

Enhancing Pipeline Maintenance Using Data Science

David Toschini
GRTgaz Direction Projets et Ingénierie
6 Raoul Nordling
92270 Bois Colombes
France

Markéta Lallementová
GRTgaz Direction du Système d'Information
9 avenue de l'Europe
92270 Bois Colombes
France

Anthony Dequick and Bernard Masson
GRTgaz Direction Technique
5 Rue Ferdinand de Lesseps
Compiègne 60203
France

Thomas Zamojski
DataStorm
5 avenue Henry Le Chatelier
91120 Palaiseau
France

ABSTRACT

We can now make improvements on the assessment of the external coating condition of underground pipelines: in fact, after excavation and examination, only a small proportion of detected coating anomalies contain also a metallic fault or worse corrosion.

GRTgaz, operator and owner of the longest high-pressure natural gas grid in Europe (approximately 32 000km (19,884 miles), has tested an analytical approach based on historical data to increase the success rate (from the understanding of the problem and the available data, to the elaboration of algorithms and the optimization of their parameter). Different data sources to extract new knowledge have been used, works have been prioritized and the excavations have been optimized in order to have fewer but more efficient interventions.

The development and validation of the predictive model is ensured by a multidisciplinary team consisting of data scientists, corrosion experts and project managers. Based on historical data from 2013 to 2017, the model was calibrated taking into account both the indications of greater probability of metal damage and those of lower probability. The output of this model gives an indication of likelihood of corrosion along the pipes which cannot be pigged.

The paper will cover the construction method (based on Random forest algorithms), the first results. We will see the enhancements of the model by injecting more data and modifying mathematical rule, and how it will be able to integrate a new decision support tool.

Key words: pipeline integrity management program, metal damage, corrosion, maintenance, machine learning, random forest, cathodic protection, indirect inspection

INTRODUCTION

GRTgaz is a major gas grid operator (TSO). It operates a grid of approximately 32 000km (19884 miles) which is one of the longest pipeline system in Europe.

The company has a legal obligation to inspect and maintain all of its pipelines with a 10 years periodicity. There are about 2 700km (1678 miles) (inspected per year which leads to 1300 excavations representing an investment of 40 million € (approximately \$ 45 million).

Two main methods are used for pipeline inspection: inline inspection and External Corrosion Direct Assessment (ECDA).^{1, 2, 3} While smart pig data gives very good and reliable information about the overall pipeline condition, ECDA (based on DCVG/ACVG testing method) can improve the process.

In the method proposed in this paper, the assessment of pipeline condition is done using an internally developed methodology called MEICA which is based on evaluation of different risk factors levels.⁴ A weighted sum of these risk factors is then calculated. The weights were determined by a group of pipeline integrity experts putting together their long experience. Today, the methodology is recognized and approved by the French administration.

When looking at past experience, the methodology detects pipeline coating anomalies very well but the proportion of excavations also containing a pipeline damage (corrosion, dents, gouges, etc.) is rather small. For this reason, the integrity expertise team decided to test an analytical approach based on data and machine learning algorithms in order to increase pipelined damage detection and be able to make a more relevant selection of defects to be repaired.

In this paper, we will describe the way a machine learning algorithm was built and how its performance was evaluated. We will then show how it was implemented and how the comparison between the traditional method and the new data is done in order to be able to assess its performance over several years.

BUILDING A MACHINE LEARNING ALGORITHM

With more than 40 years of experience in pipeline inspection and repair, a lot of data about pipelines and their repairs has been accumulated. With the boom of data science and artificial intelligence use in the pipeline industry, the question about the possibility to improve the performance of pipeline integrity assessment with these technologies has become quite obvious.

As stated above, the company uses DCVG/ACVG and associated expert based methodology for all unpiggable pipeline integrity assessment. This represents approximately 2/3 of the excavations done by GRTgaz per year.

While the coating defects are well identified by this method, the performance in the detection of pipeline damage is rather poor. This is a major concern from the safety, cost efficiency and client impact points of view; thus, the pipeline integrity expertise team decided to work together with data science and IT experts in order to build an algorithm that would improve the repair work selection by a better prediction of metal damage presence.

Data to Be Used

We can find many different data sources containing data related to:

- Pipeline characteristics: diameter, thickness, maximum pressure, operating pressure, coating, age;
- Network configuration and topology;
- Cathodic protection: potential measurements, difference in potential, current density and intensity, stray currents, alternating current;
- Special points: bridges or water crossing, casings, et al;
- Excavations results: nature of the defect (coating defect, dent, corrosion, etc.) and its dimensions, expertise results, acceptability;
- Inspection results: maximum continuous gradient, IR% which is the quantification of the decrease in potential due to suspected pipeline damage;
- Humid zones and groundwater presence.

Although this already represents quite a lot of data, it seemed useful to complete it with external information coming from open source data such as:

- Corine Land Cover⁵ for the nature of the environment soil,
- Route 500^{®6} for road and railway presence,
- RTE, French high voltage grid⁷, for electrical network presence.

It is important to note that not all of these features will be used in the end by the machine learning model. It is always useful to have as much data as possible and let the algorithms decide in the training phase about the most pertinent ones. We would also like to underline that the data must be of good quality otherwise the model will not learn the right relationship between the features.

Once all the data was collected, it needed to be compiled in order to construct a *training set* – i.e. a set of data used for the calibration of the machine learning algorithm. This was the longest and the most time-consuming work in the project. Many meetings and workshops between the experts, data scientists and IT experts were necessary in order to define the right way to combine and clean the data.

Pipeline Modelling and Data Homogenization

From a statistical point of view, it is almost impossible to be able to locate the presence of an anomaly or a defect with the precision of a GPS (1 – 2m) (1.1 – 2.2 yards). So, it is necessary to find a scale for which all features that would enter into the model would be observed.

Pipeline integrity experts determined it would be pertinent to detect a defect over a 50m (54.7 yards) interval. Thus, the entire gas grid was divided into portions of approximately 50m (54.7 yards) leading to almost 650 000 sections.

For each of these sections, the value of each available feature was either found in the corresponding database or calculated. As several values may be available for a feature within a section, a homogenization leading to one value per section was developed with the help of several expert rules:

- Precaution principle – if several values are available for one section the value that is considered as most critical is imputed,
- If criticality is not known or not applicable, “the majority takes it all” principle is applied for categorical variables and a mean value is calculated for continuous variables,

- Certain features are measured within very large intervals which are nevertheless representative of the overall condition of the portion of the pipeline, in this case intermediate values are interpolated.

Figure 1 shows an example of data homogenization.

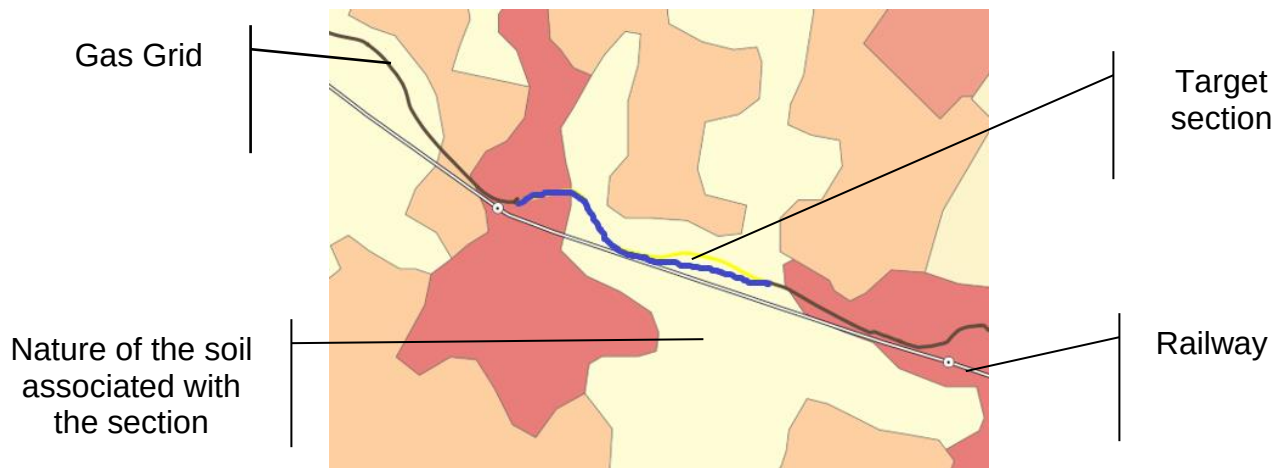


Figure 1: Example of data homogenization

Machine Learning Problem and Modelling

Once the training data set constructed, it is possible to proceed with the machine learning algorithm calibration. The selection of the appropriate algorithm depends on the nature of the problem and the data available.

Detection of corrosion or pipeline damage can be assessed with what in data science is called a **supervised classification task**. In this kind of problem, we use the data from the past that is labeled, i.e., for which the result is known, to calibrate a machine learning model, and we suppose that if we find similar data in the future, the label will also be similar.

For this type of problem, different families of algorithms can be applied and it is not possible to say in advance which one will perform the best. In our project, 3 different families were tested: logistic regression; gradient boosted trees; and random forests. It is the last family that gave the best results and that was selected for implementation. All the results presented will use this method.

Random forest algorithm

Random forest, introduced by Ho in 1995,⁸ is a set of decision trees. Decision trees⁹ are models where a decision is taken at each branch of the tree according to the values of different features and they proceed until the leaf of the tree is reached and where the tree indicates the class along with its probability.

Decision trees are generally constructed by separating the training data set into subsets according to the value of one of the input features. This process is repeated on each subset in a recursive way; it is a recursive partitioning.

The recursion is terminated if all data points in the subset have the same value of the target feature or if another partitioning does not improve the forecast. Random forest (Figure 2) is an aggregation of

individual decision trees where each tree is calibrated on a random subset of the training data. Each decision tree is used to predict a probability of the target feature and the overall prediction is then the mean value of all forecasts.

Random forests are considered as very robust and powerful algorithms and performs well on data sets on the size used (3000 – 6000 samples). They very often perform well in machine learning .

RANDOM FOREST

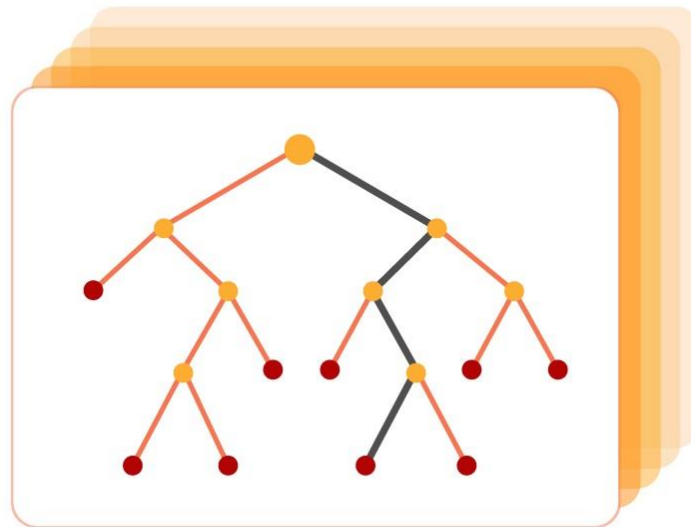


Figure 2: Random forest example

Assessment of the Machine Learning Algorithm's Performance

As explained in the previous part, the random forest gives a probability that there is corrosion or pipeline damage. But in reality, we want to make a binary decision – corrosion yes or no, mechanical damage yes or no. To be able to do that, a probability threshold needs to be set to classify a data point as positive. For instance, we might decide that there is corrosion as soon as the probability given by the model exceeds 20%.

Once the threshold is set, 3 different evaluation metrics were used to assess the performance of the model:

- Confusion matrix
- ROC curve
- AUC (area under the curve)

Confusion Matrix

Confusion matrix is a tool to measure the quality of a binary classification system. When comparing the forecasts to the reality, four different cases can occur. They are presented in Table 1 **Error! Reference source not found.**

Forecast\Reality	Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative