

Paper No.
16749



Mapping Indication Severity Using Bayesian Machine Learning from Inspection Data by Considering the Impact of Soil Corrosivity

Hui Wang, Sreelakshmi Sreeharan
University of Dayton
300 College Park,
Dayton, OH 45469
USA

Homero Castaneda
Texas A&M University
400 Bizzell St,
College Station, TX 77843
USA

ABSTRACT

This paper provides the industry with a data-driven approach that can characterize the heterogeneity of the soil corrosivity along the pipeline right-of-way by integrating data from soil survey, indirect inspection technologies, in-line inspection, and a slightly broader database of environmental (precipitation and vegetation) data. The outcome is the spatial distribution of different soil corrosivity regions, and the statistical description of the corresponding corrosion factors, corrosion defect depth, and number of defects corresponding to each corrosivity region, which can be used as a basis for excavation schedule and related risk assessment. This proposed approach considers the statistical similarity and spatial variation of soil corrosivity at different locations along the pipeline right-of-way. The proposed method is implemented and demonstrated using a database of a 110km pipeline section.

Key words: soil corrosivity; severity; unsupervised learning; clustering; Gaussian mixture model; pipeline

INTRODUCTION

Underground pipeline structures are considered as the most reliable and safest way of transporting petroleum products. These steel pipelines are subjected to many threats to integrity during their service life. Maintenance of underground pipeline systems has been an ongoing concern among industrialists. According to incident reports, corrosion is one of the leading causes of structural failures on pipeline systems.^{1,2,3} External corrosion can result in uniform and localized metal loss on the exterior surface of pipeline structures, resulting in reduction of the wall thickness of the pipe or other equipment. The loss of material from corrosion can eventually result in "pinhole" leakage, or a crack, split, or rupture of the pipeline unless the defects are repaired, the affected pipeline sections are replaced, or the operating

pressure of the pipeline structure is reduced. There exists in-line inspection (ILI) methods like Magnetic Flux Leakage (MFL),⁴ Ultrasonic wall thickness measurement and indirect inspection methods like Direct Current Voltage Gradient (DCVG), Close Interval Potential Survey (CIPS),^{5,6,7} alternating current voltage gradient (ACVG) used to detect and analyze stable as well as time dependent threats. Coating failure and surface conditions detected by these technologies may or may not pose a critical integrity threat. For example, several literature and case studies have revealed a high percentage of indications pinpointed by indirect tools resulted that the reported anomaly was electrochemically passive (with low to ignorable corrosion rate) then no repair and excavation was required.⁸ Instead, a more cost-effective solution should be implemented. Due to the lack of correlation between the indications with electrochemical and corrosion mechanism knowledge, even though some coating anomalies or surface conditions will not lead to significant mechanical damage, some companies, for conservative reason, will consider excavating all reported indications from indirect technologies. This practice not only leads to high costs, but more critically can lead to spending imbalanced with risk reduction. Currently, many pipeline companies follow NACE SP-0502 to make repair, remediation or maintenance decisions.⁹ While this is the standard practice, it does not reflect the risk-balanced decision strategy that would follow from the same outcomes supplemented by limited additional inputs integrated and aligned via an algorithm that maps active or events into metrics quantifying integrity severity or threat. We are on the track of addressing the above problem from a data-driven perspective for improved risk control and integrity management while maximizing the efficiency of resources.

Corrosion defect initiation and growth are influenced by a large number of factors and several data mining techniques have been used previously to model and predict pipeline integrity status.^{10,11,12} Among others, clustering analysis is a popular technique to categorize the data points/vectors into distinct homogeneous groups based on their similarity. Bayesian unsupervised clustering has been previously employed and found to be an effective tool to assess the soil corrosivity for external integrity management of a pipeline structure.¹³ The soil environment along the right-of-way is heterogeneous and consist of various levels of corrosivity. Hence, the severity of an external corrosion threat is position dependent. Hidden Markov random field (HMRF) models are widely used in problems where a specially constrained clustering scheme is required.¹⁴ To be more specific, the combination of the Markov random field theory and the finite mixture model can characterize the spatial correlation/variation of the soil survey data, resulting in a classification that is more realistic and easier to interpret in terms of spatial contexture constraint.¹⁵ In the present work, Markov random field and Gaussian mixture model (GMM) are employed for clustering pipeline segments buried in heterogeneous soil environment having various corrosivity levels.

EXPERIMENTAL PROCEDURE

Model setting for soil corrosivity assessment

In the present study, a pipeline system of length 110km, diameter 457.2mm and thickness 6.41mm is used. The pipeline is divided into multiple segments having 100m per segment for further analysis. In order to detect the heterogeneity of the soil environment along the pipeline right-of-way, clustering analysis is performed for grouping pipeline segments into several clusters based on some similarity measure regarding soil features (i.e., soil physical and chemical properties as well as environmental information at a larger scale).

Since the soil corrosivity corresponding to each pipeline segment is not known a priori, the topic under investigation is essentially an unsupervised learning problem. In the current context, the database consists of numerous corrosion related features: physical soil properties like bulk density of soil, percentage sand content, clay content, coarse fragment, and silt content; large scale environmental features like elevation, yearly mean precipitation and average normalized difference vegetation index (NDVI); soil electrochemical properties from site samples such as pH, different ion concentrations, half-cell potential, and resistivity. Finding a set of non-redundant soil features that contribute to corrosivity assessment is crucial.^{16,17} We use the wrapper method for feature selection. To be more specific, the

wrapper method estimates the importance of a feature subset by evaluating the likelihood of the corresponding clustering result as the quality measure after applying a specific learning model.¹⁸ Wrapper methods are preferred for the high quality of the selected feature subsets, but they have a high computational cost.¹⁹ In this work, Bayesian information criteria (BIC) along with the Gaussian mixture model (GMM) is used in the wrapper framework to determine the optimal features.

By conducting the wrapper method, ten soil features - soil resistivity, pH, Eh (soil redox potential), SO_4^{2-} concentration, Cl^- concentration, HCO_3^- concentration, and CO_3^{2-} concentration, elevation, yearly mean precipitation and soil bulk density are selected. Besides, in order to incorporate the variability of soil aeration condition along the pipeline right-of-way, the top-bottom difference and mean bulk density is used in the clustering analysis. This gives a total of eleven features, which is still too many for clustering analysis. Hence, in order to further reduce the dimensionality but retain all relevant information, principal component analysis (PCA) is applied on the eleven-feature dataset. PCA creates new uncorrelated variables that successively maximize variance.²⁰ The most critical problem in using PCA is the determination of number of principle components to retain for the data analysis. Retaining more components leads to an attempt to ascribe physical meaning to noise in the data set, while retaining too few components may cause the researcher to discard valuable information. Scree test and proportion of total variance explained are two heuristic techniques for determining the dimensionality of data sets.²¹ In a PCA model, each eigenvalue represents the level of variation explained by the associated principal component. Figure 1 shows the Scree plot of percentage of explained variance against each principal component.

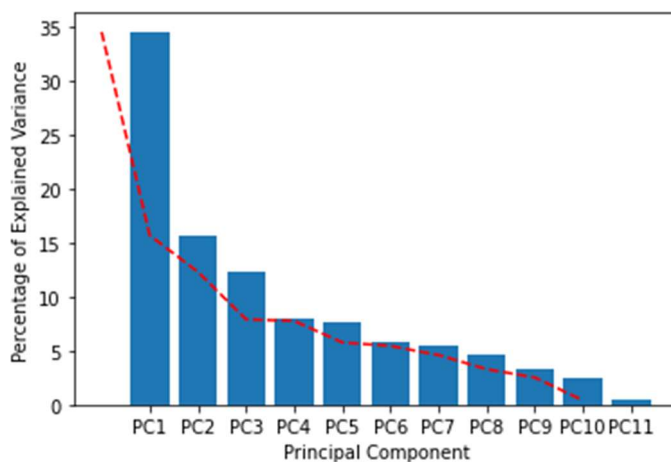


Figure 1: Scree plot of percentage of explained variance against the principle components.

The location on the graph where a sharp change in slope occurs in the line segments joining the points is referred to as an elbow. The value of principal component at which this occurs represents the number of components that should be retained in the PCA. But the Scree plot fails when there is no clearly defined break or has more than one break as in the case of the present dataset. This problem can be remedied by a simple and popular stopping rule based on the proportion of the total variance explained by the principal components retained in the model. If p components are retained, then we may represent the cumulative variance explained by the first p PC's by

$$t_p = \frac{\sum \lambda_p}{\text{trace}(S)} \quad (1)$$

where S is the sample covariance matrix and λ_p represent the p -th eigenvalue obtained from the covariance or correlation matrix. In practice, it is common to use some predefined percentage of total variance explained between 70% to 95% to decide how many PCs should be retained. Figure 2 shows cumulative explained variance against number of components with a threshold of 95% which gives us seven as an optimal number of components.

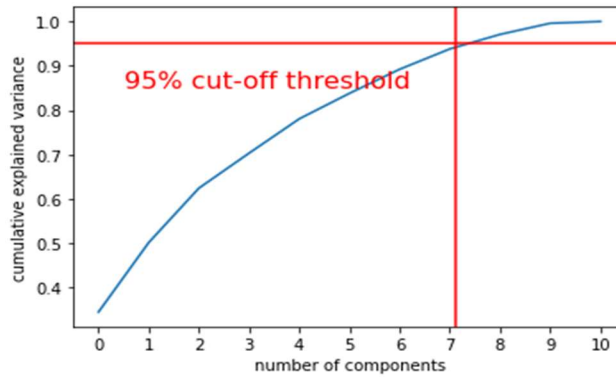


Figure 2: Cumulative explained variance against the number of principle components.

Another essential input model parameter for clustering is the number of clusters that best fits a given dataset. A proper number of clusters can cause the data points having high intra-similarities within each cluster and low inter-similarities between clusters within the feature space.²² For model-based clustering, the Approximate Weight of Evidence criterion (AWE) is used to determine the number of clusters. When Expectation-Maximization (EM) algorithm is used to find the maximum mixture likelihood for a Gaussian mixture model, an AWE called Bayesian Information Criterion (BIC) is applicable. The BIC can be used for model selection by including the number of clusters as a penalty and therefore more sensitive to the model complexity.²³ BIC has the following expression.

$$BIC = 2\log\text{like}(x, \theta) - M \log(n) \quad (2)$$

where $\log\text{like}(x, \theta)$ is the maximized log-likelihood, M is the number of independent parameters to be estimated, and n is the number of data points. A model with a higher BIC is preferred, since if data fits well to the model, the log-likelihood would be higher (more likely), and M should be a minimum (simpler). According to previous studies selecting the number of components where BIC will not significantly increase is appropriate.^{23, 24, 25}

Clustering analysis for identifying heterogeneity of soil corrosivity

The Hidden Markov Random Field (HMRF) model characterizes both the statistical pattern of the model setting described above and the spatial constraint (i.e., nearby soil environment should have a similar level of soil corrosivity) in the physical space along the pipeline right-of-way and hence is an ideal fit for grouping one-dimensional pipeline segment sequence according to their local soil corrosivity condition.

Here we consider $L = \{0, 1, 2, \dots, k-1\}$ as the set of soil states indicating different soil corrosivity conditions, and k denotes the total number of soil states. Note that the state numbers have no semantic meaning but only indicate different corrosivity conditions. Let $S = \{1, 2, 3, 4, \dots, s\}$ denote the sequential ID of segments from the beginning to the end of the pipeline. The state space of a segment, $j \in S$, is given as $X_j = \{x_j | x_j \in L\}$ and hence the configuration space of the entire pipeline can be defined as $X = \prod_{j=1}^S X_j = L^S$. The state values of all the sites $x = (x_1, x_2, x_3, \dots, x_s)$ are a configuration in the product space X and the configuration defined on S is considered as a random field given as $p(x)$. Let y_j denote the PCs at segment j . The feature space random field $p(y)$ is also a product space denoted as $Y = \prod_{j=1}^S Y_j$, $Y_j = \{y_j | y_j \in \mathbb{R}^d\}$ and $y = (y_1, y_2, y_3, \dots, y_s)$ is a realization of $p(y)$. The d -dimensional space $\mathbb{R}^d$ is the feature space framed by all PCs.

The main characteristics of an HMRF are as follows:

1. The soil configuration of all sites $x = (x_1, x_2, x_3, \dots, x_s)$ i.e. the clustering result, is unobservable.
2. The observed random field $p(y|x)$, is generated by any particular configuration $x = (x_1, x_2, x_3, \dots, x_s)$.