**CAAP Quarterly Report**

**January/5th/2022**

*Project Name: Pipeline Risk Management Using Artificial Intelligence-Enabled Modeling and Decision Making*

*Contract Number: 693JK32150001CAAP*

*Prime University: Rutgers University*

*Prepared by: Bingyan Cui (PhD student), Emad Farahani (PhD student), Dr. Hao Wang (PI), Dr. Qindan Huang (Co-PI)*

*Reporting Period: 10/1/2022 – 12/31/2022*

**Project Activities for Reporting Period:**

*Task 1 Literature Review (Completed)*

*Task 2 Data Collection from Industry Partners (Completed)*

*Task 3 Data-Driven Probabilistic Modeling of Pipeline Defects*

***Predicting Defect Growth Using Bayesian Neural Network***

According to the raw ILI data from the pipeline, external corrosion was found to be affected by soil physical parameters. The soil survey was conducted every 200 m along the pipeline (defined as one zone) in 2012, which means the pipe segments in each zone had the same soil properties. Data filter and processing were conducted based on the defect change of pipe segment from 2005 to 2010 in each zone. Two scenarios where the defect observations over time were taken into consideration to generate the processed dataset. In the first scenario, there existed defects at a specific zone during both inspections, and the defects in 2005 were smaller than those detected at the nearest location within the same zone in 2010. In the other scenario, there was no defect inspected at a specific zone in 2005, but initial defects were observed in 2010. According to the condition above, the data points with defects in two inspections were extracted to create the processed dataset for further analysis. After data processing, observations of metal loss at 296 zones and observations of defect length at 278 zones consisted of the processed dataset.

Figure 1 shows the scatter plot of metal loss and defect length along the pipeline length in the processed dataset. Multiple defect data points in each zone were included in the dataset. There were more defect data points in 2010 as compared to those in 2005, which was caused by the initial corrosion occurred during 2005-2010.

(a)

(b)

Figure 1 Scatter plot of pipe damage along pipeline length for the processed dataset: (a) metal loss; and (b) defect length

The statistical characteristics of defect data from two inspections in the processed dataset were illustrated by the boxplot (Figure 2) and distribution frequency plot (Figure 3). The comparison of statistical characteristics between the raw and processed datasets indicated that the pipe damage trend was more reasonable after data processing. Therefore, the processed dataset will be further analyzed to establish predictive models.

 

1. (b)

Figure 2 Boxplot for the processed dataset: (a) metal loss; and (b) defect length

1. (b)

Figure 3 Distribution frequency plots for the processed dataset: (a) metal loss; and (b) defect length

As Bayesian Neural Network (BNN) quantifies and propagates uncertainty, it is preferred over deterministic mathematical models like artificial neural network (ANN), and is more robust against overfitting and applicable for limited and noisy data (Neal, 1996). Compared to Gaussian process regression to establish Bayesian prediction models, BNN is more suitable for dataset with high-complexity and high-dimensionality.

Neural network consists as a set of nodes called ‘neurons’ with connections between them and is defined by the architecture of the neural network (e.g., the number of hidden layers and number of neurons per layer). Figure 4 illustrates the framework of BNN model for probabilistic modeling of degradation. where β are the network parameters. If using *w* and *b* to be the weight and bias for each neuron, the output of neuron y can be predicated by Eq.1.

$y=f\left(w∙x+b\right)$ (1)

where, *f*(·) refers to an activation function. In BNN, $β=\left\{w,b\right\}$ are modelled as random variables sampled from probabilistic distributions. With a joint prior distribution on the network parameters, the joint posterior distribution of the network parameters (including the covariance matrix with assumed zero mean error) will be kept updated after observing a set of training data. With the calculated posterior distribution, BNN can predict an output distribution.

In BNN, there are two most popular approximation methods: Markov chain Monte Carlo (MCMC) and variational inference (VI). Due to the limitation of dealing with high-dimensional posterior distribution using MCMC, VI is usually implemented as a mathematical model of nonlinear mapping between inputs and outputs to interfere the space of parameters in the BNN (Sun et al., 2019).



Figure 4 Illustration of BNN model for probabilistic modeling of degradation

The processed and zone-based datasets were used to establish BNN models for the prediction of metal loss and defect length. The dataset was divided into the training set (80%) and test set (20%) for cross-validation. In this study, the inputs included the soil parameters (Eh, resistivity, pH, concentration, soil moisture, and soil type), age of pipe segment after initial installation or replacement, elevation, and wall thickness. BNN models were developed for each defect type, and the coefficients of determination (R2) for model accuracy (based on test set) were summarized in Table 1.

 Table 1 Accuracy of BNN models for prediction of pipe defects

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Defect Type** | **R-square** | **Dataset Size** |
| Processed Dataset | Metal Loss | 0.49 | 2,222 |
| Defect Length | 0.35 | 1,659 |

 Figure 5 shows the measured vs. predicted defects with 95% conference interval by BNN models developed using the processed dataset. The blue points represented the measured values, the orange points stand for the mean value of predicted results, and the orange range corresponded to the confidence interval which was expected to cover 95% of observations from the view of statistics.

(a)

(b)

Figure 5 Measured vs. predicted defects with 95% conference intreval by BNN models developed using the processed dataset: (a) metal loss; and (b) defect length

***Predicting Defect Growth using Probabilistic Power Law Model***

As mentioned in the previous report, a power-law function of time model formation is adopted for corrosion growth as shown below.

|  |  |
| --- | --- |
| $$Y\_{m}\left(t,C\_{m}\right)=C\_{1,m}∙\left(t-t\_{0,m}\right)^{C\_{2,m}}+σ\_{m}ε\_{m}$$ | (2) |
| $$C\_{1,m}\left(θ\_{1},f\right)=θ\_{1,0}+\sum\_{i=1}^{n}θ\_{1,i} f\_{i}$$ | (2a) |
| $$C\_{2,m}\left(θ\_{2},h\right)=\frac{\left(θ\_{2,0}+\sum\_{j=1}^{m}θ\_{2,j} h\_{j}\right)^{2}}{1+\left(θ\_{2,0}+\sum\_{j=1}^{m}θ\_{2,j} h\_{j}\right)^{2}}$$ | (2b) |

Where, *m* = types of defect quantity (e.g., *m* = *D* for the maximum defect depth and *m* = *L* for the maximum defect length), *Ym*$ $= defect quantity (e.g., maximum defect depth or defect length) at a time instant *t*,**θ** = unknown model parameters, *fi* and *hj* = influencing environmental variables. In order to ensure the growth model provides descending growth rate over time, the power term of the growth model, i.e., *C*2,*m* needs to be bounded between zero and one. Thus, a transformation as shown in Eq. (1b) is adopted to meet such requirement. Therefore, regardless the value of $θ\_{2,0}+\sum\_{j=1}^{m}θ\_{2,j} h\_{j}$, *C*2,*m* obtained from Eq. (1b) will be always bounded between zero and one.

As discussed in the previous report, a Poisson process is assumed for defect generation, meaning that the time of initiation of each defect, *t*0, follows a Gamma distribution, where the shape factor, *α*, is assumed to be the ranking of each defect. This is based on the assumption that a defect with a larger detected dimension occurs earlier than others. Besides, the scale parameter, *β*, in the Gamma distribution is treated as an unknown model parameter and will be estimated through MCMC process.

In the MCMC sampling processing, the defect initiation time, *t*0, is randomly generated from the Gamma distribution. However, for some cases, the generated Gamma random number is greater than the time of inspection, *t*, meaning that the defect is initiated after the inspection time; this is unreasonable. To overcome this challenge, truncated Gamma distributions were employed so that upper bound of the distributions was considered to be equal to the time of inspection *t*.

In addition, it is found that when using MCMC to search the posterior distribution of unknown parameters (i.e., **θ** and *β*), the convergence is extremely slow or not achievable due to the large variability in *t*0: even for the same *β*, *t*0 could change dramatically since it is randomly generated for each MCMC sampling.

To overcome this issue, we remove the uncertainty associated with Gamma random number generation; at each iteration of the MCMC process, the mean value of the corresponding Gamma distribution, *µ*, is used as the defect initiation time, which is calculated as the product of Gamma distribution shape factor, *α*, and scale factor, *β*, by *t*0=*µ*= *α* × *β*. This way, the randomness associated with Gamma random numbers is eliminated. In fact, in this way, the model parameter *β*,which is updated in the MCMC technique and together with the shape parameter defines the Gamma distribution thus the initiation time.

The above-mentioned process is carried out for both corrosion depth and length growth. As mentioned in the previous report, a bi-variant Normal distribution is employed to construct the likelihood function considering the correlation between defect depth and length growth models.

Table 2 summarizes the statistics of the model parameters posterior distribution when 2 million iterations was considered for the MCMC process and soil moisture is taken into account in the model. Note that since moisture is included, for every single defect, a unique model is obtained based on the estimated *t*0 and also the corresponding soil moisture at that location.

Table 2 Posterior distribution statistics of model parameters in the damage evolution model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model for | Model parameter | Mean | Std. | Median | geweke |
| Depth | *θ*1,0 | 1.365 | 0.772 | 1.330 | 0.976 |
| *θ*1,1 | -4.262 | 2.595 | -4.131 | 0.969 |
| *θ*2*,*0 | 4.439 | 6.295 | 3.849 | 0.781 |
| *θ*2,1 | -22.478 | 28.518 | -19.434 | 0.833 |
| *βD* | 0.765 | 0.043 | 0.774 | 0.992 |
| *σD* | 0.198 | 0.008 | 0.196 | 0.987 |
| Length | *θ*1,0 | -0.181 | 0.969 | -0.203 | 0.048 |
| *θ*1,1 | 1.862 | 3.542 | 1.955 | 0.073 |
| *θ*2*,*0 | 22.621 | 568.403 | -14.901 | 0.941 |
| *θ*2,1 | 2.733 | 571.562 | 12.673 | 0.737 |
| *βL* | 0.178 | 0.028 | 0.173 | 0.953 |
| *σL* | 2.692 | 0.122 | 2.659 | 0.961 |

The predicted corrosion depth and length and the associated actual measured quantities obtained through ILI are compared in Figure 6. For a perfect prediction model, the predicted data should line up along the 1:1 line. One can see that most of the predicted data are located around the 1:1 line within the ± 1 standard deviation band (i.e., dashed lines). This indicates that the proposed defect growth models provide unbiased prediction with sufficient accuracy despite only one single growth model is used for the whole length of the pipeline. In addition, significant improvement is observable when the results are compared with those presented in the previous report by using transformation of power term, adopting truncated gamma distribution, and considering the impact of soil moisture.

Figure 7 shows the predicted corrosion damage evolution over time for all inspected defects. As can be seen the initiation time for each defect is unique, and also the corrosion rate tends to decrease over time, as expected and promised by the transformation shown in Eq. (2b). Besides, it can be observed that while depth model shows quite nonlinear, power-law behavior, length model is close to a simple linear model.

Based on the results at this stage, the correlation between depth and length model error is calculated as 2.3% implying that this correlation is not very strong; therefore, simplifying the MCMC process by separating depth and length models will not adversely affect the results.



(a)

(b)

Figure 6 Comparison between predicted versus measured defect dimension: a) corrosion depth, and b) corrosion length



(a)



(b)

Figure 7 Predictive corrosion evolution for all inspected defects over time: (a) corrosion depth, (b) corrosion length

**Project Activities with Cost Share Partners:**

Cost share is provided by Rutgers University and Marquee University during this quarterly period as budgeted in the proposal.

**Project Activities with External Partners:**

The PI attended the NDE summit organized by Pipeline Research Council International (PRCI) at Houston in October 2022. It is a good opportunity for the PI to see different types of ILI tools and technologies and interact with equipment manufacturers to understand the real operation practice of ILI.

**Potential Project Risks:**

N/A

**Future Project Work:**

The research team will continue working on Data-Driven Probabilistic Modeling of Defects.

1. The prediction accuracy of Bayesian Neural Network (BNN) based algorithm need be improved. Currently, the soil properties are measured per zone and there are multiple measurement points at each zone that cannot be explained by the same soil properties. One solution is to use zone-based data in the prediction model. Although this will reduce the size of dataset, the model accuracy can be improved by reducing the uncertainty of input data itself.
2. The prediction accuracy of probabilistic power-law models needs to be improved by incorporating further soil properties. Then, utilizing the developed predictive model, the time-dependent reliability of pipeline will be assessed as a case study through predicting the probability of failure per kilometer by defining a series system of defects considering three different failure modes such as small leak, large leak, and rupture. Finally, sensitivity analysis will be carried out to determine to which parameter(s) the reliability of the studied pipeline is most sensitive.

**Potential Impacts to Pipeline Safety:**

The ILI data will be used to develop probabilistic growth models of pipeline corrosion defects, which can aid pipeline operators better predict failure risk and make repair decisions.