Cui and Wang *J Infrastruct Preserv Resil* (2023) 4:14
https://doi.org/10.1186/s43065-023-00081-w

Journal of Infrastructure Preservation and Resilience

RESEARCH Open Access

Analysis and prediction of pipeline corrosion defects based on data analytics of in-line inspection

Bingyan Cui[1] and Hao Wang[1*]

Abstract

In-line inspection (ILI) is important to pipeline integrity management since it can detect pipeline defects and identify potential failure locations through periodical examinations. However, effectively evaluating defects based on ILI data is challenging. Measurements of ILI are easily influenced by instrument performance and maintenance activities, leading to unmatched and imbalanced data. Poor ILI data make it difficult to establish defect growth models based on multiple inspections. This study conducted comprehensive analysis of ILI data for evaluating corrosion defects of a steel pipeline. First, statistical analysis was performed on raw data to visualize distributions of corrosion depths and number of corrosions. Second, hierarchical clustering method was used to classify corrosion severity levels based on features of corrosion depth and estimated repair factor. The interaction effect between adjacent corrosions was considered. Machine learning methods, including k-nearest neighbor, support vector machine, random forest, and light gradient boosting machine were used to explore the relationship between the location parameters of adjacent corrosions and severity levels. Then, maximum corrosion depths and corrosion density were filtered from raw ILI data of multiple inspections, which were critical for pipeline failure prediction. Finally, distribution parameters were fitted to establish stochastic growth models on maximum corrosion depth and corrosion number density. This study presents data analytics based approach to obtain valid information from ILI data in practice.

Keywords Steel pipeline, In-line inspection, Corrosion, Interacting effect, Machine learning, Stochastic growth model

Introduction

Pipelines play a significant role in transporting substantial amounts of oil and gas commodities across long distances. Steel pipes may suffer from different types of defects, including corrosion, cracking, and mechanical damage. If these defects are not properly monitored and repaired, it may cause public safety issues and economic losses. Pipeline integrity management has been developed to keep pipelines in safe operating conditions. It is a program that coordinates procedures, instruments, and tasks for evaluating the condition of pipelines. It can help schedule inspection and maintenance work to lower failure risk [27]. Generally, it includes three main components: defects detection and identification, defect growth prediction, and risk-based management.

Non-destructive evaluation methods such are widely used for in-line inspection (ILI) to locate and identify anomalies on pipelines. Magnetic flux leakage (MFL) and ultrasonic tools are common ILI techniques used for corrosion inspection of steel pipes. Different ILI tools show different capabilities to identify corrosion features. Some ILI tools can identify corrosion features with unique geometries including corrosion pits, axial grooving, and general corrosion better than others. Generally, ILI tools have average accuracy within ± 10% of pipe wall thickness

*Correspondence:
Hao Wang
hwang.cee@rutgers.edu
[1] Department of Civil and Environmental Engineering, School of Engineering, Rutgers, The State University of New Jersey, New Brunswick, USA

[26]. To predict defect growth and time to failure, ILI need to be performed periodically. Defects from at least two inspections should be matched to their positions in the pipeline. However, each ILI uses its own coordinate system to locate detected corrosions in the pipeline [22]. As a result, these inconsistent coordinate systems would lead to unmatched data from multiple ILI runs. In addition, the accuracy of ILI tools is greatly influenced by instruments error and environmental conditions [4]. Changes in technologies and maintenance activities make it difficult to obtain consistent ILI data from multiple years.

Corrosion is one of the most important defects that affects the pipeline integrity directly. There are large quantities of ILI data on corrosion features. Therefore, extracting useful information from corrosion ILI data is important. Corrosion defects on pipeline can be divided into single defect and interacting multiple defects [19]. Compared to single defect, analysis of interaction between multiple corrosion defects was more complex. Chiodo and Ruggieri [7] found that interactions between adjacent defects would influence the failure pressure of pipeline significantly. Similarly, it was reported that the failure pressure of pipeline decreased significantly due to interaction effects between adjacent corrosion defects [6, 14, 24, 25]. Therefore, the assessment of interacting corrosion defects is desired from ILI data.

In addition to interacting effects that need to be considered, establishing appropriate growth model is also important for pipeline integrity management. The reliable prediction of defect growth can help schedule future inspection and maintenance activities to prevent potential pipeline failures in the future. There are two categories of defect growth models: the model-based approach and the data-driven approach. Model-based approaches primarily rely on physical models, such as finite element models, to predict defect growth. Liu et al. [15] employed Bayesian networks to update the likelihood of subsea pipeline damage and estimated the ultimate probability of damage. Based on this probability, they were also able to predict the remaining useful life of the pipeline. The data-driven approach is to use ILI data or sample data to investigate the propagation of defects. F. Caleyo et al. [5] used the Markov chain to estimate the time-dependent growth rate of pipelines. Arzaghi et al. [1] used Dynamic Bayesian network (DBN) to predict varying growth rates of pitting and corrosion degradation in subsea pipelines. Instead of calculating corrosion growth rates, Mohd et al. [18] used Weibull distribution to develop a time-dependent corrosion depth model that can predict the peak depth of pipeline at any given age. Similarly, Gumbel distribution was adopted to predict the growth of block maximum corrosion depth [13]. Further to this study, the peaks over threshold (POT) method was also used to improve the evaluation performance of extreme values [28]. Therefore, it is applicable to use different distribution parameters to establish stochastic growth models of different corrosion features.

In summary, how to process and analyze existing ILI data from multiple years is of great significance. Complex corrosion features may be unmatched on both spatial and temporal scales. Therefore, this study aimed to propose a comprehensive procedure to analyze both raw and filtered ILI data. Firstly, distributions of corrosion number and corrosion depth were visualized to provide preliminary evaluation. Then, interacting effects of adjacent corrosions were considered to find the relationship between defect locations and defect severities. Finally, stochastic growth models were established to predict the evolution of maximum corrosion depth and corrosion number density.

Data collection

The ILI dataset was obtained from Magnetic Flux Leakage (MFL) tools in 2005, 2012 and 2016, respectively. A 12-mile steel pipeline which was originally built in 1974 was inspected. Based on the history of replacements and relocations, the pipeline was divided into several segments (a-g), as listed in Table 1.

Table 1 General information about the pipeline

Line segment No	Length (feet)	Outer diameter (in.)	Wall thickness (in.)	Pipe grade	Year installed
a	51,241	30	0.562	5L×42	1974
b	613	24	0.438	5L×42	1974
c	772	20	0.375	5L×42	1982
d	5,910	30	0.562	5L×60	2005
e	1,698	30	0.562	5L×60	2005
f	41	6.625	0.28	5L×42	1974
g	654	30	0.562	5L×42	2002

In this study, external corrosion defects were selected for analysis since it is the major defect observed in this pipeline. The pipeline consisted of 1,955 girth welds in total. Corrosions did not occur in every segment of all girth weld numbers. Therefore, only girth weld number with corrosion defects were extracted from the ILI dataset. From 2005 to 2016, about 400 girth weld numbers showed external corrosions. Table 2 displays the number of corrosion defects present in each girth weld location, with some locations having multiple defects. Details of ILI dataset included girth weld number, absolute distance, peak depth, length and orientation.

Analysis methodology

Clustering

The objective of clustering is to divide observations into several clusters so that data points within the same cluster are similar to each other. In this study, hierarchical clustering method was used to separate corrosion defects with similar features. Corrosion severity levels have a hierarchical structure, as most features of defects in high level would be severer than low level. Therefore, hierarchical clustering is suitable for the classification of corrosion severity level.

Hierarchical clustering includes divisive and agglomerative algorithms. The divisive algorithm is a top-down approach. At the beginning, all the observations belong to one cluster. Then, different observations will be divided into more clusters according to the certain criterion such as distance. On the contrary, the agglomerative algorithm is a bottom-up approach. Each observation is a cluster at first. Then, similar observations will be merged to fewer clusters.

In this study, the agglomerative algorithm was used. It can determine the similarity between observations of each cluster by measuring the distance between them. Smaller distance indicates higher similarity. Therefore, the clustering algorithm merges the two clusters with the shortest distance between them to construct the clustering tree. Measurements of distance between clusters can be conducted through different methods, such as single, complete, centroid, average and ward linkages.

Table 2 Number of external corrosion defects found in different inspection years

Inspection year	Defects count no
2005	792
2012	1345
2016	2508

Single linkage clustering calculates the distance between two clusters as the shortest distance between any two data points in each cluster. In contrast, complete linkage clustering uses the maximum distance between any two data points in each cluster. Average linkage clustering calculates the average distance between all pairs of data points in each cluster. Centroid linkage clustering calculates the distance between the centroids of each cluster. These linkage methods may be sensitive to anomalous data points and easy to generate unreasonable clustering. However, data points of corrosion defects have many outliers. Therefore, ward linkage was used in this study. Ward linkage can minimize the loss of combining clusters each time. It calculates the error sum of squares (ESS) of each cluster. Small ESS value means agglomerative data points. Therefore, clusters can be combined to fewer clusters by minimizing the increase of ESS.

Classification

Machine learning methods

To find relationship between defect location parameters and severity levels, different machine learning methods were used, including k-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), and light gradient boosting machine (LightGBM).

KNN is a supervised learning method proposed by Fix and Hodges [11]. In classification, an unlabeled data point will be assigned to the label that is most commonly found among the k-nearest training data points from the target data point. Therefore, the select of k value and measurement of distance are important for KNN.

SVM is initially a binary classification approach which is aimed to construct an optimal separation hyperplane [17]. The hyperplane has the maximum distance from the nearest sample points (called support vector) on both sides. Therefore, SVM can balance the learning ability and the complexity of the model. By means of kernel functions, SVM is capable of mapping data from a low-dimensional space to a higher-dimensional space. There are three commonly used kernel functions, including the linear kernel, polynomial kernel and radial basis function (RBF) kernel [20].

RF was proposed to solve classification, clustering, and prediction problems. It is a decision tree based machine learning algorithm evolved from the bagging ensemble learning. Firstly, a decision tree consisting of multiple independent forests is randomly generated. Then, features are selected by calculating the information gain. From the root node, the tree is split according to the feature partitioning condition and the principle of minimum node purity until the rule is satisfied. Usually, information entropy is used to measure the purity of data [3]. Different from the single decision tree method, Random Forest

randomly selects m subsamples from the original dataset with put-back. And then it will train a single decision tree with k randomly-selected features. The optimal features are chosen from these k features to split the nodes. After that, t decision tree can be constructed by repeating above process t times. The final prediction result is a weighted average of each decision tree.

LightGBM is a boosting tree algorithm in the ensemble learning [12]. It utilizes a leaf-wise approach to select the best split, allowing it to identify the leaf node with the highest split gain out of all the leaf nodes in the decision tree. LightGBM optimizes training data points based on the gradient of each data point. Data point with larger gradient means larger contributions to the information gain. The algorithm employs a histogram-based method to convert continuous feature values into k integers, thereby allowing for the creation of a histogram with a width of k. Subsequently, the algorithm will iterate through the training data to compute the cumulative statistics for each discrete value present in the histogram. In this case, only discrete values of the sorted histogram are required to be traversed when choosing the splitting point of feature. Therefore, LightGBM can decrease the computation cost significantly.

Evaluation metrics

For binary classification, accuracy, precision, recall and F1 score are usually used to evaluate model performance. Accuracy, as defined by Baldi et al. [2], is the proportion of correctly classified samples in the testing dataset out of all the samples. Precision, on the other hand, is the percentage of true positive samples among all the predicted positive samples. Recall is the percentage of truly predicted positive samples out of all truly positive samples. F1 score is a balanced score that combine precision and recall. These metrics can be calculated as shown in Eq. (1) to (4) [21].

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

where, TP represents number of positive samples correctly predicted as positive; TN represents number of negative samples correctly predicted as negative; FP is number of negative samples incorrectly predicted as positive; FN is number of positive samples incorrectly predicted as negative.

For multi-classification, it can be regarded as multiple binary classifications. Therefore, average value of them can be used to evaluate the model performance. In this study, weighted F1 score was calculated, because it takes into account the importance of different categories [16].

Defects growth predictions

Data preprocessing

In the ILI dataset, not all inspection locations had external corrosions. Normal points and manufactural bend were also common. Therefore, data points of external corrosions were filtered first. After that, the segments with replacement recordings were eliminated because it will influence the defect growth.

However, to establish growth models, the filtered data still needed to be organized according to certain rules. For the growth model of corrosion depth, the maximum peak depth in each segment was selected, as maximum corrosion depth is one of the most important factors leading to pipeline failure. Then, only data showing continuous increase in maximum depth over inspection years were filtered. This approach yields a more conservative data subset, which will be used to analyze the growth of maximum corrosion depth in further analysis. For the growth model of corrosion density, data points that deviated from the mean value by more than 3 times the standard deviation were removed. Except for these outliers, all data points were used for growth prediction of corrosion density.

Distribution models and parameters were used to predict future corrosions because these distributions can capture the trend of corrosion based on previous ILI data. Corrosion growth process is complicated so that using stochastic growth models instead of simplified growth rate may be better.

Gumbel distribution

Gumbel distribution is particular useful in fitting the distribution of extreme values. Since maximum corrosion depth is the extreme value, Gumbel distribution was selected to fit corrosion depth data. Gumbel distribution is derived from the extreme value theory that developed by Fisher and Tippett [10]. The probability distribution function of the maximum value for each sample converges to the generalized extreme value (GEV) distribution. Gumbel distribution is a special form of GEV distribution, as expressed in Eq. (5) [13].

$$G_t(z) = e^{-e^{\frac{z-\mu(t)}{\sigma(t)}}} \tag{5}$$

where, $G_t(z)$ is the density when the maximum corrosion depth is equal to z; and z is the maximum corrosion

depth in this study; μ is the location parameter; σ is the scale parameter; and t is the inspection year.

Weibull distribution

Weibull distribution is a non-stationary distribution that follows Cole's method [9]. It is usually used to model the reliability. Weibull distributions can model right-skewed data, left-skewed data, or symmetric data [23]. In this study, corrosion number density is an index that reflects the number of defect per unit distance. In different segment, the number density has a large difference. Corrosion number density below 5 was the most, leading to left-skewed ILI data. In this case, Weibull distribution

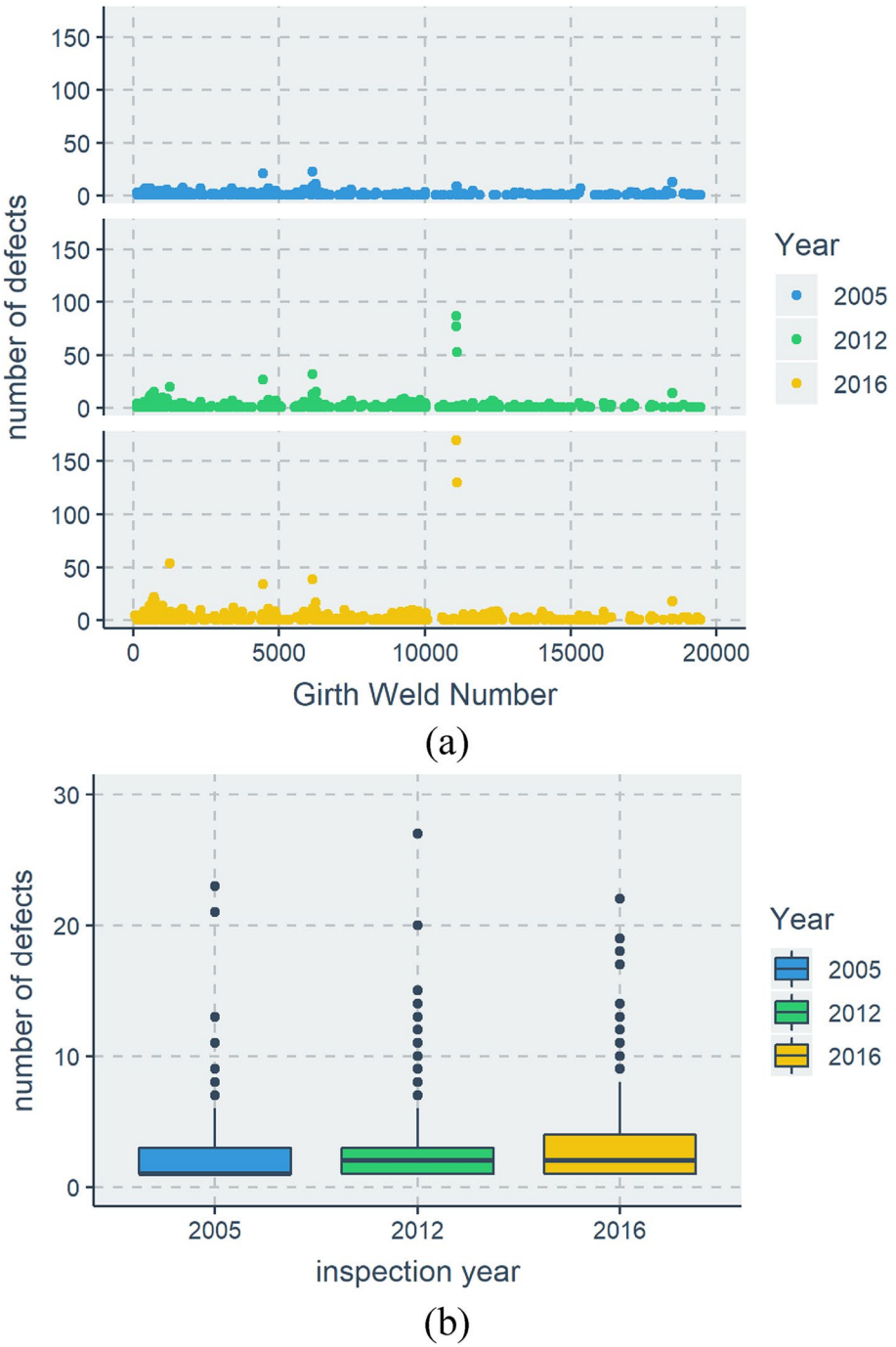

Fig. 1 Number of defects along the pipeline from 2005 to 2016 (**a**) scatter plot, (**b**) boxplot

can be of great help. The expression of Weibull distribution is shown in Eq. (6) [13].

$$W_t(x) = \frac{\xi(t)}{\sigma(t)}\left[\frac{x}{\sigma(t)}\right]^{\xi(t)-1} \times e^{-\left[\frac{x}{\sigma(t)}\right]^{\xi(t)}}, \quad x \geq 0 \quad (6)$$

where, $W_t(x)$ is the density when the corrosion number density is equal to x; and x is the corrosion number density; ξ is the shape parameter; σ is the scale parameter; and t is the inspection year.

Analysis results and discussion

Statistical analysis of corrosion depths and locations

To compare the distribution of corrosion defects, the number of defects in each girth weld number along the pipeline was counted, as shown in Fig. 1. Each girth weld

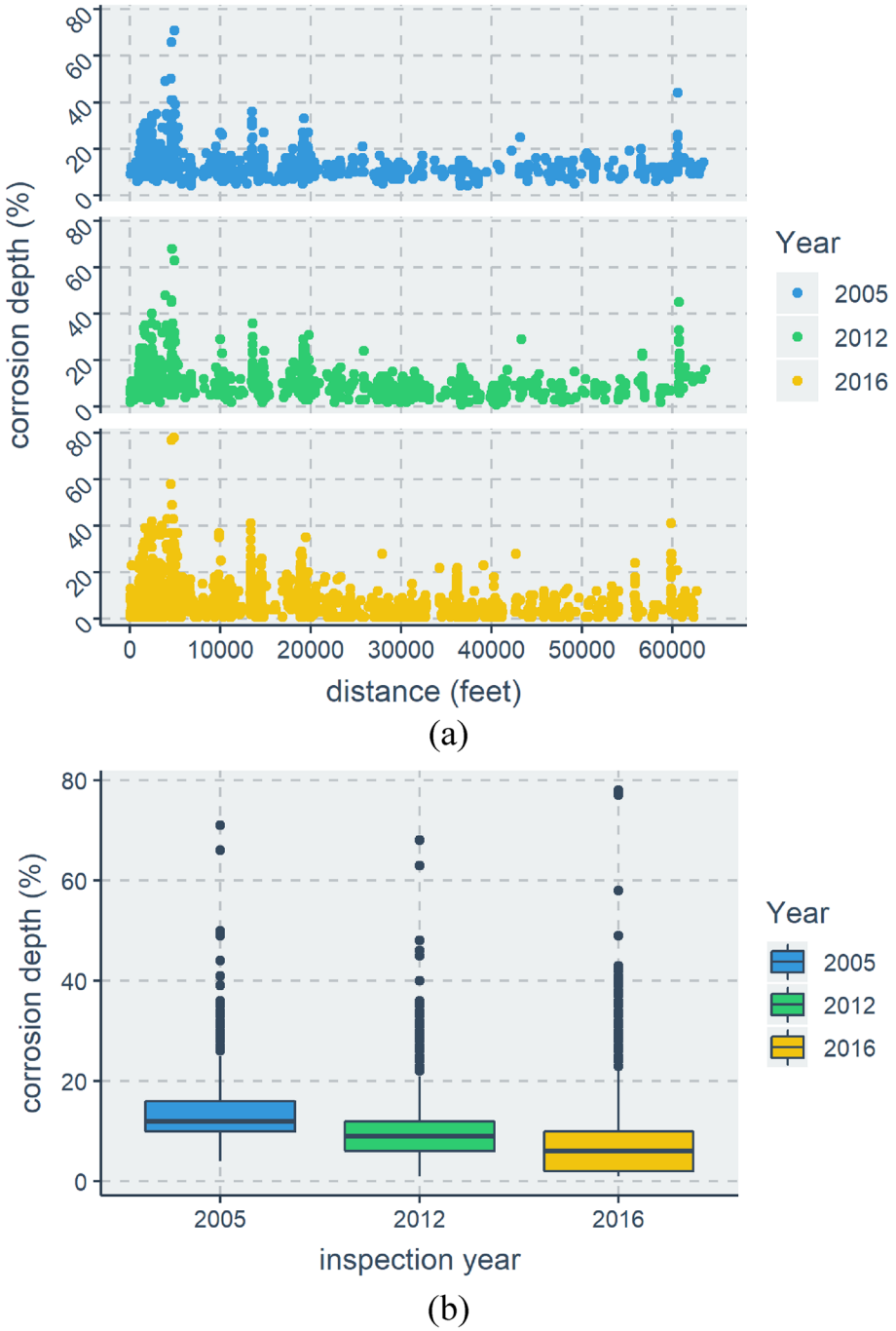

Fig. 2 Corrosion depth along the pipeline from 2005 to 2016 (**a**) scatter plot, (**b**) boxplot

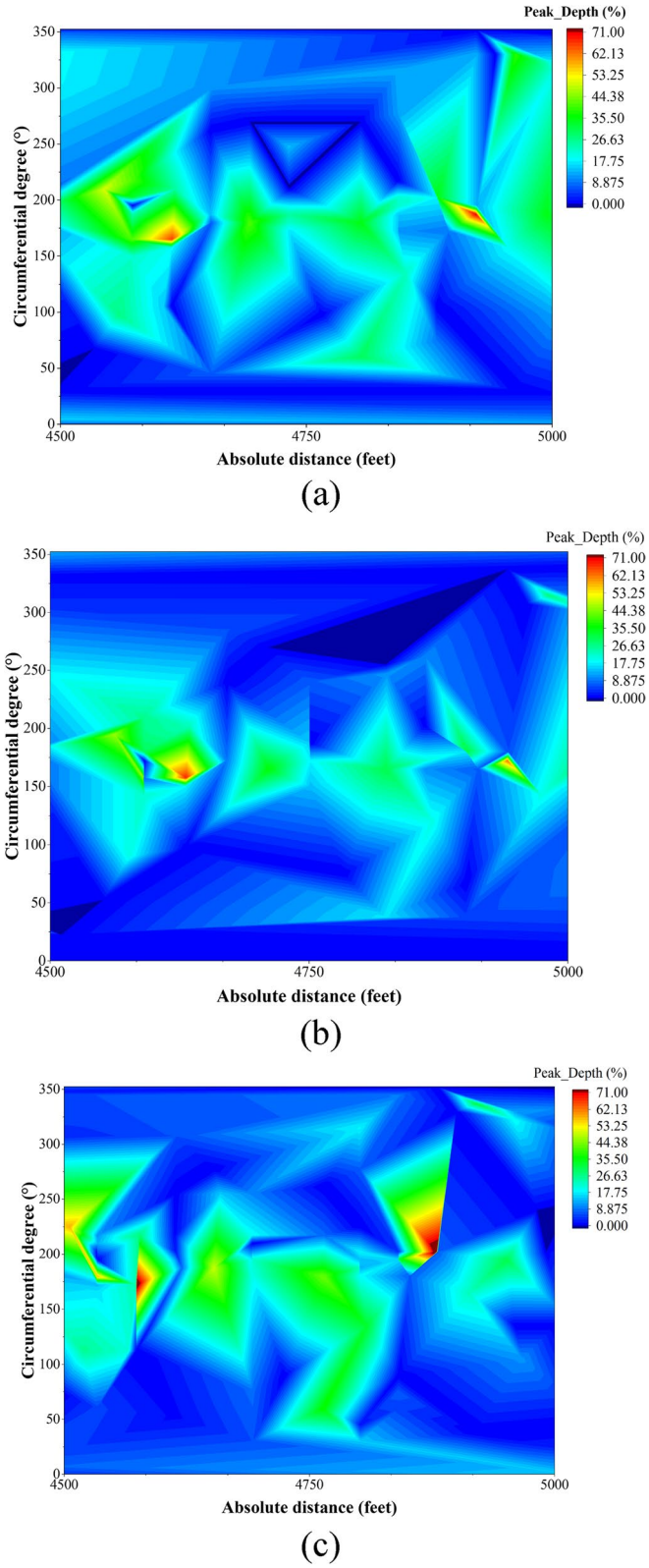

Fig. 3 2D contour plot of peak depth in (**a**) 2005, (**b**) 2012, (**c**) 2016

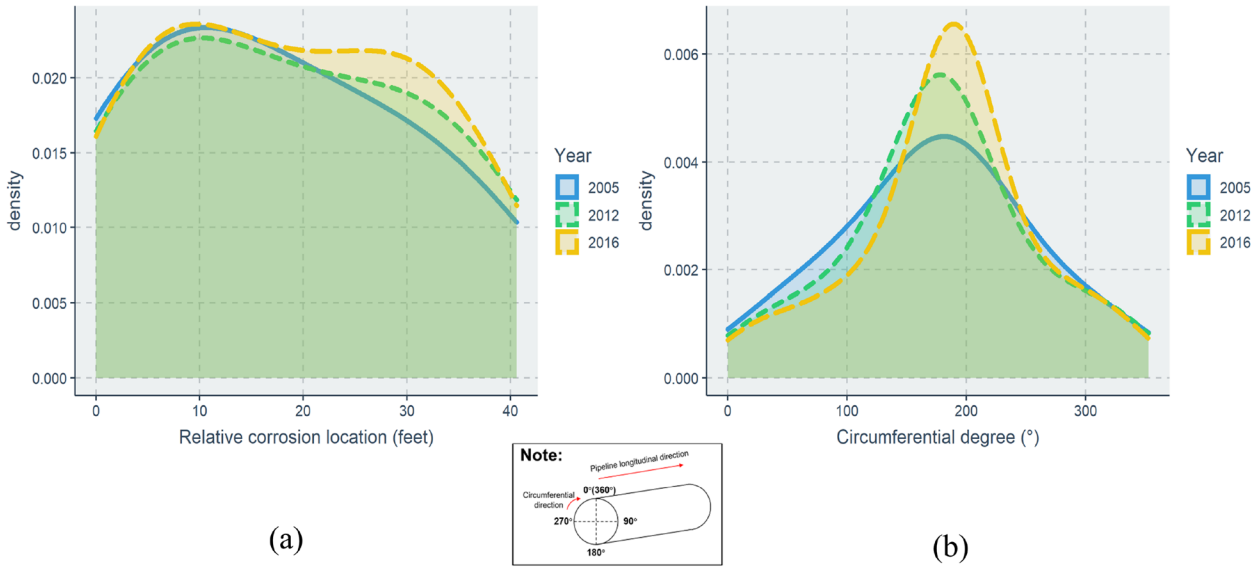

Fig. 4 Density plots of (**a**) longitudinal locations of corrosion defects; (**b**) circumferential locations of corrosion defects

represents 30–40 feet pipe length. It can be seen that the average number of defects increased from 2005 to 2016, which is consistent with the change in total number of defects. In addition, the increase of corrosion defects around several girth weld numbers was found more significant. For example, the number of defects in segments around 11,080 girth weld number was 9 in 2005. However, it increased to 77 and 170 in 2012 and 2016, respectively, indicating the soil environment in these segments for high corrosion potential. However, the soil survey data were not available.

The comparison of corrosion depth was based on peak depth in each girth weld number. The peak depth is defined as the maximum depth of the corrosion divided

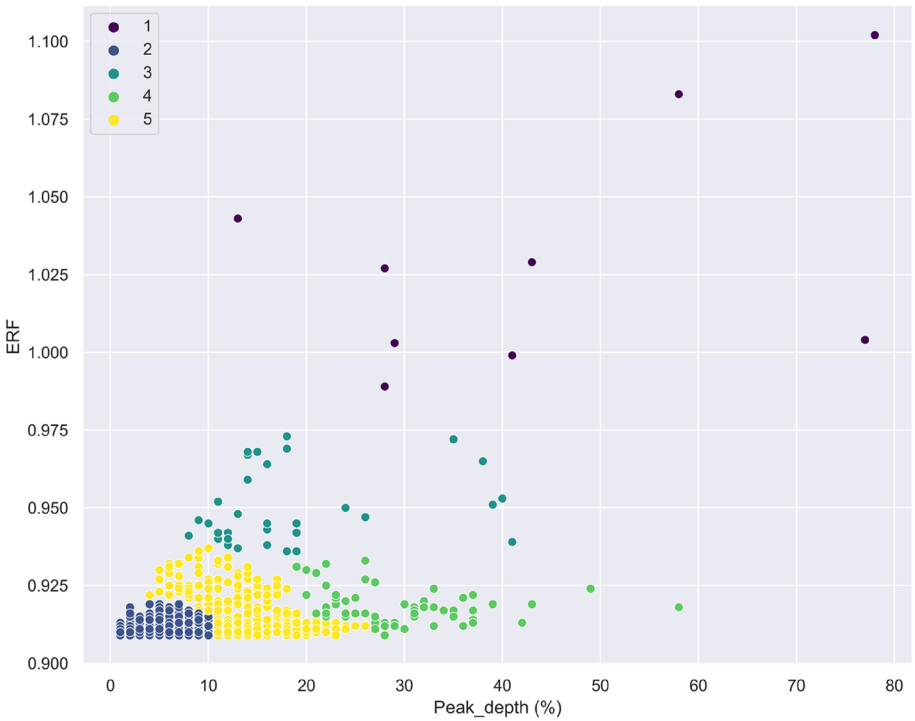

Fig. 5 Hierarchical clustering of corrosion defects based on peak depth and ERF

Table 3 Classification results of corrosion severity levels

Severity level	Average depth (%)	Average ERF	Average length (in.)	Average width (in.)
Low	4.34	0.910	1.4	1.7
Medium	14.05	0.913	1.4	1.9
High	27.31	0.937	3.3	4.8

by the wall thickness at the location of the corrosion. Therefore, the larger peak depth means the severer corrosion condition. The plot of peak depth along the pipeline is shown in Fig. 2. Interestingly, the average corrosion depth was observed to decrease from 2005 to 2016. This is reasonable because there were a lot of small corrosion defects generated in 2012 and 2016, which reduced the average depth. Ideally, the corrosion depth would

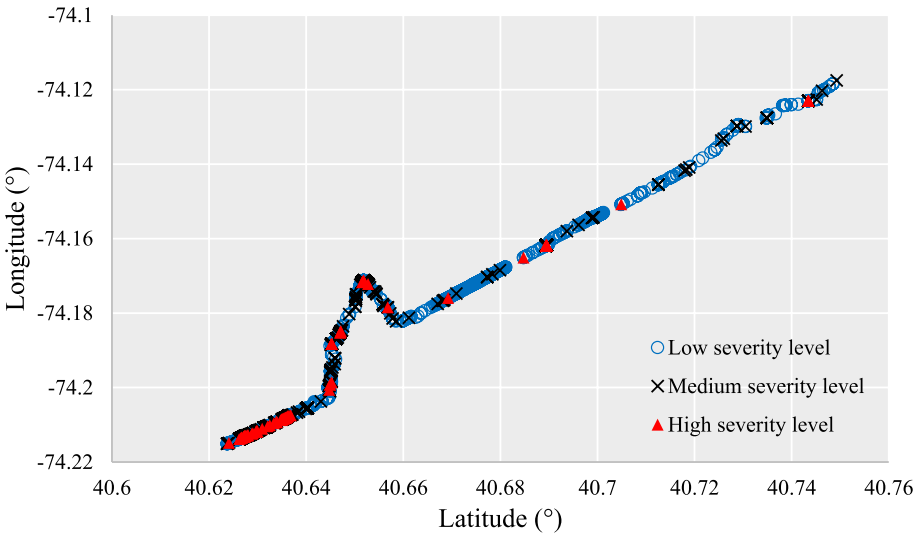

Fig. 6 Geographical distribution of corrosion severity levels

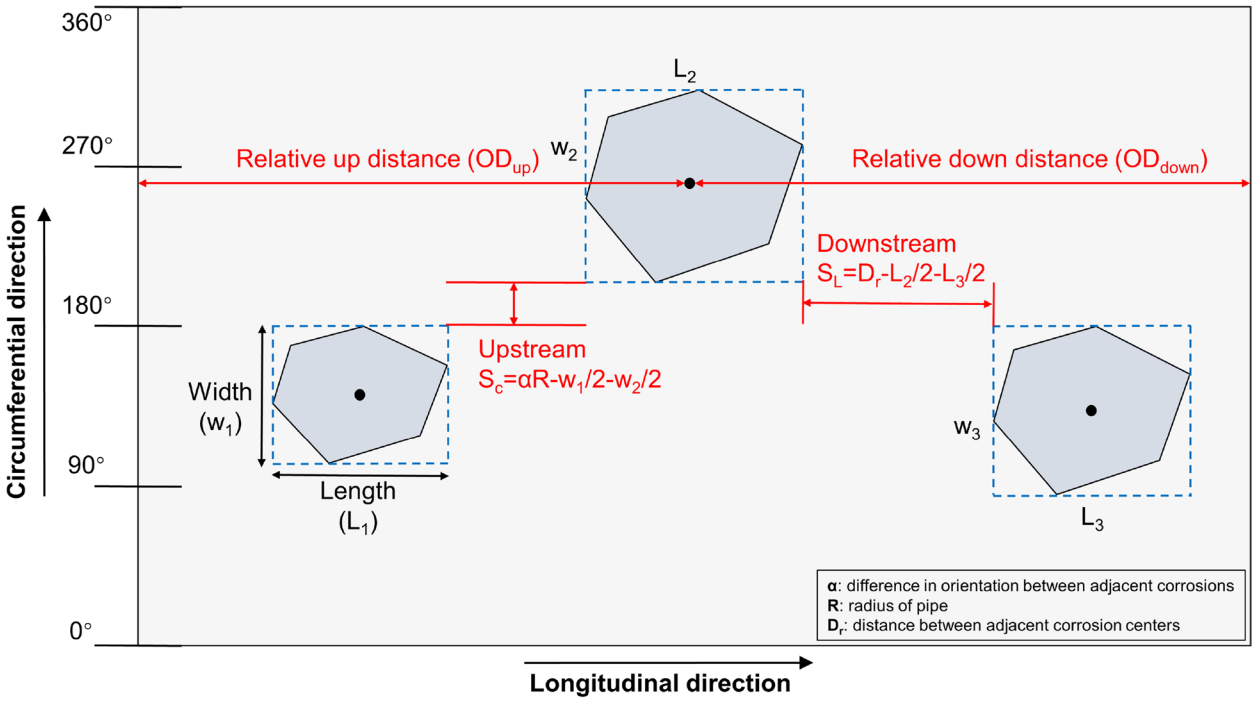

Fig. 7 Illustration of location parameters on a 2D plane for one pipe segment

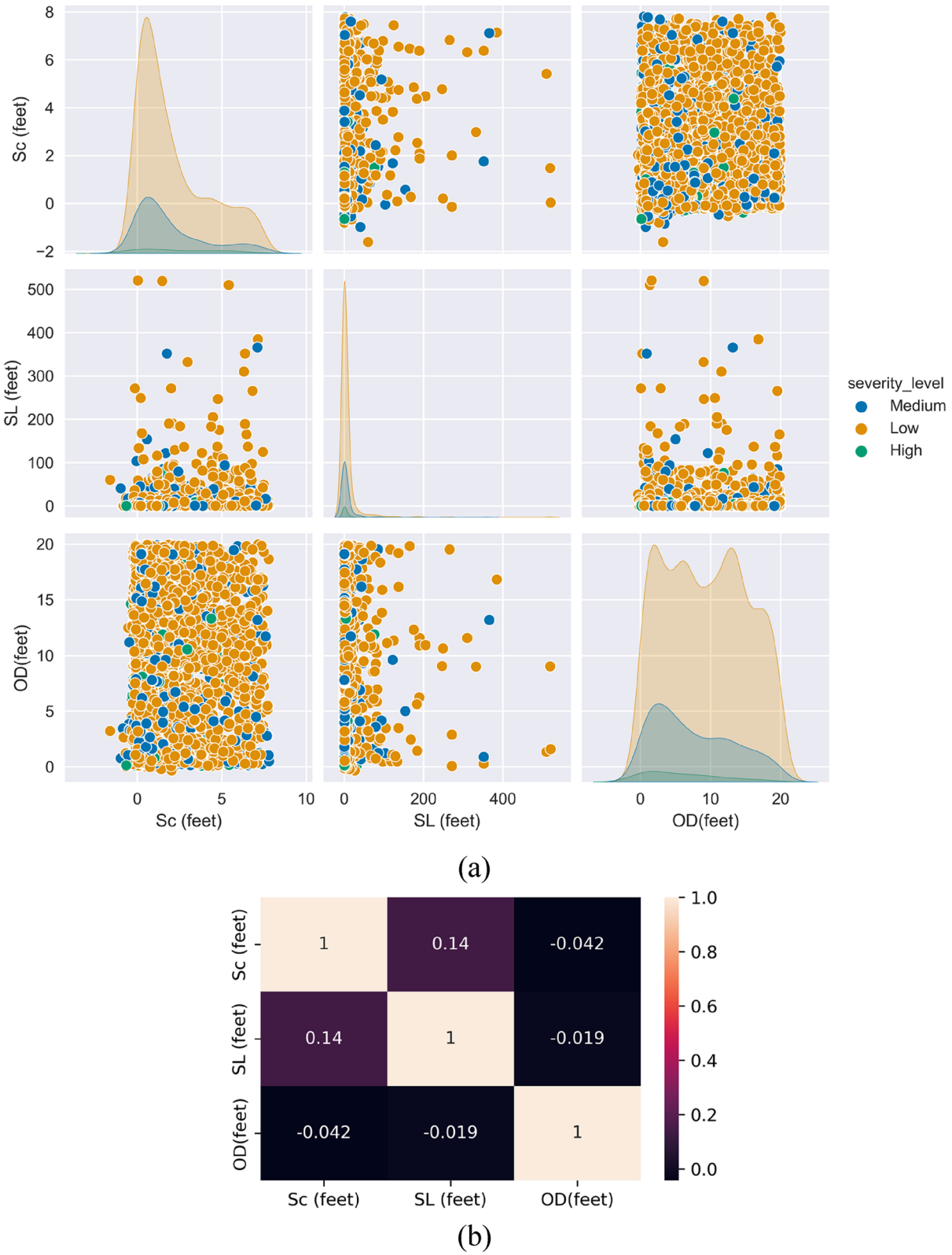

Fig. 8 Correlation plot between three input variables: (**a**) scatter pair plot, (**b**) heat map

increase over years if no repair is placed. However, this trend was not observed at each inspection location. The variations can be caused by the changes in instrument performance of ILI tools and the maintenance or repair activities between different inspections. However, the information of these changes were not available in this study. Therefore, establishing the corrosion depth growth model based on raw ILI data was not suitable.

For the localized segment, corrosion depth presented certain increasing trend. As shown in Fig. 1, the corrosion depths were the most severe around the distance of 4500–5000 feet. Therefore, 2D contours of the peak corrosion depth were plotted in these segments, as shown in Fig. 3. In Fig. 3, x-axis was the absolute distance to the original location; *y*-axis was the orientation degree in the circumferential direction. For example, 0° and 360° represented the top of pipeline, while 180° denoted the bottom of pipeline. It shows that the area of maximum peak depth increased a lot in 2016, compared to 2005. In addition, it was found that maximum peak corrosion depths were located at around 4600 and 4800 feet with circumferential degrees of 150°-200°.

To have better understanding of the corrosion distribution, the density plots of axial and circumferential locations of corrosion defects were shown in Fig. 4. It was found that external corrosions were more likely to occur at 10 and 30 feet relative to the pipeline joint. The circumferential degree was mainly around 180°, indicating external corrosion tended to happen at the bottom of steel pipe.

Interaction of adjacent defects on corrosion severity level

Classification of corrosion severity level

In this section, ILI data in 2016 was used to investigate the relationship between corrosion severity level and defect location parameters. Estimated repair factor (ERF) is the ratio of maximum allowable operating pressure (MAOP) of pipeline to the safe working pressure. Both peak depth and ERF are the significant indicators about corrosion severity level. Higher peak depth and ERF indicate defects that are more dangerous. Therefore, all defects were divided into several clusters through hierarchical clustering method based on defect depth and ERF, as shown in Fig. 5.

To better characterize the corrosion severity level, these clusters needed to be combined to fewer categories. Clustering methods can capture characteristics of data distribution based on distance criteria. However, to obtain reasonable severity levels, empirical methods should also be considered. Therefore, cluster 1, cluster 3 and cluster 4 were combined to represent the highest defect level, because these clusters had the highest value in peak depth or ERF. Similarly, cluster 5 were used to represent medium severity level. Cluster 2 were the low severity level. It should be noted that the low, medium, and high severity levels here are relative in this ILI dataset.

Table 3 shows the classification results of severity level. From the table, it is obvious that the average defect depth, ERF, length and width were the most in high severity level. This is reasonable, as higher values mean higher risk of failure. Therefore, defects at high severity level should be prioritized in the maintenance scheduling. Furthermore, the geographical distribution of three severity levels can be seen in Fig. 6. Defects with high severity level were mainly found in low latitudes, indicating the soil environment in low latitudes may have high corrosion potential.

Relationship between corrosion location parameters and severity level

As stated above, corrosion severity level was classified based on defect depth and ERF. These two indicators are geometric parameters related to defects themselves and do not take into account the interactions between multiple defects. In this study, three location parameters were selected to represent the interacting effect of adjacent defects, including OD, S_c and S_L. OD denotes the relative distance between the centroid of corrosion defect and pipeline girth weld. S_c is the distance between two adjacent corrosion defects in the circumferential direction, while S_L is the distance between two adjacent corrosion defects in the longitudinal direction. Detailed illustrations of these parameters are depicted in Fig. 7.

It should be noted that final values of OD, S_c and S_L were the minimum of upstream and downstream values. This is because the interacting effect of adjacent defects is mainly caused by the nearest ones. After obtaining location parameters, the correlation between these factors should be analyzed first to avoid co-linearity. Figure 8 (a) shows the correlation between each two variables. It can be observed that the scatter data points of them distributed randomly. No obvious linear or nonlinear relationship were found. From Fig. 8 (b), correlation coefficients between each pair were also small, indicating that the co-linearity did not exist in these variables. Therefore, there is no need to reduce the dimensionality of these variables.

Machine learning methods were used to analyze the relationship between location parameters and severity of

Table 4 Performance of different machine learning methods

Metrics	KNN	SVM	RF	LightGBM
Accuracy	72.71%	74.24%	74.83%	74.15%
Weighted F1	0.67	0.69	0.71	0.69

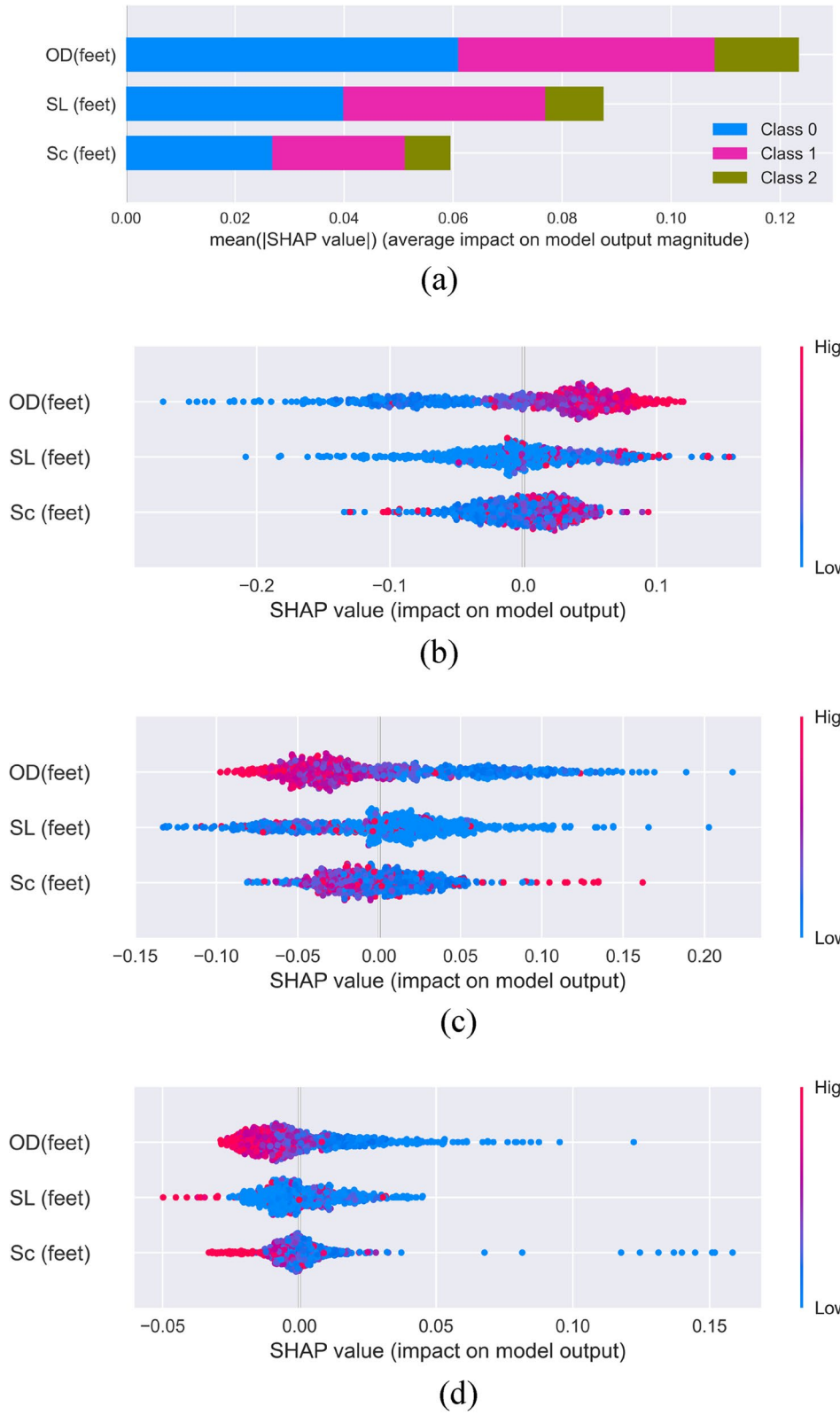

Fig. 9 (**a**) importance of input variables on three severity levels of corrosion defects; and impact of input variables on (**b**) low; (**c**) medium; (**d**) high severity level

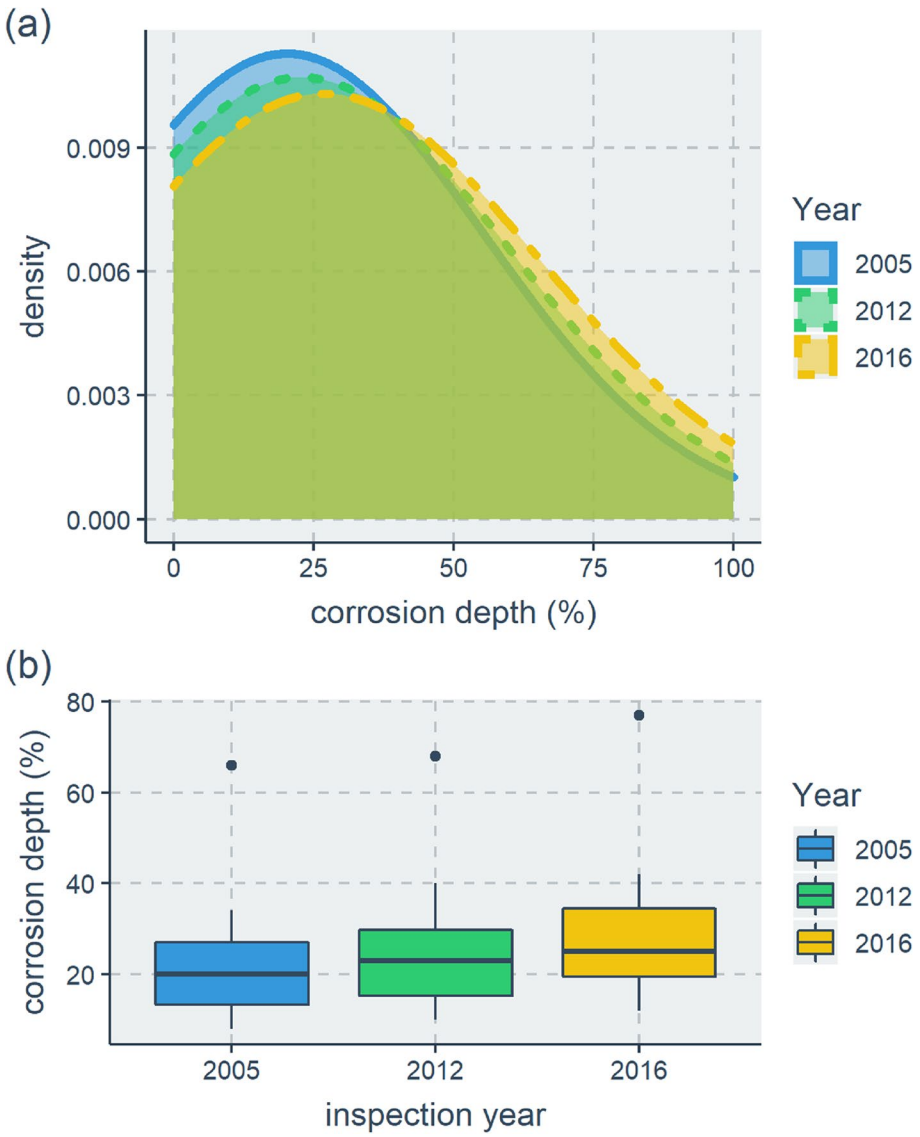

Fig. 10 Plot of maximum corrosion depth (**a**) density plot, (**b**) boxplot

corrosion. Taking OD, S_c and S_L as the input variables and three severity levels as responses, the fitting results using four different machine learning methods (KNN, SVM, RF, LightGBM) were listed in Table 4. As can be seen, random forest shows the best performance among all methods.

The importance of three location parameters were further analyzed using random forest model. Shapley Additive Explanation (SHAP) was used to interpret the classification results. It is a method derived from coalitional game theory [8]. Initially, SHAP value is developed to evaluate the contributions from each player to the game. In the model interpretation, the prediction made by a model can be explained as the sum of the contribution or attribution values of each input variable used in the model. Therefore, the impact value of each feature can be calculated as SHAP value. A higher SHAP value indicates a more important feature.

In Fig. 9 (a), class 0 represented the low severity level, class 1 represented the medium severity level, class 2 represented the high severity level. It can be seen that OD had the most significant impact on the classification, followed by S_L and S_c. Furthermore, positive and negative correlations between location parameters and severity level can be interpreted. As shown in Fig. 9 (b), high feature values of OD, S_L and S_c mainly distributed in regions greater than 0. That means greater value of OD, S_L and S_c can make more defects belong to low severity level. Similarly, in Fig. 9 (d), high feature values of OD, S_L and S_c mainly distributed in regions smaller than 0, which means smaller value of OD,

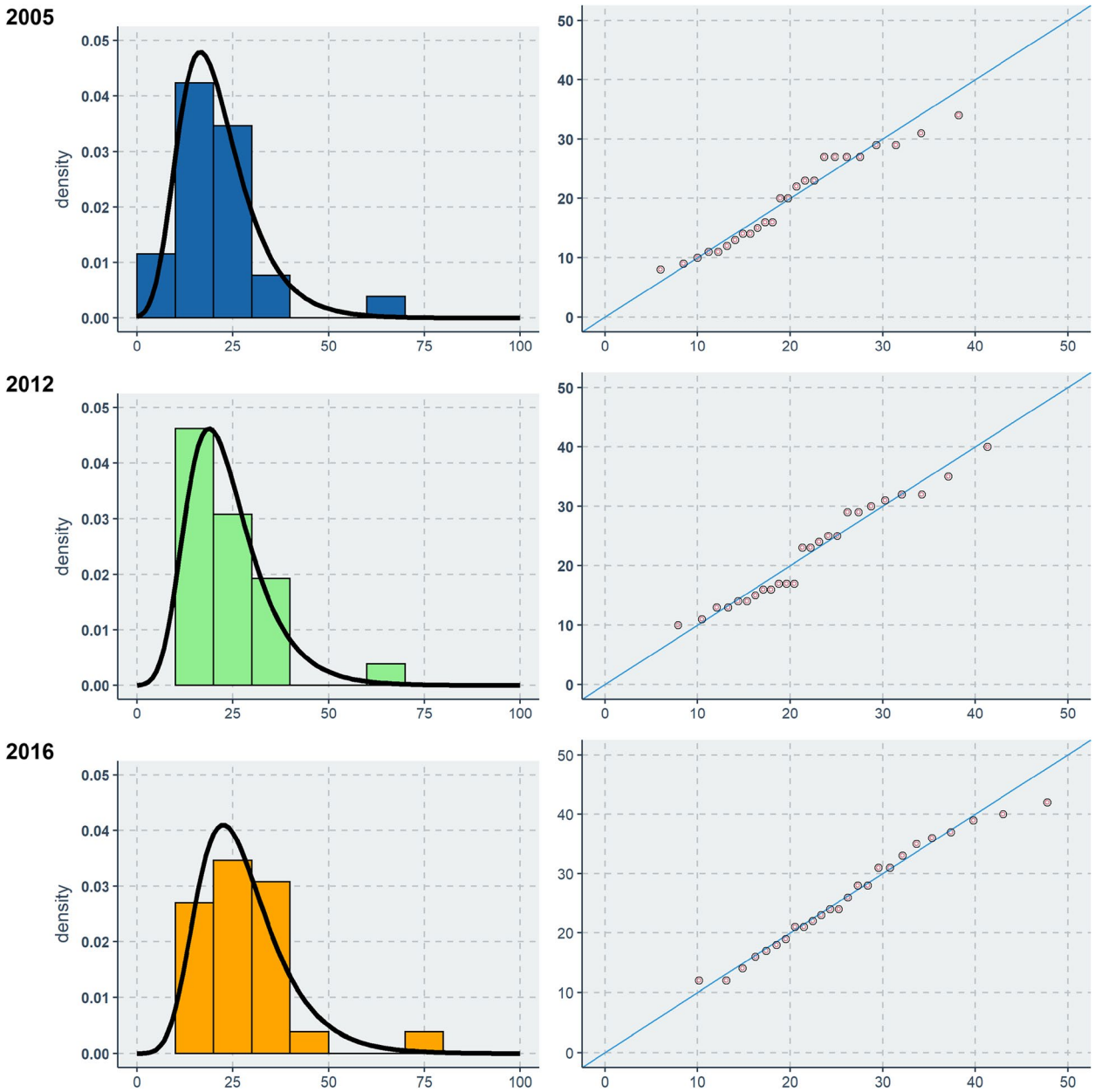

Fig. 11 Fitting of Gumbel Distribution in different inspection years (**a**) histogram plot, (**b**) Q-Q plot

Table 5 Fitting parameters of Gumbel distributions

Inspection year	Location parameter (μ)	Scale parameter (σ)
2005	16.53	7.68
2012	18.83	7.96
2016	22.47	8.97

S_L and S_c can make more defects belong to high severity level. When considering the location parameters OD, S_L, and S_c, smaller values of these parameters indicate higher potential for more critical corrosion defects. A smaller value of OD implies that the corrosion defect is located closer to the pipeline joint, increasing the likelihood of high severity. Similarly, smaller values of S_L and S_c indicate that the corrosion defects are located closer together in the longitudinal and circumferential directions, respectively, which can lead to higher potential for interaction and

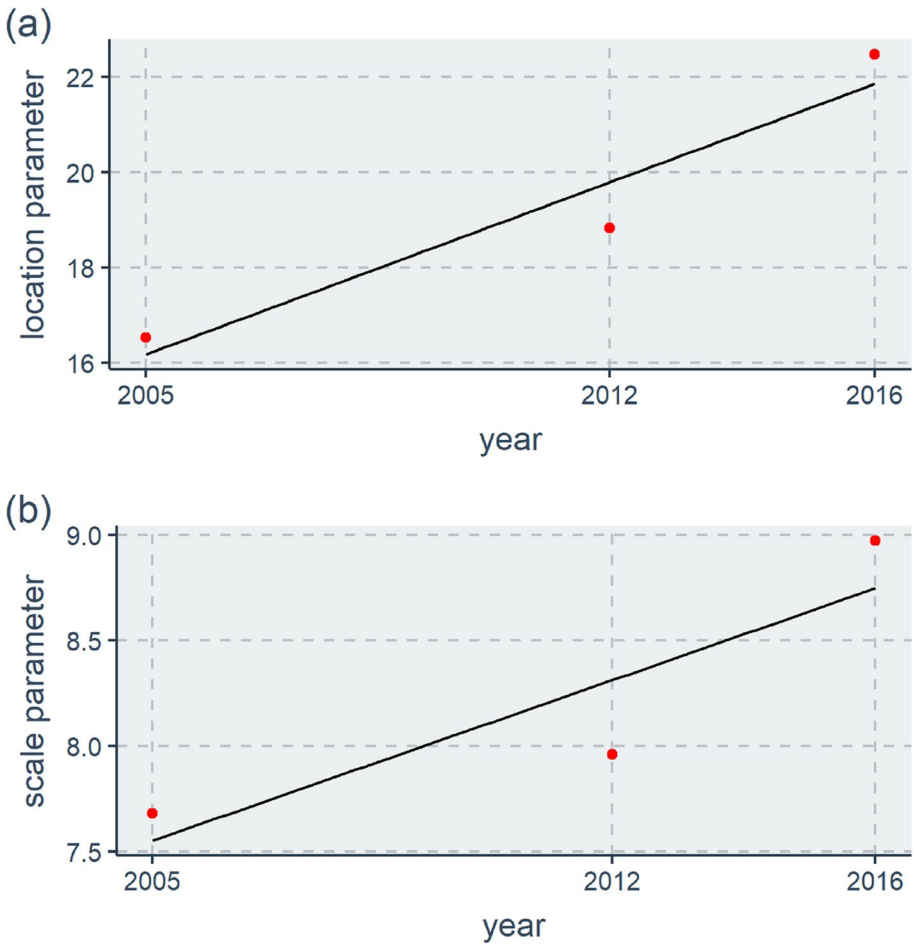

Fig. 12 Fitted lines of Gumbel distribution parameters (**a**) location parameter; (**b**) scale parameter with respect to inspection year

combined effect. Therefore, smaller values of these location parameters can cause more severe corrosion defects.

Stochastic growth models

Maximum corrosion depth

To predict the growth of maximum corrosion depth, the raw dataset was processed to obtain the reasonable subsets for further analysis. Relative distance to the girth weld number was used to locate the defect location; For the defects at the close locations, only the data that shows the continuous growth trend of maximum corrosion depth over inspection years were selected. That means, if the maximum corrosion depth keeps growing, it can be considered that this location was most susceptible to external corrosion. This approach resulted in a smaller and more conservative data subset, which was used to analyze the growth of maximum corrosion depth. The density and box plot of the extracted data subset are shown in Fig. 10. It shows that the maximum corrosion depth increases over time and can be used for growth prediction.

Considering that the Gumbel distribution is particular useful in representing the probability distribution of the maximum value in a sample, the corrosion depths in the subset were fitted to the Gumbel distribution. The theoretical and empirical quantiles were compared through the histogram and Q-Q plots as shown in Fig. 11. The data points are close between theoretical and empirical quantiles, indicating the fitted Gumbel distribution has high accuracy. The fitting parameters are shown in Table 5.

Using the linear regression to fit the Gumbel distribution parameters over the inspection year, the fitted line can be seen in Fig. 12. It can be found that the location and scale parameters have an increasing trend that indicates the growth of maximum corrosion depths. The linear model can be expressed in Eq. (7) and (8). After obtaining the two parameters, the Gumbel distribution can be used to calculate the density of maximum corrosion depth at the year of interest.

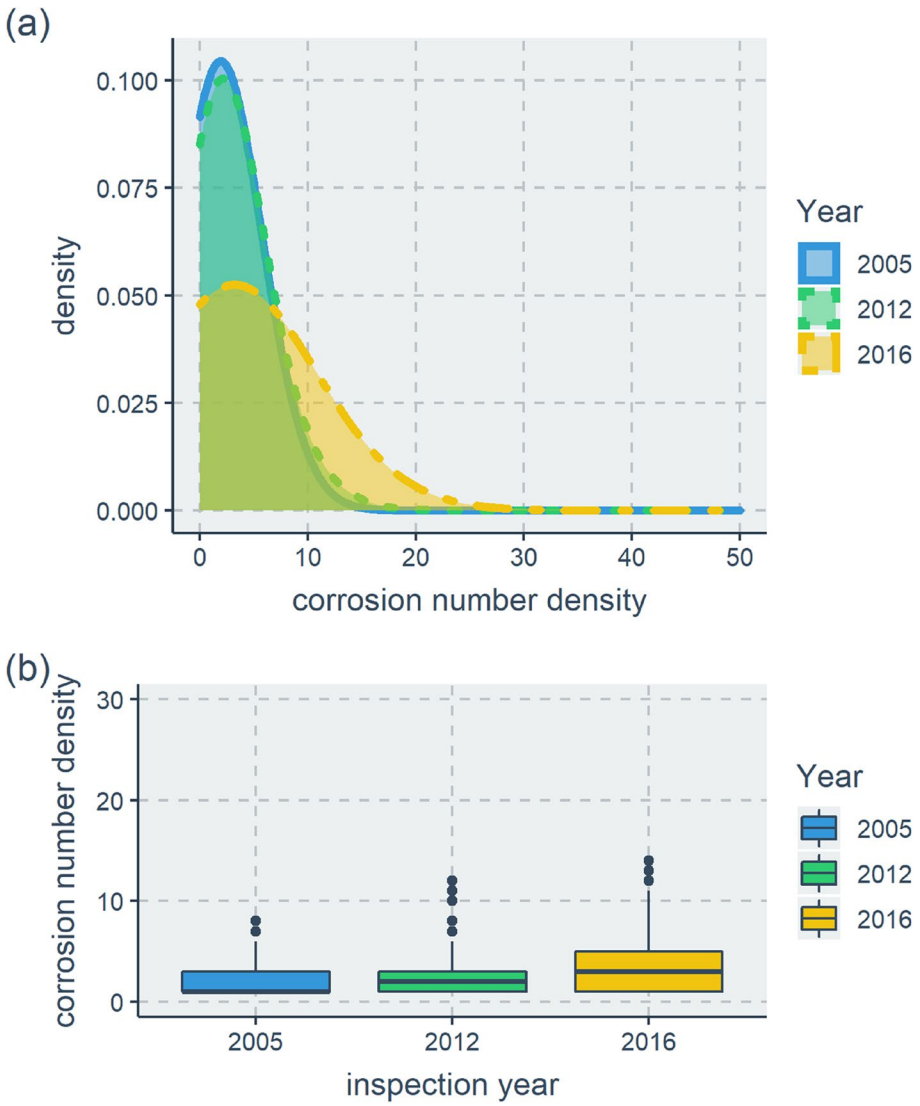

Fig. 13 Plot of corrosion number density (**a**) density plot, (**b**) boxplot

$$\mu(t) = 0.5161t - 1018.7 \tag{7}$$

$$\sigma(t) = 0.1085t - 210.1 \tag{8}$$

where, μ is the location parameter; σ is the scale parameter; and t is the inspection year.

Corrosion number density

In this study, corrosion number density denotes the number of defects per unit distance. Therefore, the number of corrosion defects in each segment of girth weld was used to construct the probabilistic model of number growth. As stated before, the number of defects tended to increase over time. However, there were many outliers in raw data, which had a negative effect on the fitting of growth distribution parameters. Therefore, raw data for the number of defects were processed to filter these outliers. Data points that deviated from the mean value by more than three times the standard deviation were deleted. The density and box plot of the processed data can be seen in Fig. 13.

Then, Weibull distribution was used to fit the corrosion number density. As shown in Fig. 14, most of the observations were located in the tails, which was consistent with non-stationary assumption of Weibull distribution [9]. In addition, it can be observed that the percentage of corrosion number density with high values increased over time. Therefore, more corrosion defects could be found in the same segment in 2016 than 2005 and 2012.

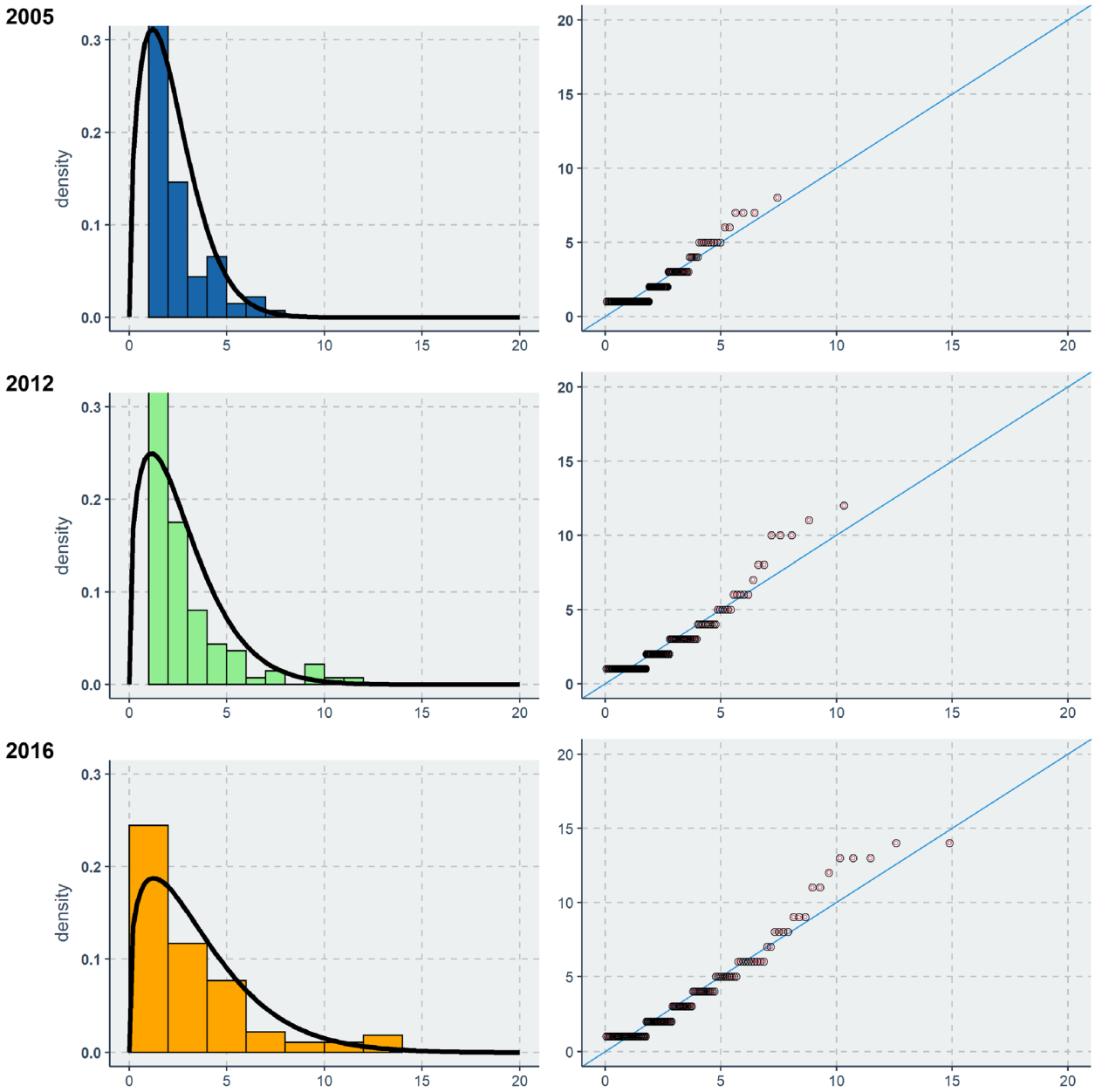

Fig. 14 Fitting of Weibull Distribution in different inspection years (**a**) histogram plot, (**b**) Q-Q plot

The fitted parameters of Weibull parameters can be seen in Table 6. And the linear regression of these parameters are shown in Fig. 15. It was found that the shape parameter decreased over time, but the scale parameters had an increasing trend. The linear model is expressed in Eq. (9) and (10). After obtaining the two parameters, the Weibull distribution can be used to calculate the density of corrosion number density at the year of interest.

$$\xi(t) = -0.021t + 43.56 \tag{9}$$

$$\sigma(t) = 0.1313t - 260.94 \tag{10}$$

Conclusions

This study used statistical analysis and data analytics to analyze ILI data of pipeline corrosions. Firstly, the distributions of corrosion depths and the number of corrosions on raw data were visualized. Then, the corrosion severity levels were classified based on the clustering of corrosion depth and ERF. Relationship between location

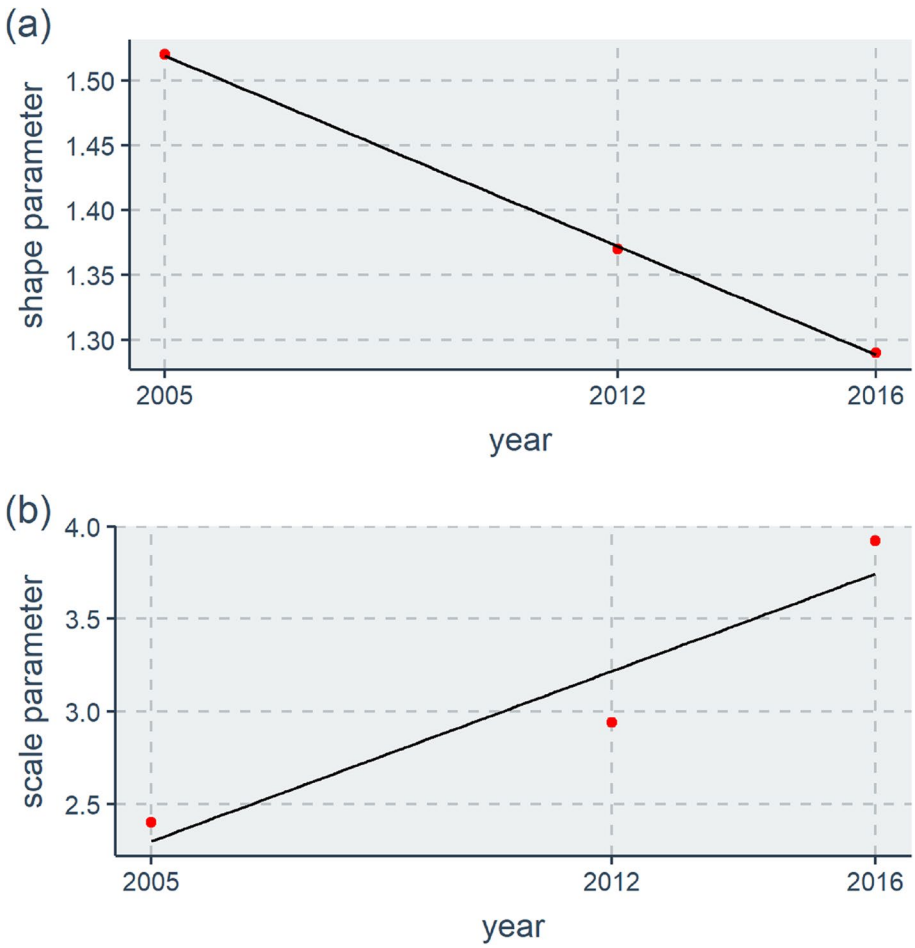

Fig. 15 Fitted lines of Weibull distribution parameters (**a**) shape parameter; (**b**) scale parameter with respect to inspection year

parameters and corrosion severity level considering interactive effects were explored. In addition, raw ILI data were processed to obtain useful data for establishing stochastic growth prediction models on maximum corrosion depth and corrosion number density.

The number of corrosion defects increased significantly over years. However, average corrosion depths decreased due to the occurrence of small corrosions and maintenance activities. In the longitudinal direction, corrosions were more likely to occur at 10 and 30 feet relative to pipeline joint; while in the circumferential direction, corrosions were prone to occur at the bottom of pipeline. In the segment of each girth weld number, the locations with shorter spacing between adjacent defects and the locations close to the girth weld were more prone to severe corrosion. For the entire pipeline, corrosion with higher severity level was mainly located in lower latitudes, indicating the soil environment in low latitudes may cause high corrosion potential.

The growth trend of two corrosion characteristics: maximum corrosion depth and corrosion number density were observed. Gumbel and Weibull distribution parameters of stochastic growth models can be used to predict the evolutions of maximum corrosion depth and corrosion number density, respectively. This study presents a detailed approach on how to obtain valid information from ILI data in practice, which can be further used for failure prediction and maintenance planning in pipeline integrity management system.

Table 6 Fitting parameters of Weibull distributions

Inspection year	Shape parameter (ξ)	Scale parameter (σ)
2005	1.52	2.40
2012	1.37	2.94
2016	1.29	3.92

Authors' contributions
B.Y. Cui: Data Curation, Investigation, Formal analysis, Original draft preparation; H. Wang: Supervision, Methodology, Writing- Reviewing and Editing. The authors read and approved the final manuscript.

Funding
USDOT Pipeline and Hazardous Materials Safety Administration (PHMSA).

Availability of data and materials
The dataset is provided by a third-party pipeline operator and can only be available after the specific request is made and approved.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 23 March 2023 Revised: 28 March 2023 Accepted: 4 May 2023
Published online: 02 June 2023

References

1. Arzaghi E, Abbassi R, Garaniya V, Binns J, Chin C, Khakzad N, Reniers G (2018) Developing a dynamic model for pitting and corrosion-fatigue damage of subsea pipelines. Ocean Eng 150:391–396. https://doi.org/10.1016/j.oceaneng.2017.12.014
2. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16(5):412–424
3. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
4. Caleyo F, Alfonso L, Espina-Hernandez JH, Hallen J (2007) Criteria for performance assessment and calibration of in-line inspections of oil and gas pipelines. Measur Sci Technol 18(7):1787
5. Caleyo F, Velázquez JC, Valor A, Hallen JM (2009) Markov chain modelling of pitting corrosion in underground pipelines. Corros Sci 51(9):2197–2207. https://doi.org/10.1016/j.corsci.2009.06.014
6. Chen H, Shu D (2001) Simplified limit analysis of pipelines with multi-defects. Eng Struct 23(2):207–213
7. Chiodo MS, Ruggieri C (2009) Failure assessments of corroded pipelines with axial defects using stress-based criteria: numerical studies and verification analyses. Int J Press Vessel Pip 86(2–3):164–176
8. Cohen S, Dror G, Ruppin E (2007) Feature selection via coalitional game theory. Neural Comput 19(7):1939–1961
9. Coles S, Bawa J, Trenner L, Dorazio P (2001) An introduction to statistical modeling of extreme values (Vol. 208): Springer.
10. Fisher RA, Tippett LHC (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. Paper presented at the Mathematical proceedings of the Cambridge philosophical society.
11. Fix E, Hodges JL (1989) Discriminatory analysis. Nonparametric discrimination: Consistency properties. Int Stat Rev/Revue Internationale de Statistique. 57(3):238–247
12. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Liu T-Y. (2017) Lightgbm: a highly efficient gradient boosting decision tree. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach
13. Khan F, Yarveisy R, Abbassi R (2021) Cross-country pipeline inspection data analysis and testing of probabilistic degradation models. J Pipeline Sci Eng 1(3):308–320
14. Li X, Bai Y, Su C, Li M, Piping (2016) Effect of interaction between corrosion defects on failure pressure of thin wall steel pipeline. Int J Press Vess 138:8–18
15. Liu Y, Hu H, Zhang D (2013) Probability analysis of damage to offshore pipeline by ship factors. Transp Res Rec 2326(1):24–31. https://doi.org/10.3141/2326-04
16. Mandl T, Modha S, Majumder P, Patel D, Dave M, Mandlia C, Patel A (2019) Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. Paper presented at the Proceedings of the 11th forum for information retrieval evaluation
17. Mathur A, Foody GM (2008) Multiclass and binary SVM classification: Implications for training and classification users. IEEE Geosci Rem Sens Lett 5(2):241–245
18. Mohd MH, Kim DK, Kim DW, Paik JK (2014) A time-variant corrosion wastage model for subsea gas pipelines. Ships Offshore Struct 9(2):161–176
19. Norske VD (2004) DNV Recommended practice. Corroded Pipelines, RP-F10.
20. Patle A, Chouhan DS (2013) SVM kernel functions for classification. Paper presented at the 2013 International Conference on Advances in Technology and Engineering (ICATE).
21. Powers DM (2020) Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv preprint arXiv:.16061.
22. Reber K, Beller M, Barbian A (2006) Run comparisons: using in-line inspection data for the assessment of pipelines. Paper presented at the Hannover: Pipeline Technology 2006 Conference.
23. Sharif MN, Islam MN (1980) The Weibull distribution as a general model for forecasting technological change. Technol Forecast Soc Change 18(3):247–256
24. Silva R, Guerreiro J, Loula A (2007) A study of pipe interacting corrosion defects using the FEM and neural networks. Adv Eng Softw 38(11–12):868–875
25. Sun J, Cheng YF (2018) Assessment by finite element modeling of the interaction of multiple corrosion defects and the effect on failure pressure of corroded pipelines. Eng Struct 165:278–286
26 Vanaei H, Eslami A, Egbewande A, Piping A (2017) A review on pipeline corrosion, in-line inspection (ILI), and corrosion growth rate models. Int J Press Vess 149:43–54
27. Xie M, Tian Z (2018) A review on pipeline integrity management utilizing in-line inspection data. Eng Fail Anal 92:222–239
28. Yarveisy R, Khan F, Abbassi R (2022) Data-driven predictive corrosion failure model for maintenance planning of process systems. Comput Chem Eng 157:107612

Publisher's Note