NRC Publications Archive Archives des publications du CNRC

Exploring the relationship between soil properties and deterioration of metallic pipes using predictive data mining methods

Liu, Z.; Sadiq, R.; Rajani, B. B.; Najjaran, H.

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

Publisher's version / Version de l'éditeur:

[https://doi.org/10.1061/\(ASCE\)CP.1943-5487.0000032](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000032)

Journal of Computing in Civil Engineering, 24, May-June 3, pp. 289-301, 2010-05-01

NRC Publications Record / Notice d'Archives des publications de CNRC:

<https://nrc-publications.canada.ca/eng/view/object/?id=18d9ce84-b1a4-457f-a5bf-ada38458dac0>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=18d9ce84-b1a4-457f-a5bf-ada38458dac0>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

<http://www.nrc-cnrc.gc.ca/irc>

Exploring the relationship between soil properties and deterioration of metallic pipes using predictive data mining methods

NRCC-53534

Liu, Z.; Sadiq, R.; Rajani, B.B.; Najjaran, H.

May 2010

A version of this document is published in / Une version de ce document se trouve dans:
Journal of Computing in Civil Engineering, 24, (3), May-June, pp. 289-301,
May 01, 2010, DOI: [10.1061/\(ASCE\)CP.1943-5487.0000032](https://doi.org/10.1061/(ASCE)CP.1943-5487.0000032)

The material in this document is covered by the provisions of the Copyright Act, by Canadian laws, policies, regulations and international agreements. Such provisions serve to identify the information source and, in specific instances, to prohibit reproduction of materials without written permission. For more information visit <http://laws.justice.gc.ca/en/showtdm/cs/C-42>

Les renseignements dans ce document sont protégés par la Loi sur le droit d'auteur, par les lois, les politiques et les règlements du Canada et des accords internationaux. Ces dispositions permettent d'identifier la source de l'information et, dans certains cas, d'interdire la copie de documents sans permission écrite. Pour obtenir de plus amples renseignements : <http://lois.justice.gc.ca/fr/showtdm/cs/C-42>

National Research
Council Canada

Conseil national
de recherches Canada

Canada

Exploring the Relationship between Soil Properties and Deterioration of Metallic Pipes Using Predictive Data Mining Methods

Zheng Liu¹, Rehan Sadiq², Balvant Rajani³, and Homayoun Najjaran⁴

Abstract

Soil corrosivity is considered to be a major factor for the deterioration of metallic water mains. Using a 10-point scoring method as suggested by the American Water Works Association, soil corrosivity potential can be estimated by five soil properties: (1) resistivity, (2) pH value, (3) redox potential, (4) sulfide, and (5) percentage of clay fines. However, the relationship between soil corrosivity and pipe deterioration is often ambiguous and not well defined. In order to identify the direct relationship between soil properties and pipe deterioration, which is defined as the ratio of the maximum pit depth to pipe age, predictive data mining approaches are investigated in this study. Both single- and multi-predictor based approaches are employed to model such relationship. The advantage of combining multiple predictors is also demonstrated. Among all approaches, rotation forest achieves the best result in terms of the prediction error to estimate pipe deterioration rate. Compared to the random forest method, which is the next to the best, the normalized mean square error decreased 50%. With the proposed approaches, the assessment of pipe condition can be achieved by analyz-

¹Research Officer, Institute for Research in Construction, National Research Council Canada, 1200 Montreal Road, Building M-20, Ottawa, Ontario K1A 0R6 Canada E-mail: zheng.liu@nrc-cnrc.gc.ca

²Associate Professor, School of Engineering, University of British Columbia (Okanagan Campus) Kelowna, British Columbia V1V 1V7 Canada Email:rehan.sadiq@ubc.ca

³Principal Research Officer, Institute for Research in Construction, National Research Council Canada, 1200 Montreal Road, Building M-20, Ottawa, Ontario K1A 0R6 Canada E-mail:balvant.rajani@nrc-cnrc.gc.ca

⁴Assistant Professor, School of Engineering, University of British Columbia (Okanagan Campus) Kelowna, British Columbia V1V 1V7 Canada Email:homayoun.najjaran@ubc.ca

ing soil properties. This study also highlights the importance for collecting more reliable soil properties data.

Keywords: cast iron pipe, condition assessment, pitting corrosion, soil corrosivity potential, deterioration rate, data mining.

Introduction

Corrosion is the most important factor that contributes to the failure of metallic water mains (Sadiq et al. 2004), while other factors, such as casting technology, manufacturing defects, and environmental conditions are contributing factors that affect the structural properties of water pipes. A number of criteria have previously been proposed to assess the structural condition of a pipe, among which breakage frequency and corrosion pit depths are common ones (Kleiner and Rajani 2001). Corrosion in cast iron pipes generally takes place in the form of graphitic corrosion, where iron leaches out leaving behind the graphite matrix (Sadiq et al. 2004). A corrosion pit develops, while the remaining carbon remains intact, therefore reducing the tensile strength of the pipe (Rajani et al. 2000).

Kleiner and Rajani (Kleiner and Rajani 2001) reviewed statistical models used to describe the structural deterioration of water mains, including both deterministic and probabilistic models that considered a multitude of factors, like pipe attributes, surrounding conditions, and water quality, and then related them to pipe breakage or failure models. The requirements and cost for collecting the required data varied with the system complexity.

The surrounding environment of the buried pipe has a significant contribution toward its deterioration, which may ultimately lead to failure. Soil corrosivity is one of the major factors linked to pipe deterioration. Robinson described the relationship between environmental factors and the corrosive nature of soil (Robinson 1993), where soil resistivity is considered the most important factor to determine soil corrosivity. Seica et al. (Seica et al. 2002) reported that underground corrosion was primarily promoted by the presence of moisture, oxygen, redox potential, pH, soil resistivity, and microbial activity. They applied four soil test procedures to assess the soil aggressiveness, including soil description and identification,

resistivity, pH, and sulfide (Seica et al. 2002). A strong correlation between average pitting rate and soil resistivity was identified. However, such a correlation has not yet been characterized.

A frequently used approach to determine soil corrosivity for cast and ductile iron pipes is the 10-point scoring method (AWWA 1999), which uses five soil properties: (1) resistivity, (2) pH value, (3) redox potential, (4) sulfide, and (5) percentage of clay fine (classified as soil type). The 10-point scoring method classifies the soil either as corrosive or non-corrosive. Another available method, which considered twelve factors, was suggested by Metalogic (1998). These factors included soil type, soil resistivity, water content, pH, buffering capacity, sulfide, chloride and fulfate concentrations, ground water level, horizontal and vertical soil homogeneities, and electrochemical potential. Metalogic divided soil corrosivity into four categories: virtually not corrosive, slightly corrosive, corrosive, and highly corrosive. Spickelmire (2002) used a similar four-part division (mild, moderate, appreciable, and severe) of soil corrosivity but employed a 25-point scoring method. In addition to soil properties, other factors such as pipe location and leak repair difficulty, pipe minimum design life, pipe maximum design surge pressure, and pipe size were also taken into account (Spickelmire 2002). These latter factors also related to failure consequences.

Instead of using a binary result to describe soil corrosivity, Sadiq et al. (2004) proposed a fuzzy-based method to classify soil corrosivity into three categories: non-corrosive, moderately corrosive, and corrosive. Membership functions for each type of soil properties were defined as three continuous ordinal states. Based on an analytical hierarchy process (AHP), a weighting matrix was defined for five soil properties. The mean of maximum method was used for defuzzification to assign the state of soil corrosivity potential. Najjaran et al. (2006) developed a multi-input and single-output (MISO) fuzzy inference system using a knowledge base developed by experts. The knowledge base was transformed into a set of IF-THEN rules. This model generated a membership value over an interval $[0, 1]$, where “zero” and “one” correspond to the non-corrosive and most corrosive soil, respectively. The advantage

of this method is that a continuous numeric value is used to represent the corrosivity potential of the soil. However, none of the aforementioned methods were able to establish a strong relationship between the soil corrosivity potential and pipe deterioration expressed as either pipe breakage or pitting rate. A summary of previous studies is given in Table 1.

[Table 1 about here.]

The schematic of previously developed approaches to predict corrosivity potential is illustrated in Fig. 1. In those studies, soil properties were used to determine a physically non-measurable quantity called soil corrosivity potential (CP). The definition of soil corrosivity potential may vary from one approach to another and is a qualitatively defined variable. Moreover, the effectiveness of the estimation for soil corrosivity potential needs to be validated or verified by a physical measurement like the pipe deterioration (breakage or pitting) rate. Therefore, it is worthwhile to establish a direct connection between soil properties and actual pipe deterioration data.

[Fig. 1 about here.]

In this study, predictive data mining approaches were used to explore the direct relationship between the soil properties and pipe deterioration, which is represented by the growth rate of corrosion pits. Both single- and multi-predictor based approaches were investigated. The state-of-the-art ensemble method, namely rotation forest, was applied to the collected soil property data (Rodriguez et al. 2006). This study was conducted in the environment of the statistical computing language *R*, which is widely used for statistical software development and data analysis (R Development Team 2007). *R* is considered as a free implementation of the S programming language (R Development Team 2007).

Data Mining Approaches

Figure 2 illustrates a comprehensive framework for the assessment of pipe performance. Multi-source information, such as soil properties, historical records, field inspection data,

and sensory information is collected and processed to achieve a higher level description or metrics related to pipe conditions. The derived information can be further combined for a comprehensive assessment through a data fusion process (Najjaran et al. 2006). The data fusion process can be applied at different levels and stages, where the need arises. The models proposed in this paper to describe the relationship between soil properties and pipe deterioration are a component of this framework.

[Fig. 2 about here.]

Based on the principles of statistics and focuses on building a model that can generate useful predictions, data mining is defined as “the nontrivial extraction of implicit, previously unknown, and potentially useful information from data” and “the science of extracting useful information from large data sets or databases” (Hand et al. 2001). Generally, the process of data mining consists of following three stages: 1) the initial exploration, 2) model building with validation/verification, and 3) deployment (Lewicki and Hill 2005). The process starts with data preparation, which may involve data cleaning, transforming, and selecting operations. In the second stage, various models are tested, and the best one is chosen. Finally, the selected model is applied to the newly extracted data set, in order to generate predictions or estimates of the expected outcome (Lewicki and Hill 2005).

In this study, soil properties were used to estimate the pipe deterioration rate, which is expressed as the ratio between the maximum pitting depth and pipe age. Once the relationship is established, the derived deterioration rate can be used as a factor to assess pipe performance.

The selected data mining methods are organized in two categories: single- and multi-predictor. The first group includes multiple linear regression, partial least square regression, principle component regression, and regression tree. The second group contains bundling method, bagging tree, random forest, and rotation forest. Some of these methods have also been discussed and applied extensively in others fields such as Hand et al. (2001), Duda et al. (2001), and Theodoridis and Koutroumbas (2006).

Variables and Data Summarization

Five soil properties, which are suggested by American Water Works Association (AWWA 1999) and described in Table 2, were considered in this study. The five variables are described and summarized in Table 2. The data set used in this analysis was collected in 1990s from 45 locations in North American cities to study the growth of corrosion pits in cast iron pipes (Sadiq et al. 2004). The summary of the data is shown in Table 3, and the corresponding scatter plot is given in Fig. 3, which gives a snapshot of how much one variable is affected by another. For each plot, the coordinates are determined by the variables on the corresponding row or column. A rough relationship of two independent variables can be found. Table 3 provides a series of statistical information including mean, median, quartiles (Qu.), and extreme values. The 1st quartile is the value below which lies 25% of the data, while 3rd quartile is the value below which lies 75% data points.

[Table 2 about here.]

[Table 3 about here.]

The histogram and box plots of resistivity, pH, soil type, and redox potential are given in Fig. 4. Sulfide is a factor variable, where the “negative”, “trace” and “positive” are represented in the experiment by -1, 0, and 1 respectively. The histogram of the sulfide data is shown in Fig. 5. The histogram plots, such as Fig. 4(a), show the frequency counts for each variable. The distribution of the data can be roughly observed. The box plot gives a quick summarization of the variable distribution. The first and third quartiles of a variable consist of the box bound. The horizontal line inside the box represents the median value. The small horizontal dashes show the largest/smallest observation that falls within a distance of 1.5 times the box size from the nearest hinge. The outliers can be identified from the box plot of soil resistivity and pH value. Soil type is mainly concentrated on three values, i.e. 22%, 30% and 42% of clay fines (Fig. 4(e) and 4(f)). These figures provide a fast screening of the raw data.

As shown in Fig. 4(b), there are outliers in the resistivity data, which can be removed to improve the result. Usually, the removed points can be replaced by the estimation from correlated variables (Torgo 2003). However, no obvious association of the variable was identified. Therefore, all the data was used in this study. It is understood that those outliers may introduce errors to the prediction.

The plots of soil properties against pipe deterioration rate are given in Fig. 6. Again, no explicit trends were observed. However, Fig. 6(a) indicates that low soil resistivity may contribute to a higher pipe deterioration rate. A higher resistivity corresponds to a lower pipe deterioration rate.

[Fig. 3 about here.]

[Fig. 4 about here.]

[Fig. 5 about here.]

[Fig. 6 about here.]

External pitting corrosion starts from the outer pipe surface. The corrosion is confined to a point or small area and forms cavities. A narrow pit with minimal overall metal loss can lead to the failure of a whole pipe (Roberge 2007). Pitting is expected to start just after installation, unless some prior observation to the contrary is recorded. There are different ways to define the deterioration rate (Najjaran et al. 2006). Therefore in this study the deterioration rate is characterized by the *maximum* pit depth and the pipe age ($t = T$), as given below:

$$D = \frac{1}{T} \max (d_1, d_2, \dots, d_i, \dots, d_n) \quad (1)$$

where D refers to the pipe “deterioration rate”; d_i represents the depth, assuming n pits were measured and recorded for a given pipe. Here, i is the number of pits on the surface

area under study and T is the pipe age at the time of inspection. As illustrated in Fig. 7, the *maximum* pitting depth is used to calculate the D value, which gives a *maximum* rate during the specific period of time. In reality, the rate of corrosion pit growth is generally not constant over time, and it is hard to determine the rate changes with time. In this study, a direct relationship between D and the five soil properties is proposed.

[Fig. 7 about here.]

Methods with One Predictor

Multiple Linear Regression (MLR)

The linear regression is a well-established method to determine the relation between the response variable e.g. pipe deterioration rate D and the explanatory independent variables including soil properties $\mathbf{x}_i = \{x_{ij}\} = \{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}\}$ ($i = 1, \dots, n$, $j = 1, \dots, 5$). The multiple linear regression model is given by equation (2).

$$D_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i \quad (2)$$

where the subscript i refers to the i^{th} individual sample. In this case, 45 samples were used ($n = 45$). The residual term ε_i , $i = 1, 2, \dots, n$, is assumed to be an independent random variable, which is normally distributed with zero mean and constant variance σ^2 . The regression coefficients $\{\beta_0, \beta_1, \dots, \beta_5\}$ are the coefficients of the linear regression model. The regression coefficients are calculated by the least square estimation. Equation (2) can also be rewritten in the vector form as:

$$\mathbf{D} = \mathbf{X}\beta + \varepsilon \quad (3)$$

where there are $\mathbf{D}^T = (D_1, \dots, D_n)$, $\beta^T = (\beta_0, \beta_1, \dots, \beta_q)$, and $\varepsilon^T = (\varepsilon_1, \dots, \varepsilon_n)$. The matrix $\mathbf{X}$ is:

$$\mathbf{X} = \begin{pmatrix} 1 & \mathbf{x}_1 \\ 1 & \mathbf{x}_2 \\ \vdots & \vdots \\ 1 & \mathbf{x}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} \end{pmatrix} \quad (4)$$

When the cross-product $\mathbf{X}^T \mathbf{X}$ is non-singular, the least square estimation of β is:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{D} \quad (5)$$

If $\mathbf{X}^T \mathbf{X}$ is singular, the model needs to be reformed as $\mathbf{y} = \mathbf{X}\mathbf{C}\beta^* + \varepsilon$ such that matrix $\mathbf{X}\mathbf{C}$ has a full rank (Everitt and Hothorn 2006). Once β is obtained, the MLR model is established.

Principal Component Regression (PCR) and Partial Least Square Regression (PLSR)

Since the predictor variables in vector $\mathbf{x}$ may be highly related to each other, the purpose of principal component analysis (PCA) is to generate features that are optimally uncorrelated so that redundant information is avoided. The PCA analysis is also known as Karhunen-Loève transform (Theodoridis and Koutroumbas 2006). Singular value decomposition (SVD) can be used to derive the principal components (Shlens 2005). The matrix of soil properties is reprinted in equation 6 as:

$$\mathbf{Z} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & x_{n5} \end{pmatrix} \quad (6)$$

The singular value decomposition of matrix $\mathbf{Z}$ can be expressed as: $\mathbf{Z} = \mathbf{S}\Delta\mathbf{V}^T$ with $\mathbf{S}^T \mathbf{S} = \mathbf{V}^T \mathbf{V} = \mathbf{I}$. Matrix Δ is a diagonal matrix with singular values as diagonal elements. The

left singular vectors (i.e. the columns of $\mathbf{S}$, are the principal components of $\mathbf{Z}$). Only a few of them will be used to predict $\mathbf{D}$ with a standard regression method (Abdi 2003). This is known as principal component regression. However, these components do not take $\mathbf{D}$ into account. Thus, the PLS regression addresses this issue through searching for a set of components that performs a simultaneous decomposition of input and output variables with the constraint that these components explain as much as possible of the covariance between them (Abdi 2003).

The PLS regression establishes the model $\mathbf{D} = \mathbf{X}\beta + \varepsilon$ by producing a weighted matrix $\mathbf{W}$ for $\mathbf{X}$ (i.e. $\mathbf{T} = \mathbf{XW}$, where $\mathbf{T}$ is known as factor score matrix). The factor score variables are a linear combination of the original variables. The weights are computed so that each of them maximizes the covariance between response (deterioration rate) and the corresponding factor scores (Lewicki and Hill 2005). The regression now becomes $\mathbf{D} = \mathbf{TQ} + \varepsilon$. Once $\mathbf{Q}$ is obtained, then $\beta = \mathbf{WQ}$ can be determined, and the prediction model is complete. There are a number of algorithms available to compute PLS regression factors. In this study, the SIMPLS algorithm was employed (Jong 1993), which calculates the PLS factors directly as linear combinations of the original variables. The algorithm is implemented in the “pls” package of *R* (Mevik and Wehrens 2007).

Regression Tree (RT)

A pattern can be classified through asking a sequence of questions (Duda et al. 2001). These questions can be asked in a “Yes/No” or “True/False” style, which does not need any notion of metric. The question sequence is managed in a directed decision tree. The classification starts from the root node. By asking for the value of a particular property of the pattern, follow the link to a descendent node. Therefore, a decision tree implements a kind of nonlinear classifier. When the decision tree is used for regression, it is known as a regression tree. In this application, a regression with a continuous outcome variable is expected. The advantage of the method is that it is relatively easy to implement and has the potential to reveal complex forms of structure (Maindonald and Braun 2003). Binary

splits on a single predictor variable are created to maximize the homogeneity of the two split groups, with respect to the response variable (Prasad et al. 2006).

[Fig. 8 about here.]

The typical structure of a regression tree is illustrated in Fig. 8. An ancestor node t is associated with a subset X_t , which is further split into two subsets, i.e. X_{t_Y} (node t_Y) and X_{t_N} (node t_N). The decision tree procedure consists of four basic components: a set of questions, splitting criterion, stop-splitting rule, and class assignment rule (Theodoridis and Koutroumbas 2006). The best question is chosen from the training data set based on the splitting criteria. A commonly used criterion is the node impurity, which can be expressed as (Theodoridis and Koutroumbas 2006):

$$I(t) = - \sum_{j=1}^M P(\omega_j | t) \log_2 P(\omega_j | t) \quad (7)$$

where $P(\omega_j | t)$ denotes the probability that a vector in subset X_t belongs to class ω_i , $j = 1, 2, \dots, M$. This expression is actually the entropy associated with the subset X_t . The regression tree performs the splitting based on the highest degree of impurity. Readers are referred to references (Theodoridis and Koutroumbas 2006) and (Duda et al. 2001) for an in-depth discussion.

In practice, a small change of the training data may result in a very different tree. This is due to the hierarchical nature of the tree classifiers. Bagging (bootstrap aggregating) can reduce variance and improve the generalization error performance. The method of bagging trees is described in the next section.

Methods with Multiple Predictors

Bundling Predictors

Combining multiple classifiers is known to improve the overall performance of classification (Hothorn and Lausen 2005). The idea can be extended to regressor bundling, where a

linear model and regression trees are combined (Hothorn 2003). The coefficients of a linear model are estimated by using a out-of-bag sample in a regression. The predictions of the regression on the bootstrap sample are used as an additional predictor for regression trees. Thus, the outcomes of multiple regressors are bundled by the regression trees (Hothorn 2003). A typical implementation with R is given in the Appendix.

Bagging Tree (BT)

The error from a single regression tree is partly due to the specific choice of the training data set. The idea of a bagging tree is to use the bootstrap technique to construct a number of training sets X^* by uniformly sampling the original training data set X with replacements (Breiman 1996; Theodoridis and Koutroumbas 2006). Each bootstrap data set is used to train a regression tree. The averaging of the multiple-tree results becomes the final output. Thus, the predictive error is reduced. The selection of the number of training sets is based on the available computational resources (Duda et al. 2001).

Random Forest (RF)

A random forest gets its name from the concept that each tree is grown with a randomized subset of predictors, and that a forest consists of a large number of trees. The use of a random forest has been reported in different applications, such as Prasad et al. (2006), Svetnik et al. (2003), Gislason et al. (2006), and Svetnik et al. (2004). According to Breiman (Breiman 2001a; Breiman 2001b), a random forest for regression is formed by growing trees depending on a random vector Θ , such that the tree predictor $\{h(\mathbf{x}, \Theta)\}$ takes on numerical values. The training set is independently drawn from the distribution of the random vectors $\mathbf{D}$ and $\mathbf{X}$. The random forest regressor is formed by averaging K of the trees $\{h(\mathbf{x}, \Theta_k), k = 1, \dots, K\}$. The algorithm is as follows (Breiman 2001a):

- Choose bootstrap samples from the original data;
- Find the best split based on randomly selected variables in the training set;
- Predict new data by averaging the predictions of the K trees.

The random forest differs from the Bagging tree approach in that the algorithm uses a random sample of variables, instead of all the variables, for splitting at each node for each tree. The advantage of the random forest method is that overfitting is not possible due to the large number of trees. Meanwhile, the random forest can reduce the variance and keep a low bias. Moreover, variable importance can be evaluated based on the performance of prediction, when the data for this predictor is randomly permuted.

Rotation Forest (RTF)

A new classifier ensemble method, namely rotation forest, was recently proposed by Rodriguez et al. (Rodriguez et al. 2006). Similar to random forest, the rotation forest builds base classifiers independently with decision trees, which are trained in a rotated feature space (Kuncheva and Rodriguez 2007). The attribute or feature is expressed as $F = \{X_1, X_2, \dots, X_p\}$. In this study, F is $\{R, P, RP, S, CF\}$ and number of feature $p = 5$. The feature set is first split into subsets, and then principal component analysis (PCA) is applied to each subset respectively. A classifier is then trained with this linearly transformed subset. A full feature set for each classifier is subsequently reconstructed in the ensemble.

Zhang et al. (2008) recently proposed using rotation forest to improve regression analysis. Assuming that there are f features, then the training data T is an $n \times p$ matrix. The basic steps include:

- Randomly split data set into K disjointed subsets. Each subset contains $M \leq K$ features;
- Let F_{ij} be the j th subset features in F_{ij} . For each subset, select the corresponding data in matrix X to constitute a new matrix X_{ij} and draw a bootstrap sample X'_{ij} from X_{ij} . Use X'_{ij} to do PCA, and store the results in matrix D_{ij} (an $M \times M$ matrix).
- Construct a sparse rotation matrix R_i with D_{ij} and rearrange the rows of R_i according to the original feature sequence. The rearranged rotation matrix is denoted by R_i^a . The training set for regressor C_i is $\begin{bmatrix} X R_i^a & Y \end{bmatrix}$.

For a given data point $\mathbf{x}$, the value obtained by regressor C_i is $C_i(\mathbf{x}R_i^a)$. The final prediction can be calculated as:

$$C(\mathbf{x}) = \frac{1}{T} \sum_{i=1}^T C_i(\mathbf{x}R_i^a) \quad (8)$$

Experimental Results

Model Validation and Development

The selected algorithms were tested with the aforementioned data set. The five soil properties were used to develop models for estimating pipe deterioration rate. The prediction error of the regression models can be calculated with the cross validation or the bootstrap methods. Here, the root mean square error (RMSE) is employed in a *k-fold cross-validation* procedure for the assessment of the regression models. The *k-fold cross validation* is useful when there is a limited number of samples available. In the experiment, k was assigned a value of 10. The choice of this number is based on the computational efficiency and size of the data. Generally, a larger group number is preferred, when less data is available. Thus, the 45 samples were partitioned into 10 subgroups. Of the ten subgroups, a single subgroup was retained as the validation data to test the model, and the other nine subgroups were used for training models. This process was repeated ten times. The true RMSE was estimated as the average of the test results. The results are provided in Table 4. The smaller RMSE value indicates a better model. Thus, the rotation forest based regression is the most successful one among all the tested models.

[Table 4 about here.]

The summary of the derived MLR model is given in Table 5, where each coefficient, its standard error, t -, and p - values are listed. The t -test value was used to check the importance of each coefficient using a null hypothesis, i.e. $\beta_i = 0$. The t -value is calculated as the ratio between the coefficient and its standard error. A p -value indicates the level at which the

hypothesis that the coefficient is null is rejected. For example, the number “0.0002” means that there is about 99.98% confidence that β_0 is not null.

[Table 5 about here.]

For both the PCR and PLSR models, two components were used for the modeling. The resulting plots are given in Fig. 10(b) and 10(c) respectively. The impact of the number of components used in the regression is reflected in the RMSE from cross validation (see Table 6). The smaller value indicates a good fit of the model. The reduced number of components can achieve a “reasonable” result compared to the use of all the components.

[Table 6 about here.]

For the regression tree shown in Fig. 11, only two variables (resistivity and pH value) are kept to estimate pipe deterioration rate by the “rpart” function in *R*. Only three numerical values of deterioration rate comprise the regression results. This explains why in Fig. 10(d) the predictions concentrate on three specific points. Obviously, this is not suitable for the application of predicting pipe deterioration rate. The growing of the trees to a “forest” may solve the problem.

The prediction results obtained with multi-predictor are given in Fig. 12. The rotation forest achieves a better prediction compared with other methods. The quantile-quantile (Q-Q) plot of the random forest prediction error is given in Fig. 9(a), which indicates the error is normally distributed. The importance of each variable is shown in a bar plot in Fig. 9(b). Among the five properties, soil resistivity appears to be the most important variable while sulfide is the least.

[Table 7 about here.]

The random forest and rotation forest were compared with different number of variables. Less important variables were removed. These two methods were applied to 5, 4, 3, and

2 variables respectively. The results are listed in Table 7. Using more variables gave a better result for both the RF and RTF methods. When five variables were involved in the prediction, the RTF achieved a better performance than the RF method. However, when the number of variables reduced, the RF achieved a better result.

From Table 7, it is found that with five variables the root mean square error reaches the minimum value, which means all the five variables contribute to the estimation. However, they are not equally important. Among the five, soil resistivity is the most important variable. The random forest can achieve a similar result with only two variables as with five. Therefore, when different variables are considered, the prediction approach must be chosen carefully.

[Fig. 9 about here.]

[Fig. 10 about here.]

[Fig. 11 about here.]

[Fig. 12 about here.]

Model Performance Assessment

To examine the goodness of fit of the model, a measure called coefficient of determination or percentage of variance (R^2) is used (Faraway 2002):

$$R^2 = 1 - \frac{\sum_{i=1}^n \varepsilon_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (9)$$

The larger value (near to 1) indicates a good fitness of the model; however, when comparing models which use a different number of explanatory variables R^2 need to be adjusted to account for the different degree of freedoms. The adjusted coefficient, which takes into account the number of parameters (q) of the regression model is given below:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n - 1}{n - (q + 1)} \quad (10)$$

The adjusted R^2 square ($\bar{R}^2$) value can be negative and will always be less than or equal to R^2 . The $\bar{R}^2$ is useful in the feature selection stage of model building. The goodness of a $\bar{R}^2$ value is application dependent. A lower value can be expected, when there is a lot of noise in the data or the variables tend to be more weakly correlated (Faraway 2002).

The criteria used to determine the goodness of fit of the prediction models include mean absolute error (MAE), mean square error (MSE), and normalized mean square error (NMSE). The definitions of the three criteria are given below:

$$MAE = \frac{1}{n} \sum_i^n |D_{pred}^i - D_{true}^i| \quad (11)$$

$$MSE = \frac{1}{n} \sum_i^n (D_{pred}^i - D_{true}^i)^2 \quad (12)$$

$$NMSE = \frac{\sum_i^n (D_{pred}^i - D_{true}^i)^2}{\sum_i^n (\bar{D}_{true} - D_{true}^i)^2} \quad (13)$$

The MSE is not measured in the same units as the target variable. The NMSE calculates a ratio between the performance of the model and a baseline predictor, usually taken as the mean value of the target variable (Torgo 2003). The $\bar{R}^2$ values for these models and evaluation results are presented in Table 8. A “better” estimation will reach a smaller value of MAE, MSE, or NMSE.

[Table 8 about here.]

As indicated in Table 8, the rotation forest method presented the best performance in terms of the selected goodness of fit criteria. Next is the random forest method. In the regression tree method, a single tree does not provide a continuous numerical result as expected. Using multiple trees, e.g. bagging tree or random forest, can solve this problem. It is observed that the bundling of multiple regressors is even better than bagging tree. The PCR and PLSR method achieved a similar performance, though the MLR’s performance

is still better. This implies that in this application, reducing the number of independent variables does not contribute to the estimation of deterioration rate. However, the PCR and PLSR methods are particularly useful when there is a very large set of independent variables.

Conclusion

Soil properties contribute to the growth of pitting corrosion of cast iron pipes. Previous studies correlated soil properties to a non-physical quantity called soil corrosivity potential, which does not have an explicit relation with the pipe conditions. The study presented in this paper is a first attempt to explore the direct relationship between soil properties and pipe deterioration rate by using data mining methods. The experimental results with single- and multi-predictor indicate an explicit connection of pipe deterioration rate with the five soil properties (resistivity, pH value, redox potential, sulfide, and percentage of clay fines), among which soil resistivity appears to be the most important factor contributing to pipe deterioration.

Methods using multiple predictors outperform those with single predictor. The bundling of a linear model and regression trees or a combination of multiple regression trees reduces the normalized mean square error up to 80%. The rotation forest achieved the best result among all the multi-predictor approaches examined in this paper. Randomly sampling rotated features in rotation forest method makes it easier to identify different patterns in the feature space.

This study also demonstrated the importance of collecting soil property data, which are essential for the pipe performance assessment, although still insufficient to fully characterize pipe performance. Data collection is a process of accumulation. The proposed models or data mining algorithms were developed and trained on available data sets. As more data become available over time, it is possible to update the models and identify other variables correlated to deterioration rate. This remains a topic to pursue in the future, i.e. integrating data collection into the model update or predicting algorithm implementation process.

A subjective model based on expert judgement introduces the human factor and model

validation issues, but an expert system is still suitable for the need of a higher-level inference in the data fusion framework as shown in Fig. 2. The data mining methods presented in this paper can be integrated into such a framework.

Acknowledgements

Dr. Chuanxia Zhang is acknowledged for helpful discussion and MatlabTM implementation of rotation forest algorithm.

REFERENCES

- Abdi, H. (2003). "Partial least square regression." *Encyclopedia of Measurement and Statistics*, N. J. Salkind, ed., Sage Publications, Thousand Oaks, CA, USA, 740–744.
- AWWA (1999). *American National Standard for Polyethylene Encasement for Ductile-Iron Pipe Systems*. American Water Works Association, Denver, CO, USA. ANSI/AWWA C105/A21.5-99.
- Breiman, L. (1996). "Bagging predictors." *Machine Learning*, 24(2), 123–140.
- Breiman, L. (2001a). "Random forests." *Machine Learning*, 45(1), 5–32.
- Breiman, L. (2001b). "Random forests." *Technical report*, University of California, Berkeley, CA 94720.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons Inc., Hoboken, NJ, USA, 2 edition.
- Everitt, B. S. and Hothorn, T. (2006). *A Handbook of Statistical Analyses Using R*. Chapman & Hall /CRC, Boca Raton, FL, USA.
- Faraway, J. (2002). "Practical regression and ANOVA using R." <http://cran.r-project.org/other-docs.html> (2008).
- Gislason, P. O., Benediktsson, J. A., and Sveinsson, J. R. (2006). "Random forests for land cover classification." *Pattern Recognition Letters*, 27, 294–300.
- Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Adaptive Computation and Machine Learning. The MIT Press, Cambridge, MA, USA.

- Hothorn, T. (2003). “Bundling predictors in R.” *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna.
- Hothorn, T. and Lausen, B. (2005). “Bundling classifiers by bagging trees.” *Computational Statistics and Data Analysis*, 49, 1068–1078.
- Jong, S. D. (1993). “SIMPLS: An alternative approach to partial least squares regression.” *Chemometrics and intelligent laboratory systems*, 18(3), 251–263.
- Kleiner, Y. and Rajani, B. B. (2001). “Comprehensive review of structural deterioration of water mains: Statistical models.” *Urban Water*, 3(3), 131–150.
- Kuncheva, L. I. and Rodriguez, J. J. (2007). *An Experimental Study on Rotation Forest Ensembles*, Vol. 4472 of *Lecture Notes in Computer Science*, 459–468. Springer.
- Lewicki, P. and Hill, T. (2005). *STATISTICS Methods and Applications*. StatSoft Inc., Tulsa, OK, USA.
- Maindonald, J. and Braun, J. (2003). *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, UK, 2 edition.
- Metalogic (1998). “Corrosion research and materials engineering. <<http://www.metalogic.be/>> (2003).
- Mevik, B.-H. and Wehrens, R. (2007). “The pls package: Principal component and partial least squares regression in R.” *Journal of Statistical Software*, 18(2), 1–24.
- Najjaran, H., Sadiq, R., and Rajani, B. (2006). “Fuzzy expert system to assess corrosivity of cast/ductile iron pipes from backfill properties.” *Computer Aided Civil and Infrastructure Engineering*, 21(1), 67–77.
- Prasad, A. M., Iversen, L. R., and Liaw, A. (2006). “Newer classification and regression tree techniques: Bagging and random forest for ecological prediction.” *Ecosystems*, 9, 181–199.
- R Development Team (2007). “The R project for statistical computing. <<http://www.r-project.org/>> (March 2009).
- Rajani, B., Makar, J., McDonald, S., Zhan, C., Kuraoka, S., C.K. Jen, and Viens, M. (2000). *Investigation of Grey Cast Iron Water Mains to Develop a Methodology for Estimating*

- Service Life*. AWWA Research Foundation, Denver, CO, USA.
- Roberge, P. R. (2007). "Pitting corrosion. <<http://www.corrosion-doctors.org/>> (March 2009).
- Robinson, W. C. (1993). "Testing soil for corrosiveness." *Materials Performance*, 32(4), 56–58.
- Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). "Rotation forest: A new classifier ensemble method." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10), 1619–1630.
- Sadiq, R., Rajani, B., and Kleiner, Y. (2004). "Fuzzy-based method to evaluate soil corrosivity for prediction of watermain deterioration." *Journal of Infrastructure Systems*, 10(4), 149–156.
- Seica, M. V., Packer, J. A., Grabinsky, M. W. F., and Adams, B. J. (2002). "Evaluation of the properties of Toronto iron water mains and surroundingsoils." *Canadian Journal of Civil Engineering*, 29, 222–237.
- Shlens, J. (2005). "A tutorial on principal component analysis. <<http://www.cs.cmu.edu/elaw/papers/pca.pdf>> (2008).
- Spickelmire, B. (2002). "Corrosion consideration for ductile iron pipe." *Materials Performance*, 41, 16–23.
- Svetnik, V., Liaw, A., Tong, C., and Culberson, J. C. (2003). "Random forest: A classification and regression tool for compound classificationandQSARmodeling." *Journal of Chemical Information and Computer Sciences*, 43(6), 1947–1958.
- Svetnik, V., Liaw, A., Tong, C., and Wang, T. (2004). *Multiple Classifier Systems*, Vol. 3077 of *Lecture Notes in Computer Science*, chapter Application of Breimans Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecule, 334–343. Springer.
- Theodoridis, S. and Koutroumbas, K. (2006). *Pattern Recognition*. Academic Press, Elsevier, USA.
- Torgo, L. (2003). "Data mining with R: Leaning by case studies.

`<http://www.liacc.up.pt/~ltorgo/DataMiningWithR/>` (2008).

Zhang, C.-X., Zhang, J.-S., and Wang, G.-W. (2008). “An empirical study of using rotation forest to improve regressors.” *Applied Mathematics and Computation*, 195, 618–629.

APPENDIX I. R IMPLEMENTATION OF BUNDLING

In the following program, pipe deterioration data is in vector “DR” and soil property data is in matrix “data”. Multiple linear regressors are bundled by the regression trees.

```
> mybundle<-list(list(model=lm, predict=function(object,newdata) predict(object,newdata)))
> mymod<- function(formula, data){
mod <- bagging(formula, data=data[,1:6], control=rpart.control(xval=0), comb=mybundle)}
> predict.bag<-function(object,newdata){
  predict(object, newdata = newdata[,1:6],type="prob") }
> errorest(DR ~ ., data = data, model = mymod, estimator = "cv",predict=predict.bag)
```

List of Tables

1	Summary of previous studies using soil properties to predict corrosivity potential.	25
2	Soil properties used for corrosivity study (Sadiq et al. 2004).	26
3	Summary of data used in the experiment (For Sulfides, the sample numbers for each category are counted).	27
4	Root mean square error (RMSE) for the 10-fold cross-validation of data mining models.	28
5	Summary of MLR model.	29
6	Impact of the number of components in PCR and PLSR models (RMSE from cross validation).	30
7	The comparison of random forest and rotation forest in terms of RMSE with different number of variables.	31
8	Model assessment.	32

Table 1. Summary of previous studies using soil properties to predict corrosivity potential.

Methods	Factors	Classification Results	Corrosivity Potential	References
10-point scoring method	resistivity, pH value, redox potential, sulfide, and soil type	binary	1) corrosive and 2) non-corrosive	AWWA (1999)
12-factor evaluation	soil type, soil resistivity, water content, pH value, buffering capacity, sulfide, chloride and fulfate concentrations, ground water level, horizontal and vertical soil homogeneities, and electrochemical potential	four categories	1) highly corrosive, 2) virtually not corrosive, 3) slightly corrosive, and 4) corrosive	Metalogic (1998)
25-point scoring method	pH value, sulfate content, redox potential, soil type, resistivity, sulfides, moisture, pipe size, pipe maximum design surge pressure factor, pipe minimum design life factor, pipe location and leak repair difficult factor, potential interference sources, pipe zone back fill materials, and additional factors to consider	four categories	mild corrosive, moderate corrosive, appreciable corrosive, and severe corrosive	Spickelmire (2002)
fuzzy-based method	same as in 10-point scoring method	three categories	1) non-corrosive (NC), 2) moderately corrosive (MC), and 3) corrosive (C)	Sadiq et al. (2004)
fuzzy inference	same as in 10-point scoring method	numerical value between [0, 1]	non-corrosive= 0 and most corrosive= 1	Najjaran et al. (2006)

Table 2. Soil properties used for corrosivity study (Sadiq et al. 2004).

Soil Properties	Symbol	Description
Resistivity ($\Omega\ cm$)	$R\ (x_1)$	Determined by the moisture content and the concentrations of the different ions and their mobilities.
pH value	$P\ (x_2)$	Determined by the contents of carbonic acid, minerals, organic or inorganic acids, and by industrial wastes or acid rain.
Redox potential (mV)	$RP\ (x_3)$	Determined by the oxygen concentration of the soil moisture.
Sulfides	$S\ (x_4)$	Group of minerals that are compounds of one or more metallic elements combined with the non-metallic element sulfur.
% clay fines (soil type)	$CF\ (x_5)$	Can be inferred from percent fines content and used as a surrogate measure for drainage characteristics.

Table 3. Summary of data used in the experiment (For Sulfides, the sample numbers for each category are counted).

Soil property	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Resistivity ($\Omega\ cm$)	410	1300	3000	6283	6599	32418
pH value	4.600	6.700	7.400	7.171	7.900	8.700
Redox potential (mV)	-166.00	1.0	205.0	165.6	294.0	400.0
% of clay fines (soil type)	15.00	22.00	30.00	29.18	35.00	42.00
Sulfides	Negative	11	Trace	17	Positive	17

Table 4. Root mean square error (RMSE) for the 10-fold cross-validation of data mining models.

	MLR	PLSR	PCR	RT
RMSE	0.0157	0.0161	0.0162	0.0180
	Bundling	BG	RF	RTF
RMSE	0.0156	0.0153	0.0144	0.0142

Table 5. Summary of MLR model.

	Coefficients	Estimate	Std. Error	t-value	p-value
(intercept)	β_0	7.278e-02	1.758e-02	4.140	0.000180
resistivity	β_1	-7.751e-07	2.726e-07	-2.843	0.007069
pH	β_2	-5.463e-03	2.180e-03	-2.506	0.016497
soil type	β_3	5.325e-04	2.605e-04	2.044	0.047751
redox potential	β_4	-2.836e-05	1.480e-05	-1.916	0.062705
sulfide	β_5	-3.633e-03	2.861e-03	-1.270	0.211737

Table 6. Impact of the number of components in PCR and PLSR models (RMSE from cross validation).

No. of Component	1 comps	2 comps	3 comps	4 comps	5 comps
PCR	0.01754	0.01767	0.01846	0.01601	0.01569
PLSR	0.01600	0.01616	0.01581	0.01516	0.01530

Table 7. The comparison of random forest and rotation forest in terms of RMSE with different number of variables.

Number of variables	5	4	3	2
RF	0.0144	0.0145	0.0149	0.0154
RTF	0.0142	0.0154	0.0150	0.0220

Table 8. Model assessment.

	MLR	PLSR	PCR	RT
adjust R-square	0.2811	0.0816	-0.0461	0.3361
MAE	0.01114	0.01216	0.01285	0.01049
MSE	0.00019	0.00024	0.00027	0.00017
NMSE	0.63720	0.81402	0.92723	0.58847
	Bundling	BG	RF	RTF
adjust R-square	0.50627	0.43820	0.73930	0.87310
MAE	0.00873	0.00936	0.00655	0.00470
MSE	0.00013	0.00015	0.00007	0.00003
NMSE	0.43762	0.51888	0.23288	0.11250

List of Figures

1	Schematic of some previously developed approaches to predict soil corrosivity potential.	34
2	Data fusion framework for pipe performance assessment.	35
3	Scatterplot matrix for the soil property data.	36
4	Histogram and box plot of soil property data: soil resistivity (Unit: Ωcm) (a)(b); soil pH value (c)(d); percentage (%) of clay fines (e)(f); and redox potential (Unit: mV) (g)(h).	37
5	Histogram plot of sulfide.	38
6	Soil properties against pipe deterioration rate (DR).	39
7	Determining pitting corrosion rate (D) in metallic pipe.	40
8	Typical structure of a regression tree.	41
9	(a) Q-Q plot of random forest prediction error; (b) the bar plot showing the importance of each soil property.	42
10	Predictions with single predictor.	43
11	Regression tree to predict corrosion rate (Unit: mm/year).	44
12	Predictions with multiple predictors.	45

Fig. 1. Schematic of some previously developed approaches to predict soil corrosivity potential.

Fig. 2. Data fusion framework for pipe performance assessment.

Fig. 3. Scatterplot matrix for the soil property data.

(a) Histogram plot.

(b) Box plot.

(c) Histogram plot.

(d) Box plot.

(e) Histogram plot.

(f) Box plot.

(g) Histogram plot.

(h) Box plot.

Fig. 4. Histogram and box plot of soil property data: soil resistivity (Unit: Ωcm) (a)(b); soil pH value (c)(d); percentage (%) of clay fines (e)(f); and redox potential (Unit: mV) (g)(h).

Fig. 5. Histogram plot of sulfide.

Fig. 6. Soil properties against pipe deterioration rate (DR).

Fig. 7. Determining pitting corrosion rate (D) in metallic pipe.

Fig. 8. Typical structure of a regression tree.

Fig. 9. (a) Q-Q plot of random forest prediction error; (b) the bar plot showing the importance of each soil property.

Fig. 10. Predictions with single predictor.

Fig. 11. Regression tree to predict corrosion rate (Unit: mm/year).

Fig. 12. Predictions with multiple predictors.