI Measurements along a Pipeline

https://doi.org/10.52202/078572-0047

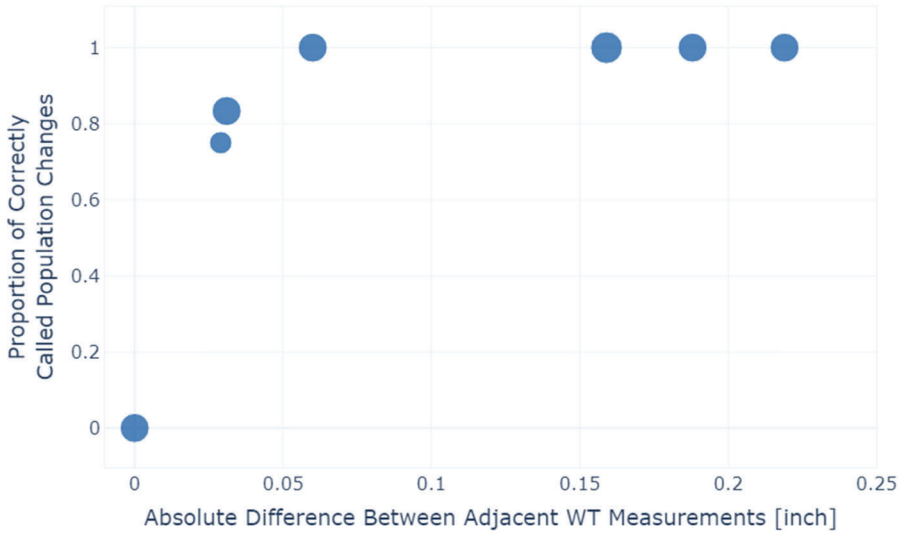

Figure 3. Proportion of Correct Calls as a Function of Wall Thickness Change

The wall thickness data in Figure 3 is used to model the probability of correctly identifying unique populations as a function of wall thickness differences. Multiple models were explored to fit this data. While conventional practice suggests fitting the data to logistic or exponential regression, a model based on first principles was developed, focusing on ILI wall thickness measurement error.

When two independent wall thickness measurements are made, we can estimate the probability that the two measurements are the same. Let X_1 and X_2 be the two measurements. We assume both measurements have a normally distributed measurement error with mean 0 and standard deviation σ. The true values of the measured quantities are μ_1 and μ_2. We also assume that the nominal wall thicknesses accurately represent the measurements. Any deviation between the nominal and true wall thickness is negligible compared to the ILI measurement performance. The measured values can be expressed as:

$$X_1 = \mu_1 + \epsilon_1 \text{ and } X_2 = \mu_2 + \epsilon_2$$
$$\text{where } \epsilon_1 \sim N(0, \sigma^2) \text{ and } \epsilon_2 \sim N(0, \sigma^2) \text{ are the measurement errors}$$

Using the measurements X_1 and X_2, we can determine what is more likely: $\mu_1 = \mu_2$ or $\mu_1 \neq \mu_2$. We calculate the difference, d, between the two measurements X_1 and X_2:

$$d = X_1 - X_2 = (\mu_1 + \epsilon_1) - (\mu_2 + \epsilon_2) = (\mu_1 - \mu_2) + (\epsilon_1 - \epsilon_2)$$

If $\mu_1 = \mu_2$, d should be close to 0. Since ϵ_1 and ϵ_2 are normally distributed with the same variance, d is normally distributed:

$$d \sim N(\mu_1 - \mu_2, 2\sigma^2)$$

Under $\mu_1 - \mu_2 = 0$:

https://doi.org/10.52202/078572-0047

$$d \sim N(0, 2\sigma^2)$$

The Z-score can then be calculated using the difference between X_1 and X_2:

$$Z = \frac{X_1 - X_2}{\sigma\sqrt{2}}$$

The probability that the two measurements are the same can be interpreted as the likelihood of observing a value as extreme as d under the assumption $\mu_1 = \mu_2$. The standard normal distribution can be used to find the probability:

$$P(-d \leq Z \leq d) = 2\Phi\left(\frac{|d|}{\sigma\sqrt{2}}\right) - 1$$

Where:

Z = standardized Z-score
Φ = cumulative distribution function of the standard normal distribution
d = observed difference between two wall thickness measurements
σ = standard deviation of the ILI wall thickness measurement error

The final probability is the p-value corresponding to the Z-score. The p-value is interpreted as the probability of correctly identifying two populations as distinct based on the difference in wall thickness measurements observed by the ILI tool. The only unknown parameter in the model is σ, the standard deviation of the ILI measurement performance. Note that σ is different from ILI performance tolerance. The performance tolerance is a predefined and known threshold that specifies the maximum acceptable deviation between the ILI measurements and wall thicknesses that can be distinguished, whereas σ represents the statistical measure of how much the wall thickness measurements typically vary. The model for correctly identifying populations based on wall thickness has two parts:

$$y_i \sim \text{Binomial}(n, p_i)$$
$$p_i = 2\Phi\left(\frac{|d_i|}{\sigma\sqrt{2}}\right) - 1, \text{ for } i = 1, \ldots, I.$$

Figure 4 is a graph showing the curve for some potential values of the parameter σ compared to the observations. The size of each data point is proportional to the number of observations.

A Bayesian model is used to fit the most likely curve. Given the data, Markov Chain Monte Carlo calculates the posterior distributions over the model parameters (Abadi M, et al., 2015). After calculating the posterior of σ, the posterior of the binomial regression curve is used to estimate the probability of correctly identifying unique populations as a function of the difference in wall thickness measurements. Figure 5 shows a posterior predictive of the binomial regression using the posterior of σ.

https://doi.org/10.52202/078572-0047

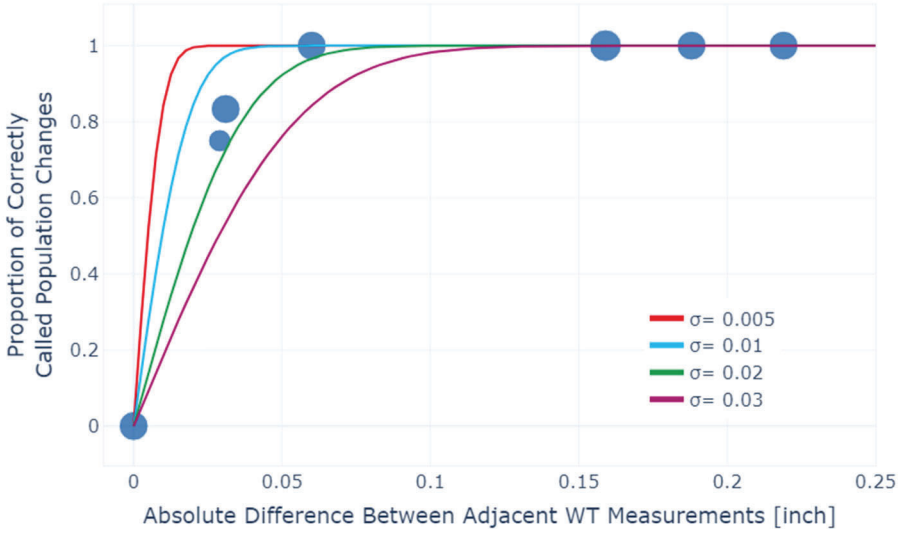

Figure 4. Probability of Correct Call Model Priors – Wall Thickness

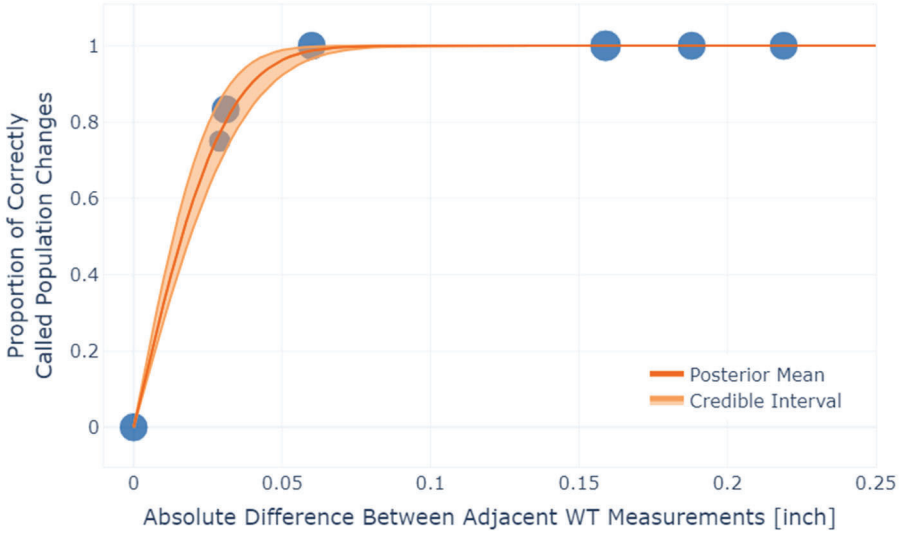

Figure 5. Probability of Correct Call Model Posterior – Wall Thickness

The dark orange line in Figure 5 shows the mean predicted probability of correct call. The light orange area represents the curve's credible interval, which can be interpreted as the uncertainty in the true value of the parameter σ based on the sample size of the observed data.

The curve fit is used to determine where populations identified based on a difference in wall thickness are most likely or least likely to be incorrect. When the difference in wall thickness is low, confidence that two populations are unique is also low. Other attributes may be needed to identify unique populations or additional field verification efforts may be warranted.

https://doi.org/10.52202/078572-0047

Seam Detection

Once pipe segments are grouped and classified into sub-populations based on measured wall thickness, the next step is to divide the populations further based on whether the segments have a detected seam. Seam welded and seamless pipes with the same wall thickness represent two unique populations. Discussions with subject matter experts indicate that ILI tools can accurately identify the presence or absence of seams, making it unlikely that detecting seams leads to incorrect population identification. In a typical scenario, the probability of correct identification based on seam detection is estimated as the proportion of correct calls.

$$\hat{p} = \frac{x}{n}$$

Where:

$\hat{p}$ = probability of correctly identifying unique populations based on detecting a change in the presence of seam between adjacent pipes
x = number of correct populations identified based on the change in the identified seam
n = number of changes in identified the presence or absence of seam

An estimated proportion may be inaccurate when x = 1.0, x = 0.0, or the sample size is small. A closed-form estimate exists when using the Beta distribution to model the prior of $\hat{p}$. The Beta distribution is a continuous probability distribution defined over the interval [0, 1]. It is characterized by two shape parameters, denoted as α and β. When using a Beta distribution to characterize the probability of correctly identifying unique populations based on the presence of a seam, the mean estimator is:

$$\hat{p}_b = \frac{x + \alpha}{n + \alpha + \beta}$$

Where:

$\hat{p}_b$ = mean probability of correctly identifying a unique population based on the presence of a seam
α = Beta distribution first shape parameter
β = Beta distribution second shape parameter

Using a Bayesian estimate with an uninformed prior (α=1, β=1), the distribution of the mean estimator aligns with the Clopper Pearson interval used for calculating a binomial confidence interval.

Yield Strength

Given yield strength measurements, the model identifies whether a single identified population is more than one by estimating the recall (true positives divided by true positives plus false negatives).

https://doi.org/10.52202/078572-0047

Yield strength measurements from ILI runs are used to subdivide populations. Characterizing the performance of ILI tools that estimate yield strength and ultimate tensile strength is ongoing, and there is uncertainty regarding the bias and measurement error. However, even without precise performance characteristics, measurements are used to distinguish populations using the following assumptions:

- The bias and uncertainty of yield strength measurements are consistent across an ILI run, and
- Yield strength for an individual pipe grade follows a unimodal distribution, similar to how it is characterized in API 1176 Table D.1 (2016) and CSA Z662 Table O.6 (2023).

Unlike comparing just two measurements, as in the wall thickness example, two populations are now being compared. The increased sample size makes it more challenging to determine whether individual samples come from one population or another, particularly if the distributions of the two pipe grades overlap by less than four standard deviations.

Figure 6 is an example of overlapping yield strength distributions. It is difficult to conclude if a single sample (dashed green line) is from pipe grade X52 or X60. In these instances, the more conservative pipe grade (lower yield strength) is typically selected.

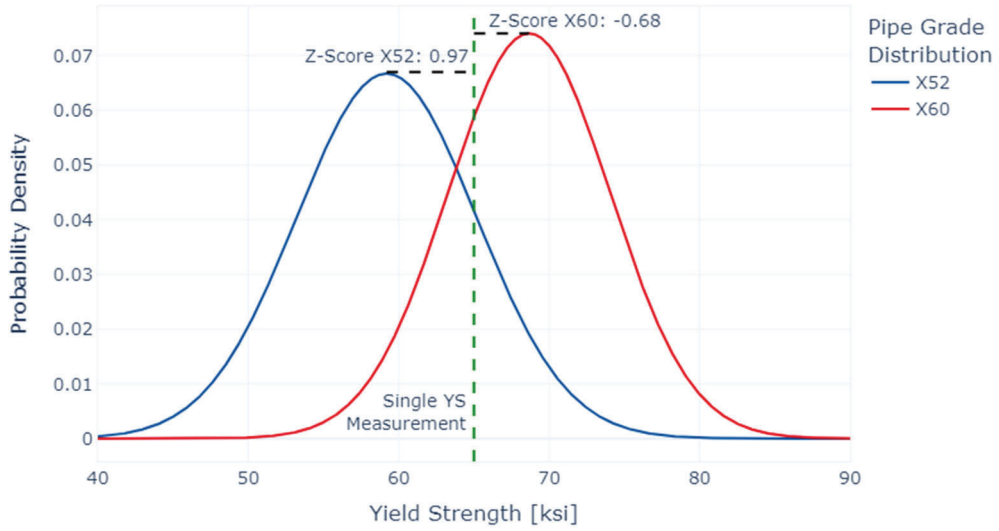

Figure 6. Comparing the Z-score of Two Yield Strength Distributions from API 1176

Although pipe grade is challenging to determine from a single measurement, the distribution of many samples can provide insight into whether one or more populations are within a group. This model aims to determine the likelihood that a population identified by the ILI vendor is a single population or multiple sub-populations.

As part of this study, cases that led to incorrect identification of populations based on yield strength were reviewed. An example is shown in Figure 7. A sample of pipe segments with similar wall

https://doi.org/10.52202/078572-0047

thickness and identified seams was initially assigned to a single population when records indicated two pipe grades interspersed along the chainage. When plotting the data in Figure 7 as a histogram in Figure 8, the multimodality of the distribution is observed; this population is likely a mixture of two populations with different yield strength distributions.

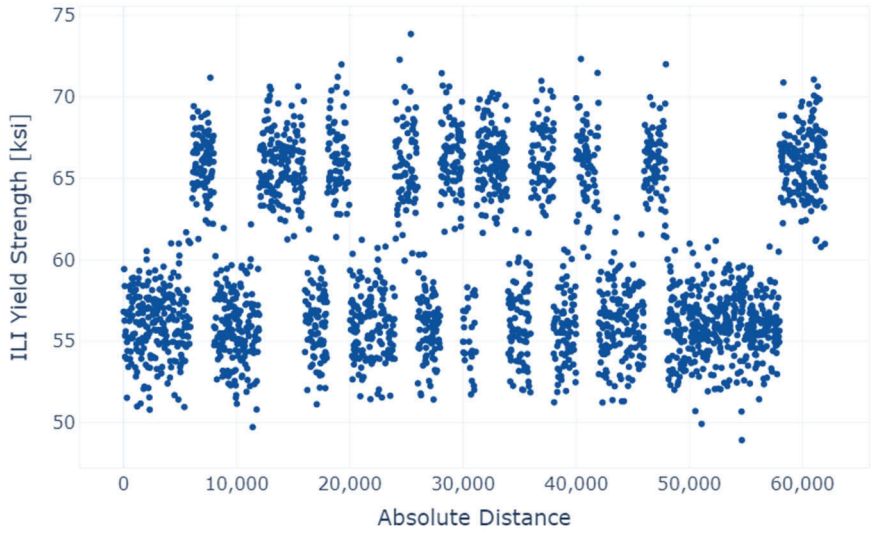

Figure 7. Yield Strength of Pipeline Segments by Chainage

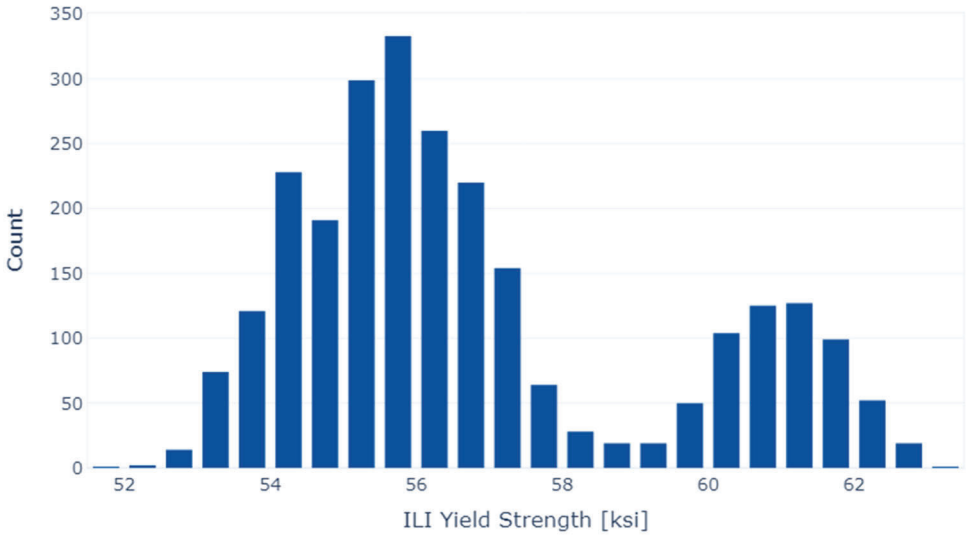

Figure 8. Yield Strength Histogram of Pipeline Segments

Applying a Gaussian Mixture Model (GMM) clustering algorithm to the pipe data in Figure 8 produces a Bayesian Information Criterion (BIC) difference when the number of subpopulations is specified. The BIC is a heuristic for model selection among a finite set of models, where models with a lower BIC are typically preferred. Figure 9 shows the results of applying a GMM to the population in Figure 8. The left sub-chart shows the results of a GMM with two clusters (k=2), and the right plot shows the change in BIC depending on the number of clusters applied to the data. In this case, using

https://doi.org/10.52202/078572-0047

a GMM with two distributions (k=2) versus a single distribution (k=1) significantly lowers the BIC, suggesting that modelling the data as two distributions with unique means and standard deviations is a better fit than treating it as a single distribution. The BIC values also suggest that there are not more than two populations in the sample since adding additional clusters does not further reduce the value of the BIC.

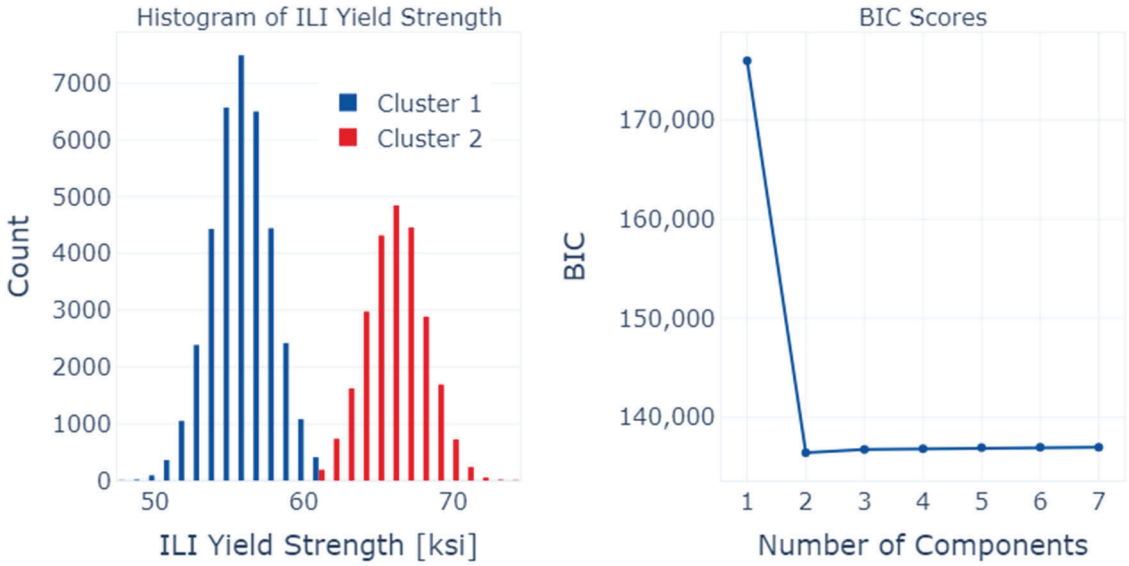

Figure 9. GMM with k=2 on Yield Strength Data

Even though a Gaussian Mixture Model suggests two populations, an engineering decision may group them differently. Data in the middle could arise from either distribution. In this case, assuming a lower pipe grade may be appropriate when performing engineering analysis. Regardless, we quantify the likelihood of correctly identifying unique populations in this project, even if combined populations are justified by engineering judgment.

Data reviewed indicates that distribution modality is a useful indicator for sub-populations. The estimate of the probability of whether there is one pipe population in a sample of segments is characterized as a function of the distribution modality. There is no universally agreed-upon summary statistic to quantify the modality of a sample; Sarle's bimodality coefficient is used to characterize the data in this paper (Ellison, 1987). The bimodality coefficient is calculated as follows:

$$\beta = \frac{\gamma^2 + 1}{\kappa}$$

Where:

β = bimodality coefficient
γ = skewness
κ = kurtosis

The formula for a finite sample is:

https://doi.org/10.52202/078572-0047

$$b = \frac{g^2 + 1}{k + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

Where:

b = bimodality coefficient for a finite sample
g = sample skewness
k = excess kurtosis
n = number of samples

The value of β for the normal distribution is 1/3 (skewness = 0, kurtosis = 3). The value for the uniform distribution is 5/9 (skewness = 0, kurtosis = 1.8). Values greater than 5/9 may indicate a bimodal, multimodal, or heavily skewed unimodal distribution.

Figure 10 shows the bimodality coefficient for each group of pipes identified as a unique population based on pipe grade. The populations are classified as either 1.0 (correctly identified as one population) or 0.0 (incorrectly identified as one population). The data point with the highest bimodality coefficient is the same example shown in Figures 7 to 9. This data indicates that a performance prediction as a function of the bimodality coefficient offers good discrimination; the model can distinguish correct and incorrect calls. However, the data are also very imbalanced between correct and incorrect calls. The model may need to be revisited if additional incorrect calls that do not show multimodality are identified.

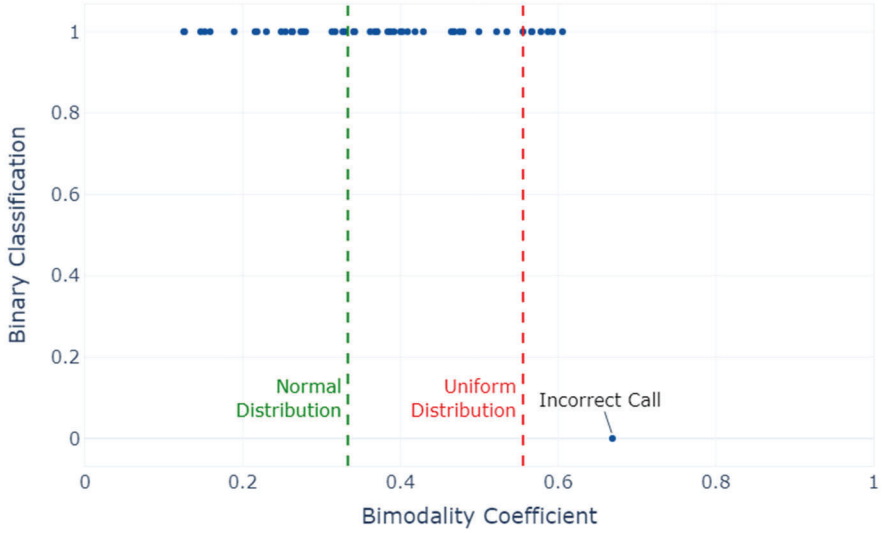

Figure 10. Correct and Incorrect Calls as a Function of Yield Strength Bimodality

The data was fit to the inverse logit function to provide the probability of incorrect call as a function of the bimodality coefficient:

$$p_i = g^{-1}(-[\beta_0 + \beta_1 \cdot b_i])$$

https://doi.org/10.52202/078572-0047

Where:

g^{-1} = inverse of the logit function (logistic sigmoid function)
β_0 = intercept of the untransformed linear model
β_1 = slope of the untransformed linear model
b_i = bimodality coefficient

A Bayesian model is fit with the following process:

$$\beta_0 \sim Normal(loc, sigma)$$
$$\beta_1 \sim TruncatedNormal(loc, sigma, lower = 0)$$
$$y_{observed,i} \sim Binomial\left(n_{observed}, InverseLogit(-[\beta_0 + \beta_1 \cdot b_i])\right), for\ i = 1, \dots, I$$

Where:

$y_{observed,i}$ = number of observed successes for bimodality coefficient
$n_{observed}$ = number of observations for bimodality coefficient
b_i = bimodality coefficient for observation

The data imbalance highlights the advantages of a Bayesian approach. While the limited number of incorrect calls indicates the effectiveness of ILI measurements in distinguishing populations, it also poses challenges in model fitting and predicting potential errors. For Bayesian models, it is crucial to determine the priors. Prior knowledge suggests that a bimodality coefficient greater than 5/9 indicates subpopulations. This aligns with our data; however, we use caution in our prior selection to avoid selecting priors that would overly influence the outcome. Figure 11 shows an example of a posterior prediction based on simulated data.

This model will not identify a large population that is incorrectly divided into two populations. Rather, it only identifies false negatives: one population is called, but more than one population is present.

The bimodality coefficient was chosen as a suitable heuristic to characterize the same inductive reasoning that a subject matter expert might employ to subdivide populations following a review of ILI measurements. It was chosen over BIC and GMM clustering due to its simplicity. If there is insufficient confidence to subdivide a population, all pipe segments are assigned to a single grade, typically the grade represented by the lower tail of the entire sample.

https://doi.org/10.52202/078572-0047

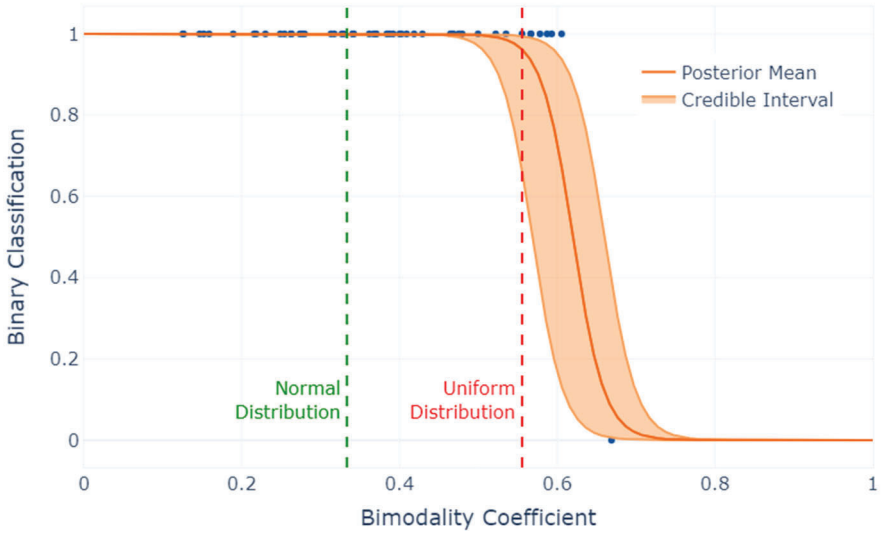

Figure 11. Probability of Correct Call Model Posterior – Yield Strength

Performance Model Results

Table 1 summarizes the models that characterize the performance of an ILI tool in identifying unique pipeline populations. For populations divided based on wall thickness, the most likely error and the only one observed in the dataset is a false positive. This occurs when two similar populations are incorrectly separated due to uncertainty in ILI wall thickness measurements. Similarly, false positives are the most likely error for seam detection, where two unique populations are identified when only one truly exists. In contrast, the observed errors are false negatives for populations identified by yield strength measurements. A single population is identified in these cases, while the underlying data contains multiple sub-populations.

Table 1. Model Summary

Model Order	Attribute	Probability Model	Estimating
1	Wall Thickness	$p = 2\Phi\left(\frac{\lvert d\rvert}{\sigma\sqrt{2}}\right) - 1$ Where d = difference in WT between two populations	Probability that two populations identified as different are actually different (false positive)
2	Seam Detection	$\hat{p}_b = \frac{x + \alpha}{n + \alpha + \beta}$	Probability that two populations identified as different are actually different (false positive)
3	Yield Strength	$p = \text{InverseLogit}(-(\beta_0 + \beta_1 \cdot b))$ Where b = Bimodality Coefficient	Probability that one identified population is not actually made up of more than one sub-population (false negative)

https://doi.org/10.52202/078572-0047

Each model captures precision or recall. Seam detection uses a simple binomial confidence interval. Wall thickness employs a single-parameter prediction based on tool error. Yield strength fits a sigmoid curve based on statistical summaries that mimic ILI data interpretation and the shape of pipe grade distributions.

One challenge in developing a population identification model is the imbalanced dataset, which contains only a few incorrect calls. The Bayesian models used produce credible intervals that reflect the uncertainty inherent in the data imbalance. Future ILI vendor performance specifications for population identification could be incorporated as informed priors to address this uncertainty. Additionally, model updates may be necessary to account for potential changes in ILI vendor procedures and the inclusion of new measurement data. Future ILI runs, and population identifications can be validated against these models. A validation framework and statistical test are proposed next to determine if future performance fits these models.

Validation Framework

Performance Validation

In the future, new ILI runs may be validated using the models presented in Table 1. Validation refers to evaluating the accuracy of the population identifications compared to the expected performance from the models. The validation procedure uses a statistical test to determine whether population identifications are performing as the models suggest.

In the validation process, the main objective is to determine if the ILI vendor performs worse than the past performance in which the model was calibrated. As is typical of a statistical hypothesis test, the validation does not confirm with statistical confidence that the identifications met the expected performance. Instead, it evaluates whether there is statistical evidence that the population identifications did not meet the performance specified by the model.

The performance validation is calculated using a model calibration measure. Calibration measures how close the predicted probabilities are to the observed rates for any configuration of the model's independent variables (D'Agostino et al., 1998; Harrell et al., 1996). Perfect calibration results when the predicted number of correctly identified populations aligns with the model's predicted outcomes.

Hosmer Lemeshow Test

Measures of calibration for binomial outcomes are often statistics which partition the data into groups and check how the average of the predicted probabilities compares with the observed proportion of success in each group.

Hosmer and Lemeshow (1980) produced a widely used statistic to test a given model's ability to fit a dataset. Let the sample size be N. The most common version of this test arranges the subjects

https://doi.org/10.52202/078572-0047

according to ascending predicted probabilities. It divides them into Q groups of the same size so that the first group contains the N/Q subjects with the smallest estimated event probabilities; the second group includes the following N/Q subjects with the next smallest estimated event probabilities. The Q-th group contains the N/Q subjects having the largest estimated event probabilities.

The grouping is usually out of Q=10 deciles, but any other choice of several groups is possible. Given the groups, the Hosmer-Lemeshow test compares the observed number of positive outcomes (prevalence or observed frequency) with the mean of the predicted probabilities (expected frequency) in each group. The more the groups' observed frequencies are close to the corresponding expected frequencies, the better the model calibration. The goodness of fit is quantified using the Hosmer-Lemeshow formula:

$$H = \sum_{j=1}^{Q} \frac{\left(O_j - n_j P_j\right)^2}{n_j P_j \left(1 - P_j\right)}$$

Where:

H = Hosmer-Lemeshow test statistic
Q = groups of data
n_j = number of observations in the jth group
O_j = number of positive outcomes (correct calls) for the jth group
P_j = average probabilities predicted by the model for the jth group

Under the null hypothesis that the regression model is correct, the statistic H has approximately an asymptotic chi-squared distribution with $Q - d$ degrees of freedom (Giancristofaro and Salmaso, 2003).

$$p = 1 - F_{\chi^2}(H, Q - d)$$

Where:

p = p-value from Hosmer-Lemeshow test statistic
H = Hosmer-Lemeshow test statistic
F_{χ^2} = cumulative distribution function of chi-squared distribution
$Q - d$ = degrees of freedom
d = number of model parameters (wall thickness: $d = 1$, yield strength: $d = 2$)

Consequently, in the Hosmer-Lemeshow goodness of fit test, an observed chi-squared value less than the critical value of the chi-squared distribution with Q-2 degrees of freedom at the 0.05 significance level indicates a good fit of the model. Because of the attributes used in this project, many variables will result in expected probabilities close to 1.0, which results in large H values. To address this, an adjusted Hosmer-Lemeshow test is proposed that adjusts for small sample sizes (Giancristofaro and Salmaso, 2003):

$$H^* = \sum_{j=1}^{Q} \frac{\left(O_j - n_j P_j\right)^2}{n_j \left(P_j + \frac{1}{n_j}\right)\left(1 - P_j + \frac{1}{n_j}\right)}$$

https://doi.org/10.52202/078572-0047

Where:

H^* = Modified Hosmer-Lemeshow test statistic

The modified Hosmer-Lemeshow statistic can be computed for new verified datasets to determine whether the data follows the model. More specifically, the test determines the goodness of fit of the latest data to the performance model (from the previous section). The model does not characterize the data if the test's null hypothesis is rejected.

One important consideration when calculating the Hosmer-Lemeshow test is that the model may be rejected if the verified populations performed better than the model; the test does not evaluate if the new observations are performing better or worse, just whether the model is a good fit. When the model is rejected, the data should be reviewed to determine if this is due to better-than-expected performance.

The model validation procedure outlined in this section is similar to the API 1163 Level 2 performance validation procedure, and the model fitting procedures in Section 2 are similar to the API 1163 Level 3 assessment.

Conclusion

49 CFR §192.607 (PHMSA 2019) defines the data requirements for pipeline material properties and attributes. Recent ILI technologies have shown promise in identifying pipe populations and are likely to become a central tool in material verification programs. The framework's key objective is to validate ILI tools' performance to define pipe segment populations based on the 49 CFR §192.607 criteria using statistical tools similar to the performance validation steps for ILI measurement performance defined in API 1163.

ILI vendors determine individual pipeline populations based on ILI measurements and follows a hierarchical procedure to subdivide populations based on measured pipeline attributes. The performance of ILI population identification based on wall thickness, seam identification, and yield strength was assessed, and models were developed for each attribute of varying complexity.

Identifying pipe populations is unique compared to other measurement performance methods in the pipeline industry, as it relies more so on Type B uncertainty through subject matter expert judgment. Certain patterns in ILI measurement data are difficult to discern solely through systematic statistical tests or rule-based algorithms, necessitating data interpretation from the ILI vendor. In this project, the bimodality coefficient is proposed as a heuristic to estimate the probability of sub-populations within a sample of pipe segments with yield strength measurements. Mixture models and BIC are also considered as potential heuristic candidates, albeit with increased computational complexity.

https://doi.org/10.52202/078572-0047

The models fit for each of the measured pipeline attributes (wall thickness, seam detection, yield strength) can be used to estimate the performance of future population identifications and to determine which identified populations are most likely to be incorrectly called. In practice, the models are used in an evaluation spreadsheet tool to help an operator prioritize the review of ILI identified populations and assist in identifying false positives and false negatives.

The modified Hosmer-Lemeshow test is proposed as a method to validate future runs and determine if performance aligns with models fit to previous ILI population classifications, which are adopted as the current ILI performance specification. This test may also be applicable to other ILI performance evaluations.

References

Abadi M, et al. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from https://www.tensorflow.org/probability/examples/Eight_Schools

American Petroleum Institute (API). 2021. In-line Inspection Systems Qualification Standard, API 1163, Third Edition.

American Petroleum Institute (API). 2022. Recommended Practice for Assessment and Management of Cracking in Pipelines, API 1176, First Edition.

Canadian Standards Association (CSA). 2023. Oil and Gas Pipeline Systems, CSA Z662. Toronto, ON.

Giancristofaro, Rosa & Salmaso, Luigi. 2003. Model performance analysis and model validation in logistic regression. Statistica. 2. 10.6092/issn.1973-2201/358.

Joint Committee for Guides in Metrology (JCGM). 2008. Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM). 100:2008.

McElreath, Richard. 2020. Statistical Rethinking: A Bayesian Course with Examples in R and STAN.

Tarbă, Nicolae, Mihai-Lucian Voncilă, and Costin-Anton Boiangiu. 2022. On Generalizing Sarle's Bimodality Coefficient as a Path towards a Newly Composite Bimodality Coefficient. Mathematics 10, no. 7: 1042. https://doi.org/10.3390/math10071042

Title 49, Code of Federal Regulations. 49 CFR § 192.607. https://www.ecfr.gov/current/title 49/section-192.607. Accessed 2024.

Vincent, Benjamin. Accessed 2024. Binomial regression. In: PyMC Examples. Ed. by PyMC Team. DOI: 10.5281/zenodo.5654871.