Big Data Analytics for Rapid Assessment of Pipeline Condition

By M. CAPEWELL, K. PESINIS and M. SMITH*

* Matthew Capewell, Konstantinos Pesinis and Michael Smith, ROSEN Group. E-Mail: arichards@rosen-group.com

0179-3187/19/06 DOI 10.19225/190607

Abstract

Inline inspection (ILI) has served the midstream oil and gas industry extremely well for decades. Using traditional techniques such as defect assessment and corrosion growth assessment, ILI data can be used to inform corrective actions and generate robust integrity management plans for future pipeline operation.

Given the age of the industry, a wealth of ILI data has now been accumulated, and at the same time the digital technologies used to manage data have become increasingly sophisticated and accessible. We can now collect, store, structure, and extract value from "big data" with relative ease.

In this article the value of a large database of ILI data and asset metadata for pipelines from all across the world will be demonstrated. As an example, it is shown how such a database can be used to classify corroded pipelines based on their condition, without completing a detailed assessment. Simple condition metrics can then be related back to the properties of the asset – for example, its age or coating type. These simple descriptive analytics techniques serve as a prelude to more rigorous classification techniques using machine learning.

Introduction

It can be thanked to Moore's law [1] for the recent (but inevitable) rise in interest in "big data" and machine learning methods. As time has progressed, our computers have become more and more powerful in an exponential fashion. Programs that used to require a computer the size of a room can now be run on a device no more than a few inches in size – such as the Raspberry Pi [2] – and in a fraction of the time.

Enabled by advances in digital technology, many businesses now collect huge amounts of data, often for the purpose of promoting or developing their products and services. The appetite for data has grown to such an extent that a multitude of commercial solutions now exist to tackle typical "big data" problems (see e.g. cluster computing, parallel computing, cloud computing [3, 4]). Additionally, many of the tools to perform data analytics are open source and have become more accessible than ever (see Python's scikit-learn module [5] or one of the forty-six packages that R offers [6]).

The oil and gas industry is not overly concerned with promoting products, but instead uses data to establish and maintain asset performance. To this end, businesses capture much of their data using measurement technology. The pipeline industry is an excellent example, generating enormous volumes of in line inspection (ILI), above ground survey and process monitoring data every day. Pipeline operators are primarily interested in their own datasets (using them to understand the condition of their assets) but there is much to learn by assessing the industry's data as a whole. With simple descriptive analytics techniques, the global dataset can be analysed comparatively in order to catalogue and benchmark the world's pipelines in terms of their condition. It can be even imagined to have a fully automatic, empirical procedure, where new data points are added to the population in real time.

This is not far from being a reality. The ILI data warehouse already holds hundreds of millions of anomaly records (e.g. corrosion and crack features) for tens of thousands of pipelines across the world. Many files on individual anomalies are upwards of 30 dimensional (with attributes such as depth, length, width, classification, inspection date etc.), resulting in billions of individual tabular entries. Pairing this data with machine learning methods, it can then be categorised pipelines based on pre-determined single-valued numerical condition metrics. Some examples of valid and relevant machine learning methods in this application are:

- *Principal Component Analysis* (PCA) to reduce the high dimensionality of the data and visualise the variance,
- *Perceptron/logistic regression* to determine predictors of category allocation,
- *Density-Based Spatial Clustering* of Applications with Noise (DBSCAN) to cluster the n-dimensional data,

with scope for many additional methods such as large margin classifiers or neural networks. Such methods are complex and often computationally expensive on large datasets (varying amongst algorithms) with large amounts of time and effort spent on prototyping models and optimising parameters. Thus, it has been decided for an introductory insight by grouping points based on population performance alone. Similar to a school class where performance in a subject is measured relatively, this ethos has been adopted towards measuring the condition of pipelines and use empirical methods only. A clear extension would be to go the extra mile and utilise both supervised and unsupervised machine learning methods, thus aiming to improve categorisation and uncover non-intuitive hidden patterns in the data. Note that applying these methods must be done with care, since the problem needs to be formulated in the correct way to be compatible with a machine learning algorithm.

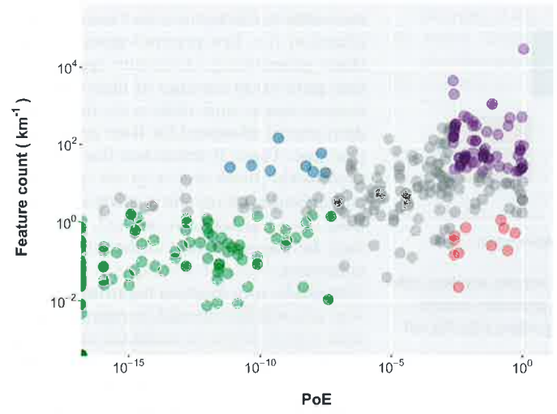

Fig. 1 Scatter plot of condition metrics with categories coloured (grey indicates insufficient confidence to categorise)

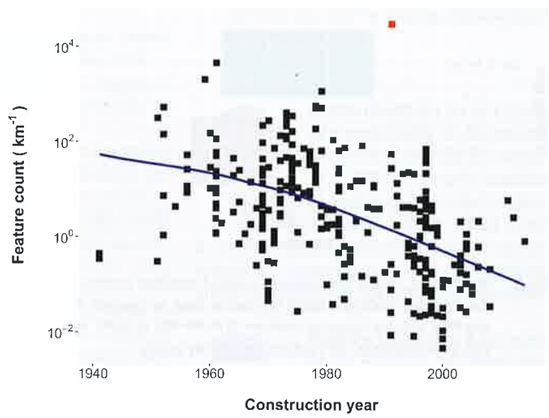

Fig. 2 Scatter plot of construction against external feature count, with fitted LOESS smoother (blue) and highlighted outlier (red)

Tab. 1 Category boundaries. Here, P75 denotes the 75th percentile, and P50 denotes the 50th percentile i.e. the median

Category 1	Category 2	Category 3	Category 4
PoE < P50 Feat. count < P50	PoE < P50 Feat. count > P75	PoE > P75 Feat. count < P50	PoE > P75 Feat. count > P75

Descriptive Analytics

Our efforts will be focussed on corrosion, given that this remains a major threat to pipelines and accounts for the majority of detected anomalies in the data warehouse. Specifically external corrosion has been explored, since the complexity of internal corrosion makes it a troublesome candidate for a purely empirical analysis. Analysis on internal corrosion is best when boosted by theoretical models, something that can be (and has been) done via Bayesian learning [7, 8]. The analysis and data visualisation is performed in the statistical computing software R [9], with the bulk of data collection having been performed in Python [10].

Condition Metrics

Quantifying a pipeline's condition numerically can be a challenge. However, it has been opt for two main metrics applicable to corroded pipelines [11]:

- *Feature Count* – the number of individual corrosion features detected per kilometre
- *Probability of Exceedance* (PoE) – a value quantifying the probability of any feature in a pipeline exceeding a critical depth threshold

Paired together these metrics can be used to categorise pipelines based on their relative condition within a population. The values can help to gauge whether particular subpopulations (e.g. pipelines with different coating types) are performing better or worse than others, therefore allowing to visualise the effect of "metadata" on the state of the asset. In a machine learning application, the training data can be labeled using these metrics, and then attempt predictions by using the metadata supplied with those training data.

Previous work by the authors [11] has described an entirely empirical approach to categorisation, using percentiles of the data in two dimensions. From this it has been continued by extending the number of dimensions to three and four to create a "condition cube", and focus on uncovering general trends in the associated metadata.

Figure 1 shows a 2D-plot for pipelines, with each point referring to a unique externally corroded asset. Categories are assigned such that category 4 (purple) is the worst, with categories 3 (red), 2 (blue) and 1 (green) following suit. The combination of relatively high PoE and feature count results in a bad condition pipeline, and thus category 4 is deemed the worst category. Category 3 pipelines have a relatively smaller feature count but with high PoE, meaning failure is more likely than for a category 2 pipeline. Category boundaries are illustrated

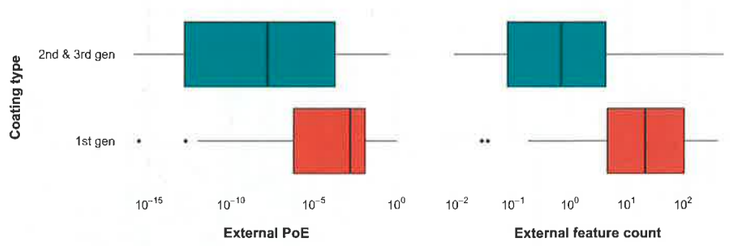

Fig. 3 Box plots of coating type measured against condition metrics. First generation includes any lines with the following coatings: asphalt tar, coal tar, tape, or concrete. Second and third generation includes any lines with the following coatings: 3LPE/PP, FBE or HDPE. Box plots show the values P25, P50 and P75 as vertical lines, with outliers denoted by points.

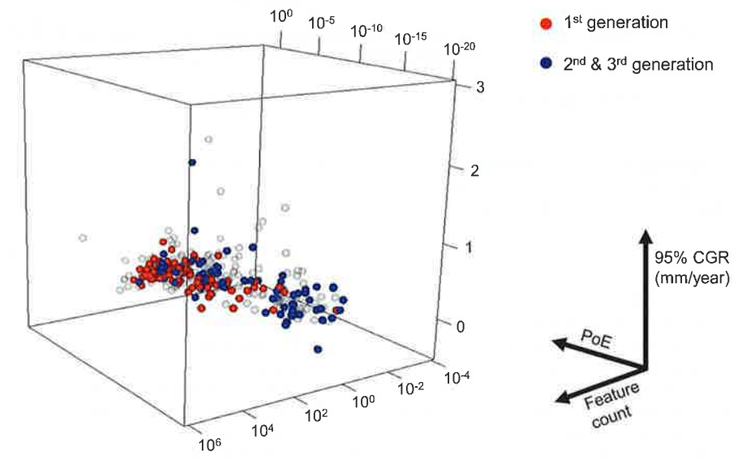

Fig. 4 Four-dimensional extension of the condition space, now with 95th percentile of CGR (mm/year) and coloured by coating

in Table 1, with P50 and P75 values being inferred from the axes in the figure.

Metadata

With external corrosion, asset "metadata" may be expected to contain valuable information in explaining the variability present in the dataset. Two pertinent variables are the age of a pipeline, and its coating type.

Figure 2 shows, for example, how the feature count metric varies with the age of a pipeline (i.e. the time period since it was constructed).

An intuitive trend can be seen in blue in Figure 2 obtained via LOESS [13], in which older pipelines appear to have a higher average external feature count. Explicitly, the trend value decreases from 24 km^{-1} in 1960 to 1.5 km^{-1} in 1990 – a notable reduction. Not all pipelines fit this trend, of course, and one particularly interesting outlier is highlighted in red. This pipeline has a feature count three orders of magnitude higher than average for a pipeline constructed in the 1990s. However, further investigation shows that this pipeline does not have any external coating, and is therefore much more prone to external corrosion than the general population. Age is thus a good predictor of condition, but it should never be considered marginally since this is a multivariate problem.

Figure 3 explores further the influence of coating type. First generation coatings such as coal tar enamel (CTE), asphalt, concrete, and tape clearly result in higher feature counts and PoE values than second and third generation coatings such as fusion bonded epoxy (FBE), and 3 layer polyethylene/polypropylene.

These are two examples of how these patterns can tease out of the data, showcasing the hidden relationships between variables that can eventually lead to improving the prediction of pipeline condition.

Predictive Analytics

In Figure 4, a direct extension of Figure 1 can be seen with two extra dimensions: the 95th percentile corrosion growth rate for that pipeline (also known as the characteristic growth rate), and coating classification (i.e. first generation vs. second/third generation). A cluster appears for first generation coatings at high values of feature count and PoE, with more random scatter observed for later generation coatings. Thus, if there has been a concern with predicting a given pipeline's condition based on the coating classification, first generation coated pipelines may be expected first to be in a worse condition.

Ultimately, any analysis regarding prediction of condition should incorporate multiple aspects of the metadata as so to build in the inter-correlation between variables. As demonstrated in the previous section, there are general trends present in both construction year and coating type, which we expect to interact with each other. From what we have seen, we may infer:

- As the age of a pipeline increases, it is more susceptible to corrosion
- As the coating type becomes more modern, it is somewhat less susceptible to corrosion than the legacy counterparts

However, is the latter effect due to a more protective coating, or is it because the pipeline is younger?

These natural correlations can make marginal inference challenging, and so these relationships must be built in to the models used for categorisation. A Bayesian approach is recommended when fitting a logistic regression model, since metadata are often missing or difficult to extract. A logistic regression model can be applied in this context of categorisation by fitting four separate regressors to obtain a probability of belonging to each category, for a given pipeline. Then, the maximum probability can be chosen to allocate a category. The variables used for prediction would be a mixture of qualitative and quantitative data, and so care must be taken to either convert the qualitative variables into binary variables, bin into factor variables, or use a different non-trivial technique.

With this set-up, a neural network can instead be employed by inputting as many variables as possible as nodes, resulting in a probability table for any given pipeline quantifying its probability of belonging to each category. Then, non-linear complex hypotheses can be learned rather than the linear divide proposed in Figure 1.

Conclusions

With large amounts of data, empirical methods prove to be a powerful tool to non-parametrically categorise the data and therefore classify pipeline by their condition. The absence of parameter tuning and model prototyping means inference is not only simple, but far quicker. However, large amounts of complete data are required, where specific entries are often missing.

To prevent this, missing metadata could be obtained by adopting a Bayesian framework, in which expert knowledge is combined with the data to model the missing values which can thus bolster our chances of determining relationships between metadata and pipeline condition.

To tackle the grey buffer zone, existing categorised points can be supplied as training data to supervised machine learning algorithms to colourise the grey points in a non-intuitive way. An alternative approach would be to use an unsupervised clustering algorithm to automatically determine the categories in the condition space, eliminating the need for labelled training data.

Using either of these methods, the value of data can be harnessed to improve the classification of pipeline condition and gain a valuable tool for benchmarking pipeline condition – thus proving useful for prioritising pipeline inspections and repairs.

References

[1] Moore, G (1998): Cramming More Components Onto Integrated Circuits. Proceedings of the IEEE, 86 (1), pp.82–85.

[2] Raspberrypi.org: https://www.raspberrypi.org/ [Accessed 13 May 2019].

[3] Spark.apache.org: Apache Spark™ - Unified Analytics Engine for Big Data. https://spark.apache.org/ [Accessed 13 May 2019].

[4] Apache Cloudstack: Apache Cloudstack. https://cloudstack.apache.org/ [Accessed 13 May 2019].

[5] Pedregosa et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12 (1), pp. 2825–2830.

[6] Cran.r-project.org: CRAN Packages By Name. https://cran.r-project.org/web/packages/available_packages_by_name.html [Accessed 13 May 2019].

[7] Smith, M S; Pesinis, K; Barton, L and Laing, I.: Intelligent Corrosion Prediction using Bayesian Networks. NACE CORROSION Conference and Expo 2019, March 24-28, 2019, Nashville, Tennessee.

[8] Smith, M. S.; Cronjaeger, S.; Ershad, N.; Nickle, R. and Peussner, M.: Pipeline data analytics – enhanced corrosion growth assessment through machine learning. Proceedings of the 2018 12th International Pipeline Conference, September 24–28, 2018, Calgary, Alberta, Canada.

[9] R Core Team: R a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[10] Python Software Foundation: Python Language Reference, (Version 2.7). URL https://www.python.org/.

[11] Palmer-Jones, R.; Smith, M. S.; Pesinis, K.; Santana, E; Capewell, M. L.: The good, the bad and the ugly – categorizing pipelines using big data techniques. Pipeline Pigging & Integrity Management (PPIM) Conference, February 18–22, 2019, Houston, Texas, United States of America

[12] Laney, D.: 3D data management: Controlling data volume, velocity and variety. META Group Research Note. 6 (70)

[13] Cleveland, W. and Devlin, S.: Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Journal of the American Statistical Association, 83 (403), p.596.

Michael Smith is a Senior Engineer with responsibility for the development of new products and services within ROSEN's integrity consultancy business in the United Kingdom. As a subject matter expert in post ILI corrosion growth assessment, Michael has contributed to numerous publications on corrosion growth rate estimation, optimization and prediction.

Matthew Capewell is a data scientist for ROSEN in the United Kingdom, having completed a Master's degree in Mathematics and Statistics at Newcastle University. Matthew's expertise includes Bayesian statistics, extreme value theory and algorithm design using the programming languages R and Python.

Konstantinos Pesinis is a data engineer for ROSEN in the United Kingdom. He holds a PhD in Engineering with applications in pipeline integrity management and is enthusiastic to be part of ROSEN's data analytics and machine learning projects. His research interests include big data and cloud computing.