International Journal of Pressure Vessels and Piping 126-127 (2015) 37–47

Contents lists available at ScienceDirect

International Journal of Pressure Vessels and Piping

journal homepage: www.elsevier.com/locate/ijpvp

A clustering based method to evaluate soil corrosivity for pipeline external integrity management

Ayako Yajima [a], Hui Wang [a], Robert Y. Liang [a,*], Homero Castaneda [b,1]

[a] Department of Civil Engineering, The University of Akron, Akron, OH 44325-3905, USA
[b] Department of Chemical and Biomolecular Engineering, National Center for Research in Corrosion and Materials Performance, The University of Akron, Akron, OH 44325-3906, USA

ARTICLE INFO

Article history:
Received 21 July 2014
Received in revised form
6 December 2014
Accepted 18 December 2014
Available online 26 December 2014

Keywords:
Soil corrosivity
Clustering analysis
Gaussian mixture model
Pipeline
Data mining

ABSTRACT

One important category of transportation infrastructure is underground pipelines. Corrosion of these buried pipeline systems may cause pipeline failures with the attendant hazards of property loss and fatalities. Therefore, developing the capability to estimate the soil corrosivity is important for designing and preserving materials and for risk assessment. The deterioration rate of metal is highly influenced by the physicochemical characteristics of a material and the environment of its surroundings. In this study, the field data obtained from the southeast region of Mexico was examined using various data mining techniques to determine the usefulness of these techniques for clustering soil corrosivity level. Specifically, the soil was classified into different corrosivity level clusters by k-means and Gaussian mixture model (GMM). In terms of physical space, GMM shows better separability; therefore, the distributions of the material loss of the buried petroleum pipeline walls were estimated via the empirical density within GMM clusters. The soil corrosivity levels of the clusters were determined based on the medians of metal loss. The proposed clustering method was demonstrated to be capable of classifying the soil into different levels of corrosivity severity.

1. Introduction

Several factors may contribute directly or indirectly to the structural failure of a buried pipeline, and corrosion is one of the critical causes [1]. Corrosion propagation is due to the direct contact of the pipeline structure with an aggressive soil environment. To assess the external deterioration of the pipeline structure and guide maintenance practices, determining the soil corrosivity in the right-of-way of a pipeline structure is a matter of great concern. In addition, pipeline structures as a kind of high capital investment must be free from risk of degradation, which could cause threats to life and environmental hazards. Hence, assessing the corrosivity of soil is important for designing a pipeline structure as well as performing the risk assessment during its service period.

Metal deterioration in an aggressive environment has been well studied by using different approaches (e.g., power functions as a deterministic approach, extreme value distributions of probabilistic models, and by artificial neural networks (ANNs) as a data mining approach [2–9]). Previous studies also reveal that the surrounding environment and the physicochemical characteristics of metal greatly affect the corrosion propagation [8,10,11]. The National Association of Corrosion Engineers (NACE) and the American Society for Testing and Materials (ASTM) suggest soil resistivity as an important factor in the degradation process for buried pipelines, [12,13]. In general, lower resistivity soils seem to be more corrosive than those with higher resistivity, since the electrical current within the soil phase dominates the corrosion degradation at the structure/soil interface [10]. However, some researchers have pointed out that the correlation between soil resistivity and the metal loss rate under field conditions often seems to be weak [14]. The topography and physical/chemical characteristics of the soil are essential for designing a pipeline structure and selecting protection systems [6,10,12]. Therefore, a comprehensive approach which can take more factors into consideration is considered to be more robust and realistic under *in situ* conditions. The American Water Works Association (AWWA) recommends using a 10-point scoring method with five factors (soil resistivity, pH, soil redox potential, sulfate concentration, and moisture content) to assess the soil

* Corresponding author. Tel.: +1 (330) 972 7190; fax: +1 (330) 972 6020.
E-mail addresses: ay11@zips.uakron.edu (A. Yajima), hw34@zips.uakron.edu (H. Wang), rliang@uakron.edu (R.Y. Liang), homeroc@uakron.edu (H. Castaneda).
1 Tel.: +1 (330) 972 7398.

http://dx.doi.org/10.1016/j.ijpvp.2014.12.004

corrosivity for a cast iron alloy [14–17]. Several modified scoring methods that include a larger number of factors are also available [14]. The drawback of using a scoring method is that it is additive and hence cannot handle nonlinear relation among the factors. In addition, scoring methods neglect some important factors – such as the corrosive effect of chlorides – that are critical to the deterioration rate of metals.

Data mining techniques are considered to be more powerful and flexible in dealing with multi-factors corrosivity assessment, and several approaches (e.g., artificial neural network, fuzzy based method, and random forest) have been previously studied to assess the soil corrosivity [11,14–16]. However, the limitation of the previous approaches is that the implementation requires class labels (e.g., corrosive or non-corrosive) in order to model the data, but typical field data do not provide this information in advance. Clustering is a major technique of data mining that subgroups the data points based on their similarity, and it has wide application in many areas [18–21]. Clustering is used to find hidden patterns in the data when a response variable is not explicitly given. Oil and gas operators routinely use in-line inspection (ILI) to detect, size and locate the corrosion defects, and clustering techniques make full use of the inspection data and enable the operators to gain more information about the latent patterns of the external corrosion degradation. Therefore, employing clustering techniques is considered to be a superior choice for developing an assessment approach for soil corrosivity. Within the scope of clustering techniques, distance-based and model-based clustering methods are preferred because 1) they classify the soil data in an unsupervised manner and hence overcome the drawbacks mentioned above; 2) they are easy to implement. Establishing a clustering technique in the context of assessing soil corrosivity for pipeline structures is a novel approach.

In this study, we aim to establish a clustering based assessment approach for external integrity management by incorporating corrosion depth data obtained from routinely performed ILIs. Two types of clustering models, the Gaussian mixture model (GMM) and the k-means algorithm, are considered and compared with conventional criteria established by Peabody recommended by NACE [12]. The present approach is applied to the soil survey data within the right-of-way of a 110-km underground pipeline structure located in the southeast region of Mexico. The clustering approach is proved to be a proper technique for soil corrosivity classification because both soil survey data and ILI inspection results for defect depth show multimodality (with several modes/peaks), which implies that the soil environment along the right-of-way is heterogeneous and may consist of various corrosivity levels [22].

This paper is organized as follows: In Section 2, we describe two types of clustering approaches, k-means and the Gaussian mixture model, and we consider the model selection criteria. In Section 3, we illustrate the procedure of clustering for soil corrosivity assessment. The analysis of two field data sets is presented in Section 4. The clustering results are also tested by the Kruskal–Wallis test [23] for the statistical significance and depicted in Section 4. The discussions on applications in industry practice are provided in Section 5. The article is concluded in Section 6.

2. Clustering methods

Clustering models are used to separate the data into distinct homogeneous groups which, in this study, should ideally represent different corrosivity levels in the soil. The process of soil corrosivity assessment is based on clustering models using two sets of data, 1) soil survey data and 2) ILI measurements. Before we describe the complete procedure of the clustering approach, we introduce two different clustering models that can be used in the assessment approach: a distance-based model and a finite mixture model.

2.1. Distance-based model

1) The *k*-means algorithm

The *k*-means clustering algorithm is the most prevalent and intuitive clustering algorithm that partitions the data by assigning each data point to its nearest center of *k* clusters. This algorithm has been used for over 50 years [20] and has numerous variations (e.g. fuzzy *c*-means, *k*-means, nonparametric Bayesian clustering and others [20,24]).

The standard *k*-means algorithm was applied to the data in this study. This algorithm requires the number of components k to be given in advance. At the initial step, the cluster label $\{c_i\}_{i=1}^{n}$ will be randomly assigned to all observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where each observation $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ is d-dimensioned. The center μ_c for the cth cluster is the mean of the observations belonging to cluster c. Let n_c be the number of observations in the cluster c. Then, for each data point, the center μ_c of the nearest cluster in Euclidean distance is identified. At the next step, the center of each cluster is updated by the average of all data points in the cluster. This process is repeated until the centers stop moving. Fig. 1 summarizes the *k*-means algorithm.

Initialize: the cluster label $\{c_i\}_{i=1}^{n}$ randomly assigned to $x_i, i = 1, \ldots, n$

Step 1: Update cluster label: $i = 1, \ldots, n$

$$c_i \leftarrow \arg\min_c \sum_{j=1}^{d} \| x_{ij} - \mu_{cj} \|^2$$

Step 2: Update cluster centers: $c = 1, \ldots, C$ and $j = 1, \ldots, d$

$$\mu_{cj} \leftarrow \frac{\sum_{i:c_i=c} x_{ij}}{|\{i : c_i = c\}|}$$

Repeat until converge

Fig. 1. The *k*-means algorithm.