MACHINE LEARNING FOR HEALTH MONITORING OF UNINSPECTED PIPELINES

Michael Smith	Principal Data Scientist	ROSEN Group
Matthew Capewell	Senior Data Scientist	ROSEN Group
Andy Linsley	Data Scientist	ROSEN Group
Jonathan Martin	Senior Data Scientist	ROSEN Group

Abstract

For decades, the pipeline industry has been collecting in line inspection (ILI) data for pipelines all around the world, in addition to an abundance of historical data on design, construction, operations and the environment. The availability of these datasets has led the industry naturally towards machine learning as a supporting technique for various integrity management activities. By observing trends from the past, we can better understand our current and future assets.

One particularly promising application of machine learning is condition prediction in pipelines that cannot be inspected using ILI. This is the case for approximately 40% of the world's pipelines.

Previous work by the authors has demonstrated this concept using historical data from approximately 80,000 pipe joints in a gas distribution network in North America. The study resulted in a set of machine learning models for external corrosion prediction in the target network, with outputs of a sufficient accuracy and resolution to support decisions on repair, mitigation and pipeline modifications.

This paper expands upon the previous work with a much larger dataset of almost 3.5 million pipe joints from multiple operators and countries in Europe. This represents a significant step towards the ultimate goal of "Virtual ILI": a health monitoring solution that can be deployed to almost any pipeline system in the world.

Introduction

Increasing volumes of in-line inspection (ILI) information have led the industry towards *supervised machine learning* as a complementary health monitoring solution for pipelines that cannot be inspected, or where ILI is otherwise contraindicated due to economic or operational constraints. If proven to be accurate and reliable, such a technique – a so-called *Virtual ILI* – could become an invaluable addition to thousands of pipeline operators' integrity management programs.

The fundamental idea is to learn trends observed within pipelines that have already inspected, and formalise these trends within a set of predictive models that can be applied to uninspected pipelines (Figure 1).

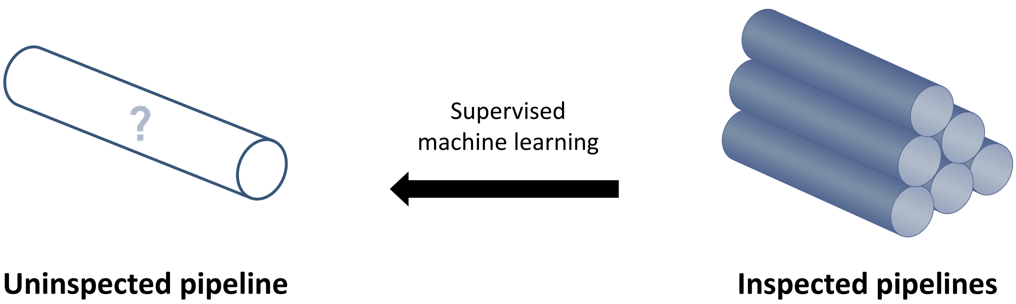

Figure 1: Principle of "Virtual ILI"

This can be applied for any pipeline threat that can be detected reliably using ILI, but the present study is focused on external corrosion. A Virtual ILI model for external corrosion may learn simple trends – for example that older pipelines tend to have a greater density of corrosion anomalies – or more intricate ones – such as the complex interactions between cathodic protection, coating, and the local environment.

Previous work by the authors [1] has demonstrated this concept using historical data from approximately 80,000 pipe joints in a gas distribution network in North America. The study resulted in a set of machine learning models for external corrosion prediction in the target network, with outputs of a sufficient accuracy and resolution to support decisions on repair, mitigation and pipeline modifications.

This paper expands upon the previous work with a much larger dataset of almost 3.5 million pipe joints from multiple operators and countries in Europe. Successful machine learning with a much more diverse, inhomogeneous dataset such as this would represent a significant step towards the ultimate goal of a health monitoring solution that can be deployed to almost any pipeline system in the world.

Data Preparation

The data for the study represented 755 onshore pipelines in Europe, inspected for metal loss between 2010 and 2020. A summary of the dataset is provided in Table 1. In addition to the inspection results (i.e. anomaly records) for these pipelines, the dataset included design and construction information at a pipe joint resolution, such as construction year and coating type.

Table 1: Summary of dataset

Number of pipeline sections	755
Number of pipe joints	3,443,896
Inspection date range	2010–2020
Number of external corrosion anomalies	1,157,386

A number of *condition metrics* were calculated from the available ILI results. Condition metrics are aggregated values describing some aspect of the external corrosion condition within a section of pipeline. In the present case, the metrics were calculated at pipe joint resolution, i.e. one value per joint. Three condition metrics were selected: *anomaly density* (anomalies m^{-2}), *relative corroded area* (%) and *maximum depth* (mm).

The dataset was further enriched using a number of open-source geospatial datasets from Europe, as described in Table 2. These provided additional environmental variables for the models. Example maps are shown in Figure 2.

Table 2: Geospatial datasets

Variable category	Dataset source
Crossings (roads, railways etc.)	Geofabrik – OpenStreetMap
Terrain	Copernicus – EU-DEM
Land use	Copernicus – CORINE Land Cover
Soil properties	Esri – World Soils Harmonized Soil Database
Precipitation	Deutscher Wetterdienst – Open Data
Socioeconomics (e.g. income group)	Natural Earth

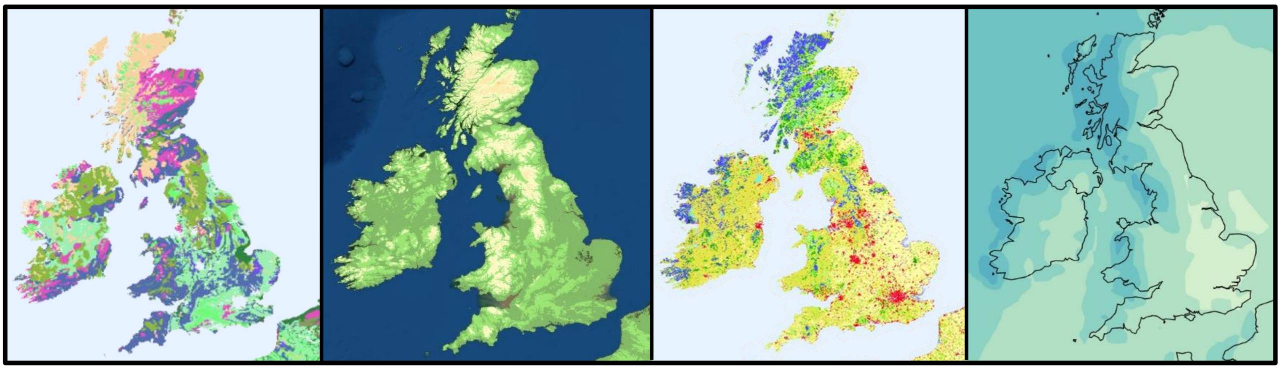

Figure 2: Geospatial dataset examples – from left to right, soil type, elevation, land use and precipitation

The data were stored in a Microsoft® SQL Server database, which could be queried in such a way as to provide a list of pipe joints with associated properties. The database was created using a series of Extract Transform Load (ETL) operations authored programmatically in Python™. Geospatial datasets required further processing using the Esri software ArcGIS™, in combination with the FME® data integration platform.

Figure 3 illustrates the structure of the final dataset.

Pipe joint properties

3,443,896 pipe joints

Joint ID	Construction year	Coating type	Precipitation (mm)
1	1982	Tape	790
2	1982	Tape	790
3,443,896	1965	Coal tar	822

Condition metrics

Anomaly density (m^{-2})	Corroded area (%)	Maximum depth (mm)
n/a	n/a	n/a
1.1	0.28	1.7
n/a	n/a	n/a

Figure 3: Illustration of final dataset

Distributions of the condition metrics are shown in Figure 4. Note that anomaly density and corroded area are presented on a logarithmic scale for clarity.

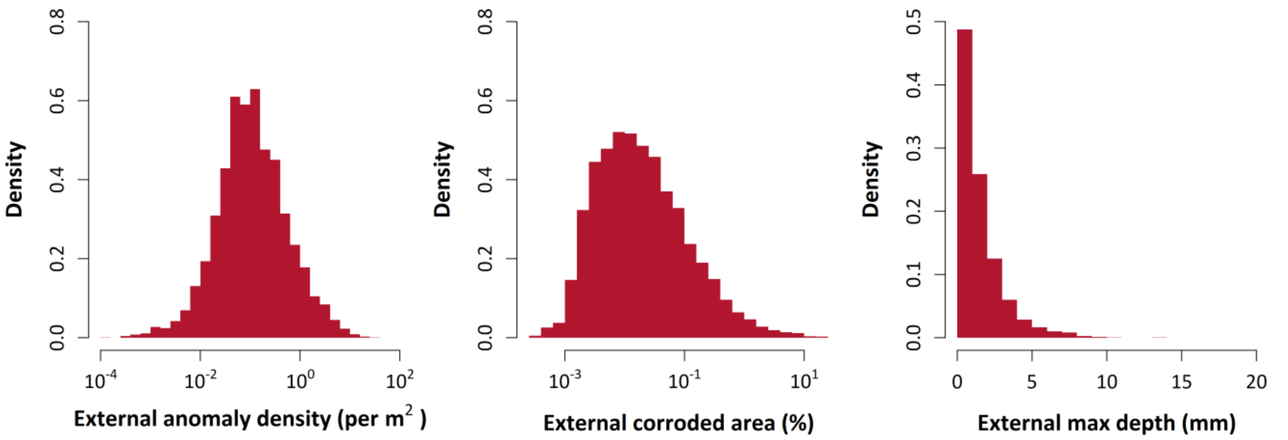

Figure 4: Distribution of condition metrics

Machine Learning

The purpose of machine learning is to embed trends within a predictive model that can be applied to new, unseen cases. In general this is achieved by defining a function, f, that maps a set of predictor variables, $\{x_i\}$, to a target variable, y, i.e.

$$y = f(x_1, x_2 \ldots x_n)$$

More detailed discussion on the key concepts of supervised machine learning is provided in previous work by the authors [1, 2].

In the present case, the predictor variables were properties of the pipe joints and environment (such as construction year, coating type, precipitation and elevation), while the target variables were the condition metrics (anomaly density, corroded area and maximum depth). The target was a set of three different models that could predict each of the condition metrics at a pipe joint level of resolution (Figure 5).

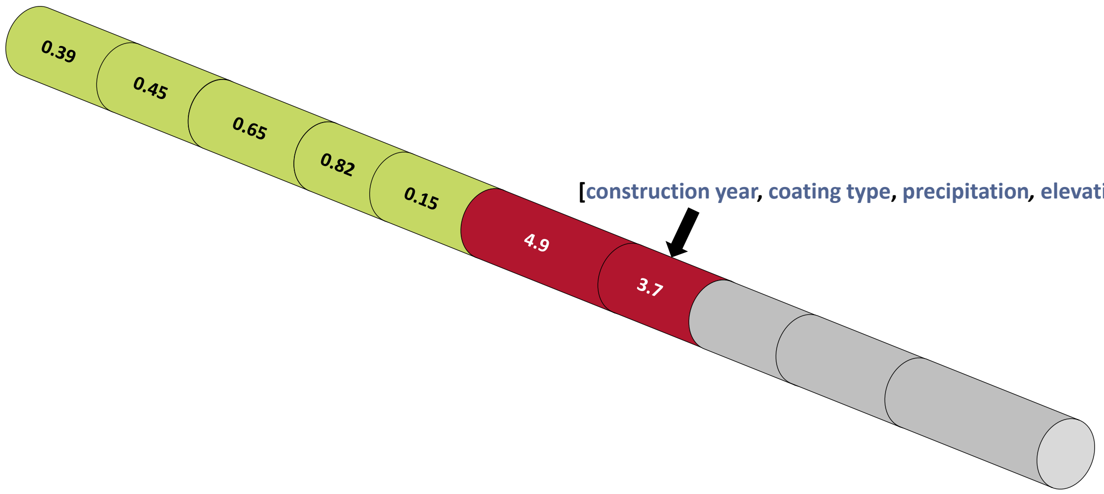

Figure 5: Prediction of maximum corrosion depth (mm) at pipe joint level

The selected model type was a *deep neural network*. A neural network is a directed graph constructed of nodes and arcs – as illustrated in Figure 6 – that describes a non-linear function between a set of inputs and one or more outputs. The function is governed by a set of network parameters (weights and biases) that are incrementally adjusted as the network "learns" from new data.

A *deep* neural network is a specific type of neural network with more than one *hidden layer* (i.e. multiple layers between the input and output layers). This allows for modelling of very complex and highly non-linear functions.

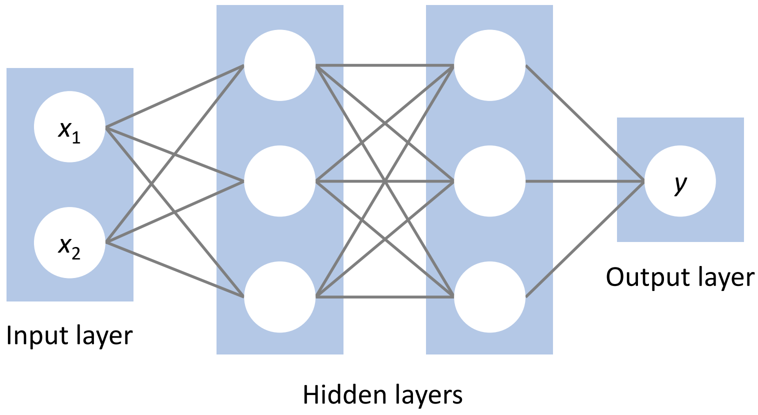

Figure 6: Deep neural network

Each of the condition metrics (anomaly density, corroded area and maximum depth) was predicted with its own deep neural network. Prior to training, however, these target variables were engineered in order to maximize performance. This involved filtering and clustering of the dataset. Further details on these processes are provided in [1].

Development and training of the deep neural networks was achieved using the R library H_2O [3], a framework for parallelizing machine learning and deep learning algorithms. The models were trained on a random 80% of the dataset and tested on the remaining 20%.

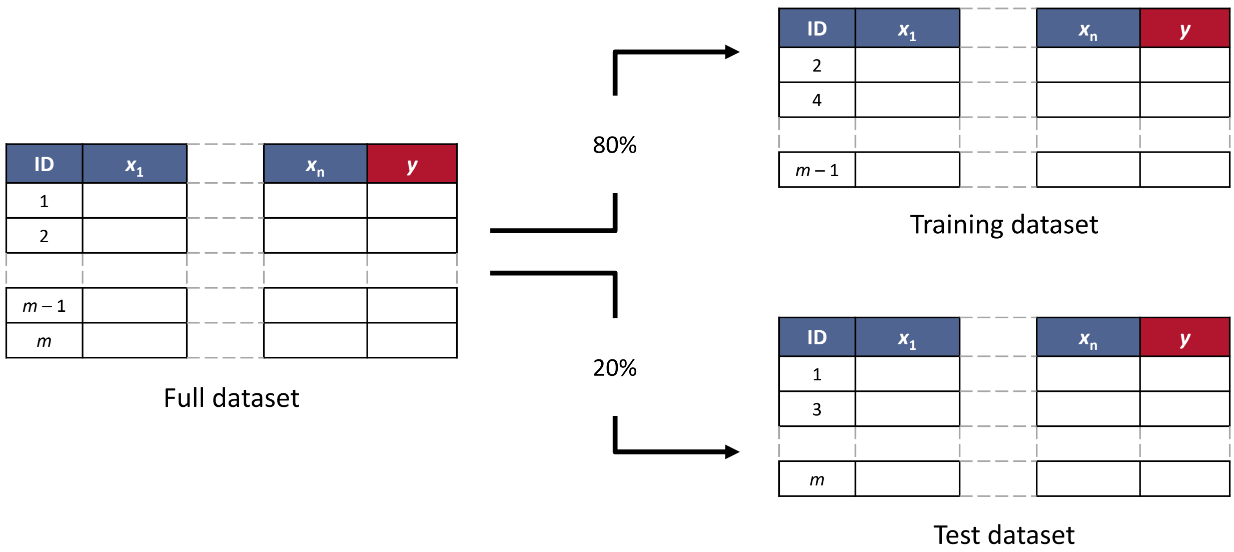

Figure 7: Train test split

Unity plots for the test dataset are shown in Figures 8-10 and the performance of the models (RMSE) is recorded in Table 3. Note that the models for anomaly density and corroded area were designed to predict the base 10 logarithm of the metric (rather than the metric itself). This was necessary due to the heavy positive skew in the distributions (see Figure 4). The result is that the RMSE values reflect orders of magnitude rather than absolute quantities.

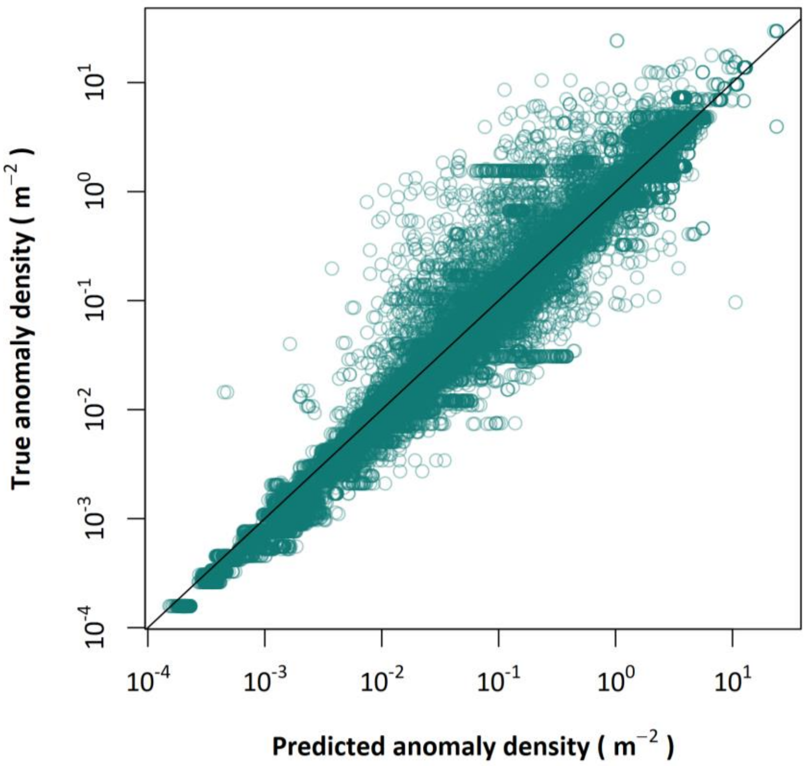

Figure 8: Unity plot for anomaly density model

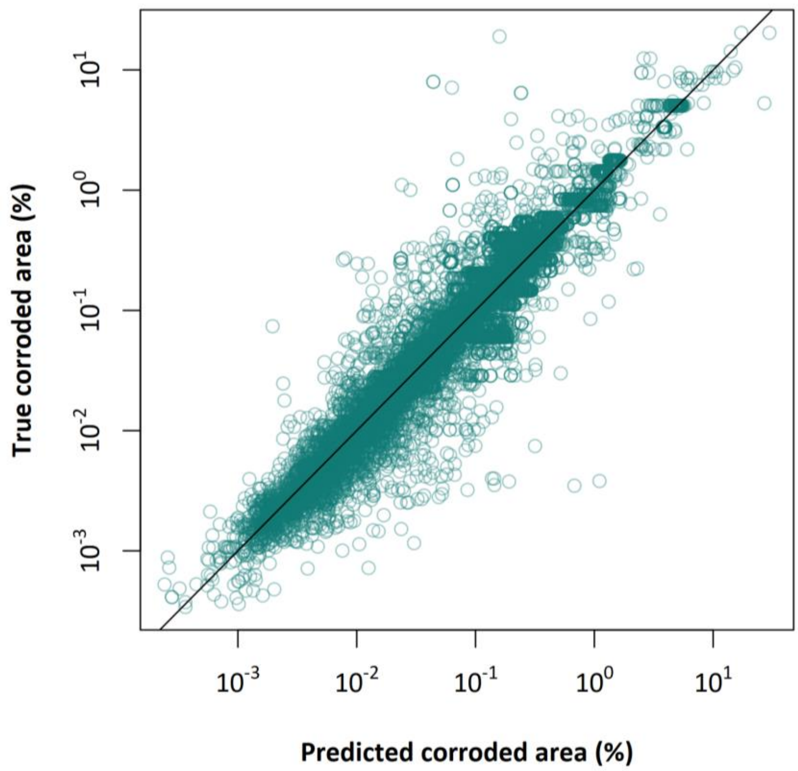

Figure 9: Unity plot for corroded area model

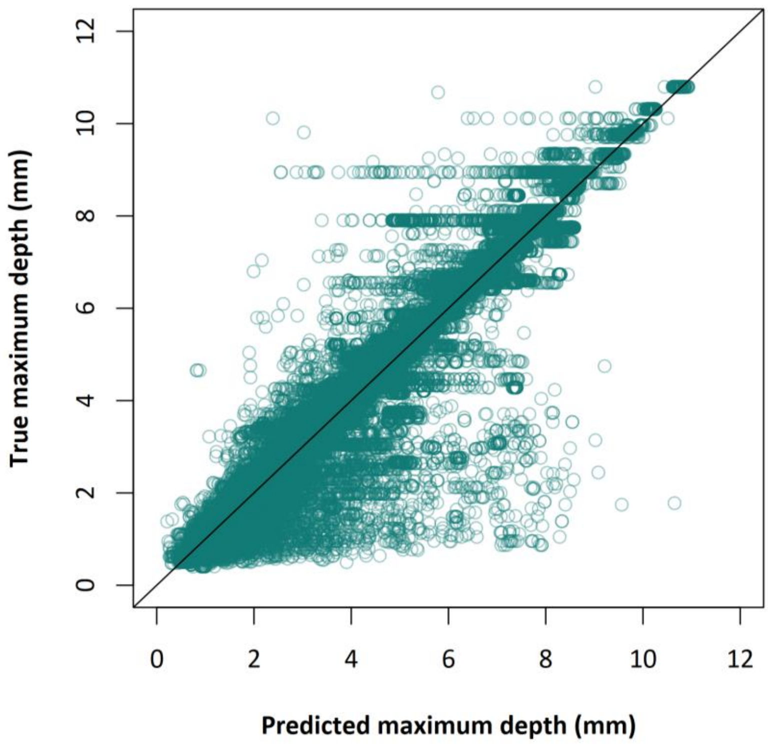

Figure 10: Unity plot for maximum depth model

Table 3: Model performances

Condition metric	RMSE for test dataset
$\log_{10}$(anomaly density)	0.10
$\log_{10}$(corroded area)	0.12
Maximum depth	0.35 mm

Discussion

The performance of the deep neural network models on the unseen test dataset is extremely promising, with almost 100% of anomaly density and corroded area predictions lying within one order of magnitude of the true values, and ~98% of maximum depth predictions lying within ±1 mm. This slightly exceeds the performance of the models developed as part of the previous study, suggesting that the relative inhomogeneity of the European dataset is compensated by its size. The results are expected to be favourable in comparison to the conventional modelling procedures used as part of External Corrosion Direct Assessment (ECDA) [4], and hence highly valuable for integrity management decision support.

The presence of outliers should not be overlooked, however. In the case of corrosion prediction these are most likely caused by unique corrosion processes that the model has never seen, or by "hidden" variables that are not currently captured in the models. This serves to highlight the importance of continuous improvement. Virtual ILI models must continuously improve as new ILI datasets become available, and new predictor variables are collected. Over time, the anomalous predictions will be brought into line and performance will improve.

Further insight can be gained by considering *variable importance* values. These reflect the extent to which each predictor variable influences the prediction. In the previous study, the Gedeon method [5] for variable importance identified several key variables that were influential on the prediction of external corrosion condition metrics, including:

- CP potential
- Elevation
- Mean annual precipitation
- Construction year
- Drainage
- Rail intersection
- External coating
- Road intersection

The importance of these variables can re-evaluated for the current set of models. Figure 11 shows a comparison of variable importance between the previous models (single operator/network in North America) and the current set of models (multiple operators/networks in Europe).

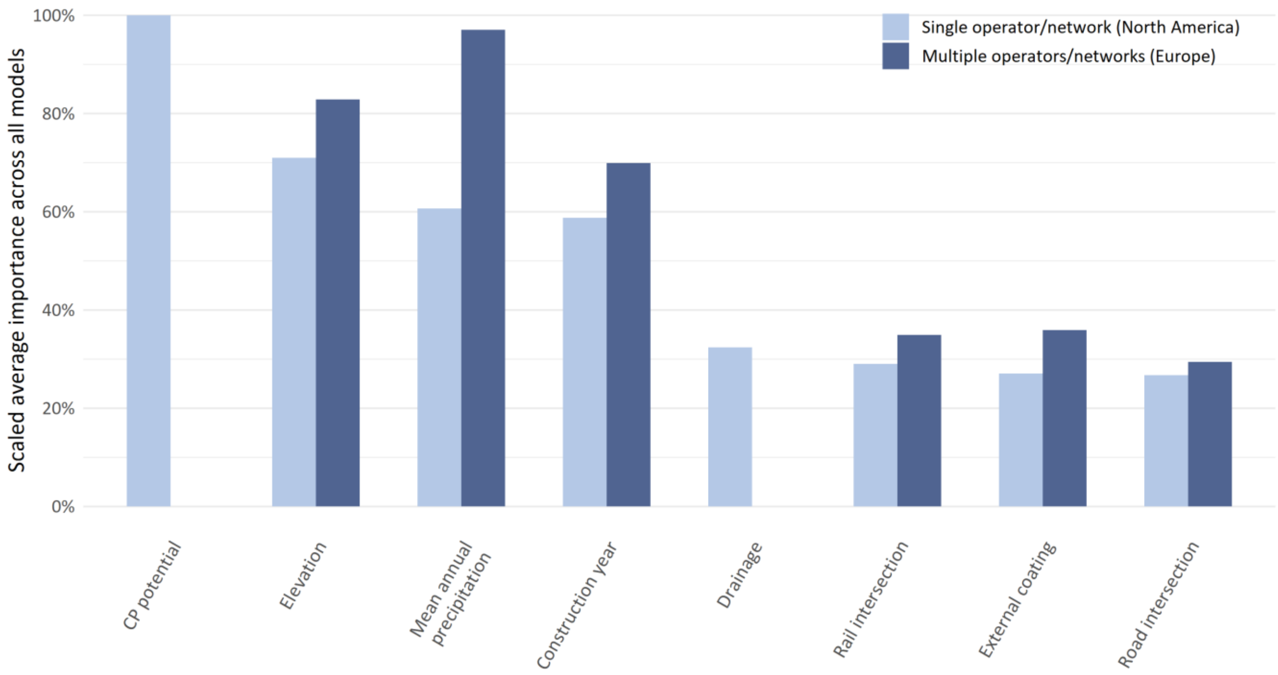

Figure 11: Comparison of variable importance between previous and current models

Roughly speaking, those variables that were previously identified as important were also identified as such in the present study, with the exception of *CP potential* and *drainage* which were not included in the new models. The ranking of these variables was also similar. Not only does this give confidence that the new models are behaving in a physically meaningful way, but it also confirms that external corrosion (at least in onshore pipelines) has universal and generalizable causes.

Given that CP potential was identified as the most important variable in the previous study, it is natural to assume that the model performances would be reduced by its absence. The same is true of drainage, although this is a less influential variable. Such a reduction in performance is not observed, however, and Figure 11 is useful in understanding why this may be the case.

From the plot it is clear that the absence of two key variables has systematically increased the variable importance values for the remaining variables. This suggests that the new models have placed a greater reliance on these variables in order to extract useful information. This is particularly true of *mean annual precipitation*, which is the most important variable in the new models, and almost 40 percentage points higher in importance than it was previously.

The most likely explanation is that *precipitation* (rainfall) is encoding information about the performance of the CP system. As identified in the previous study – and reproduced in Figure 12 – higher values of rainfall tend to correlate with more electronegative CP potentials. This results in the somewhat counterintuitive trend of higher rainfall leading to less corrosion.

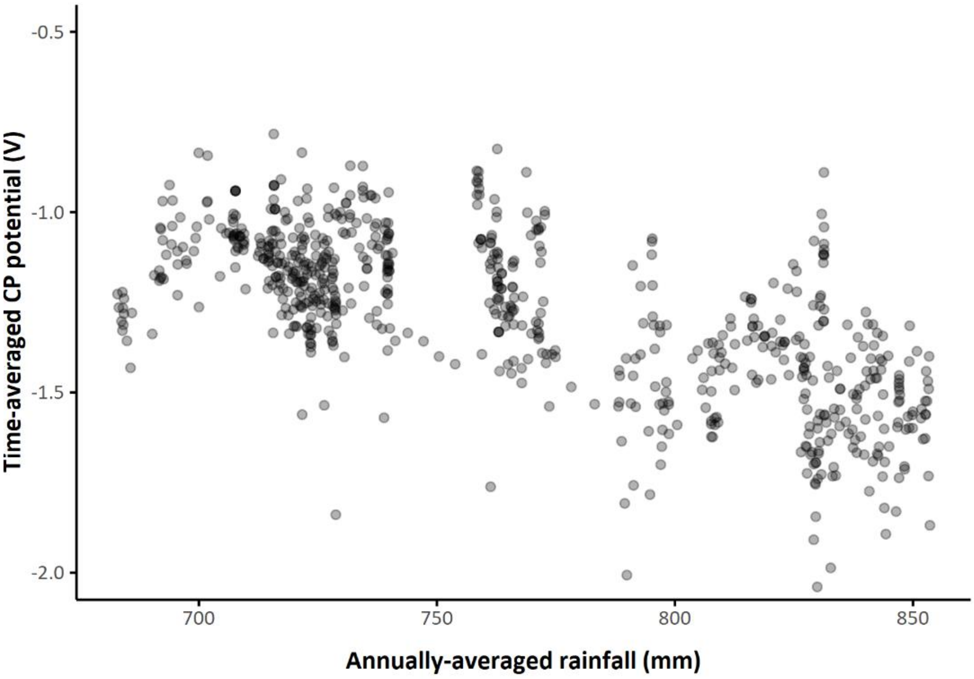

Figure 12: CP potential vs. mean annual precipitation

This finding reflects one of the great advantages of machine learning methods, namely their ability to compensate for the absence of direct, causal trends and instead use indirect correlations to make predictions. While it is good engineering practice to use causal variables, an insistence on this approach can sacrifice predictive performance. It is the experience of the authors that oversight from a competent statistician and a subject matter expert is sufficient to ensure that correlations are meaningful.

Conclusion

This paper described the production of three machine learning models applied to external corrosion trained on a large, diverse population of in-line inspection (ILI) results from 755 pipelines and almost 3.5 million pipe joints. External corrosion was measured using *anomaly density*, *corroded area* and *maximum depth*, resulting in three (deep learning) models. Comparisons were made with previous work that used a smaller, but homogeneous dataset in order to determine whether the main predictors of external corrosion were consistent.

The models' performances at pipe joint resolution were quantified using an unseen test dataset. Root Mean Squared Error (RMSE) values of 0.10, 0.12, and 0.35 mm were achieved for anomaly density (log_{10} transformed), corroded area (log_{10} transformed) and maximum depth, respectively. In addition, almost all anomaly density and corroded are values were predicted within one order of magnitude of their true value, while ~98% of maximum depths were predicted within ± 1 mm of their true value. These metrics indicate a solid performance overall.

Variable importance measures were similar when compared with the previous models, suggesting that generalizable trends have been established. Despite missing some key variables seen in previous models (e.g. cathodic protection potential), there was no corresponding drop in performance. This is likely due to the much larger sample size, and the inherent ability of machine learning models to extract information from proxy variables – for example, the use of rainfall measurements to infer the efficacy of a cathodic protection system.

Further data collection efforts are expected to reduce outliers and improve overall performance, bringing the "Virtual ILI" concept – a health monitoring solution that can be deployed to almost any pipeline system in the world – closer to reality.

References

[1] Smith, M S; Capewell, M; Clyne, A and De Leon, C (2021). *Deep Learning for High-Resolution External Corrosion Prediction in Uninspected Pipelines*. Pipeline Pigging & Integrity Management (PPIM) Conference, February 2021, Houston, Texas, United States of America

[2] De Leon, C and Smith, M S (2020). *Machine Learning to Support Risk and Integrity Management*, Pipeline Pigging & Integrity Management Conference. February 17–21, 2020, Houston, TX, United States of America

[3] H2O.ai, "Deep Learning (Neural Networks)," 17 November 2020. [Online]. Available: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/deep-learning.html. [Accessed 29 September 2021]

[4] NACE International (2010). *Standard Practice, Pipeline External Corrosion Direct Assessment Methodology*. ANSI/NACE SP0502 2010, Item No. 21097.

[5] Gedeon, T (1997). *Data mining of inputs: Analysing magnitude and functional measures*, International Journal of Neural Systems 24(2), 123-140