Trial of damage prediction for telecommunication conduits using machine learning

A. Ito[1], M. Okutu[2], M. Yukishima[3], R. Matsushita[4], N. Hayashi[5], A. Furukawa[6], G. Shoji[7], T. Suzuki[8]

[1] *Research Engineer, Nippon Telegraph and Telephone Corp, akira.itou.dp@hco.nt.co.jp*
[2] *Senior Research Engineer, Nippon Telegraph and Telephone Corp, masaru.okutsu.ef@hco.ntt.co.jp*
[3] *General Manager, NTT DATA Mathematical Systems Inc., yukisima@msi.co.jp*
[4] *Senior Member, NTT DATA Mathematical Systems Inc., matsuhshita@msi.co.jp*
[5] *Member, NTT DATA Mathematical Systems Inc., hayashi@msi.co.jp*
[6] *Associate Professor, Kyoto University, furukawa.aiko.3w@kyoto-u.ac.jp*
[7] *Professor, Tsukuba University, gshoji@kz.tsukuba.ac.jp*
[8] *Professor, Toyo University, tsuzuki@toyo.jp*
…

Abstract

In Japan, telecommunication conduits protecting cables extend over about 620,000 km. When a large earthquake occurs, however, some conduits are broken, leading to cable damage. To protect cables efficiently, it is important to predict which pipelines are most susceptible to damage.

At present, conduits' vulnerability is evaluated in terms of the peak ground velocity (PGV), joint type, length, and micro-topography classification[1]. In addition, studies have evaluated vulnerability by adding parameters such as land assumed to be artificially flattened and equivalent predominant periods. According to this type of model, if countermeasures had been taken in preparation for the 2016 Kumamoto earthquake, 29% of conduit damage could have been prevented by taking countermeasures for the top 15% most vulnerable conduits. Nevertheless, there seems room for improving the accuracy[2].

In this study, to improve the accuracy of damage prediction for conduits, machine learning was used to develop a model on the basis of data with many parameters. A total of 25651 inspection results were used to examine the damage caused by the 2011 earthquake off the Pacific coast of Tohoku and the 2016 Kumamoto Earthquake. Of these, 337 instances had damage. The parameters in the learned data included earthquake data (PGV, PGA, IJ, etc.), ground data (micro-topographic divisions, AVS30, etc.), and facility data (length, joint type, etc.).

For learning, this study used a gradient-boosting decision tree: a supervised learning method known as a nonlinear model with comparatively easy accuracy. The model's area under the curve (AUC) by this method was 0.82. The results showed that 79% of damage could have been prevented if countermeasures had been applied for the top 15% most vulnerable conduits according to this index. Even with this method, however, when another earthquake was evaluated by a model learned on a different earthquake, the AUC dropped to 0.59, and the prediction accuracy greatly decreased.

In the future, we will investigate why this model generally contributes to accuracy improvement and what kinds of variables specifically contribute to accuracy improvement. We will also aim to construct a generalizable prediction model that physically matches data.

Keywords: machine learning, gradient-boosting decision tree, telecommunication conduit, damage estimation

1. Introduction

There are many underground facilities to protect communication cables in Japanese telecommunication networks, as shown in Fig. 1. These telecommunication facilities have high reliability, and the basic policy is to keep using as many current facilities as possible without building new ones [1]. On the other hand, some conduits are damaged during earthquakes, which may damage important cables. Although cable damage can be prevented by applying a cable protection lining in advance [2], that is difficult to do for all such pipes, which have a total length of about 620,000 km. Therefore, it is important to grasp the locations where pipeline damage is likely to occur in an earthquake beforehand and apply countermeasures efficiently.

A statistical method for predicting pipeline damage due to earthquakes was proposed by Shoji et al., and a method for predicting pipeline damage in Japan's Tohoku region was also proposed. This method statistically classified the susceptibility to damage according to the conduit type, length, and engineering geomorphic classification [3]. A similar method improved the accuracy by adding the equivalent predominant period to artificially flattened land in damage data from the 2016 Kumamoto earthquake, and then extracting the top 15% from all pipelines with damage. About 30% of the pipelines could be extracted [4].

To further improve on those techniques, this paper examines a prediction model using machine learning. For example, Bagriacik et al. used four methods to predict damage to water pipes in the 2011 Christchurch, New Zealand earthquake [5]. Taki et al. estimated damage by using density ratio estimation, an unsupervised learning method, for gas pipes in the 2011 earthquake off the Pacific coast of Tohoku [6]. Arita and colleagues used four types of machine learning to predict damage to water and sewage systems in Sendai, Miyagi Prefecture, following the 2011 Tohoku earthquake [7]. On the other hand, there have been few cases of using machine learning for damage prediction with multiple types of earthquakes.

In this paper, we apply machine learning to analyze data for communication pipelines damaged in the 2011 Tohoku earthquake and the 2016 Kumamoto earthquake, and to verify the predictability. Specifically, we verify whether a model can be constructed using both datasets, and what kind of prediction is obtained if a model learned with one dataset is applied to the other earthquake.

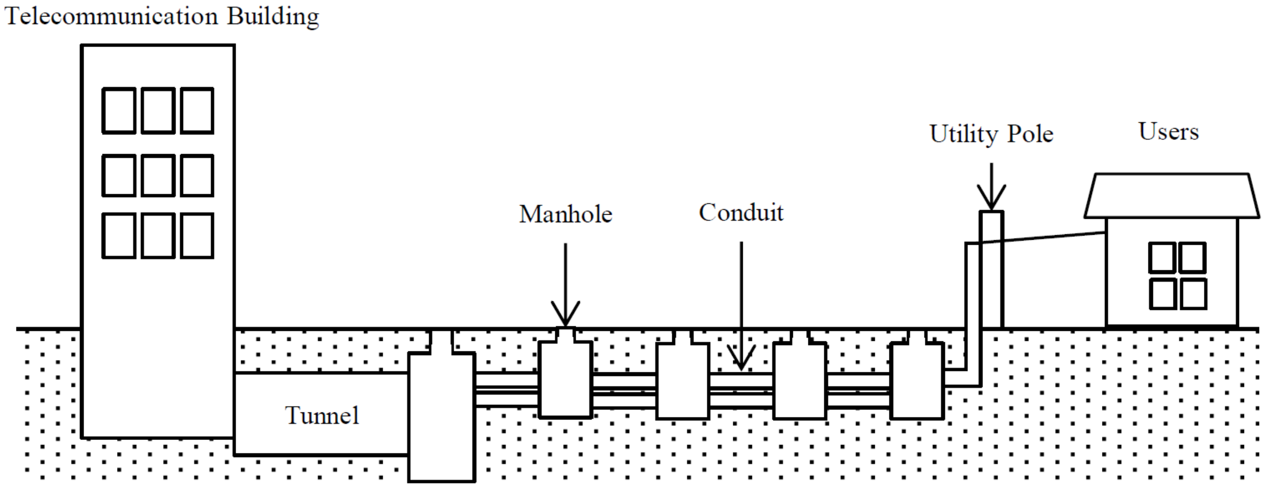


Fig. 1 Underground facilities for Japanese telecommunication

2. Method

2.1. Dataset

The target of analysis in this study was conduit damage data confirmed by inspection after the 2011 Tohoku earthquake and the 2016 Kumamoto earthquake. Communication pipelines are inspected in the case of conduits belonging to telecommunication buildings in municipalities where seismic motion with a Japanese seismic intensity of at least 6 was observed and the road surface was deformed. For the 2011 Tohoku earthquake, areas that were off-limits because of the nuclear accident were not inspected, and inundated areas considered to have been affected by the tsunami were also excluded from the data. The total number of records was 25651, with 337 indicating internal damage. Of these, 18600 were from the Tohoku earthquake, and 7051 were from the Kumamoto earthquake.

Pipeline damage was first confirmed in terms of whether a specified mandrel did not pass inspection; when it passed, it was considered sound. In cases of failure, a pipe camera was used to visually check the failure point. Then, in cases of bending, breaking, separation, sediment inflow, flattening, and so on, the damage was considered due to an earthquake. In contrast, when rust or artificial scars were visually confirmed, no earthquake damage was assumed. The presence of earthquake damage was used as an objective variable. Fig. 2shows an example of such damage.

Table 1 lists the parameters used as explanatory variables. Of the ground and landform information, the elevations and inclination angles from a 250-m mesh were obtained from Japan's National Land Numerical Information download service [8]. As for the engineering geomorphic classification, the data was used for places estimated to have been artificially flattened according to the J-SHIS map by the National Research Institute for Earth Science and Disaster Resilience (NIED) [9]. The basic natural period of the ground was obtained by two methods. Following Senna et al., one method was based on the horizontal-to-vertical (H/V) spectrum as calculated from microtremor observations [10] and determined for each engineering geomorphic classification [11]. The other method integrated bore data into the first method. The average shear-wave velocity (AVS30) was taken from the J-SHIS map and integrated with the bore data [12]. The records for peak ground acceleration (PGA), peak ground velocity (PGV), spectral intensity (SI), and Japanese seismic intensity were interpolated by kriging from observations by NIED's K-NET and KiK-NET seismograph networks, the Japan Meteorological Agency (JMA), and municipal seismic intensity meters [13]. The converted displacement was calculated as PGV^2/PGA. The ground strain was calculated in two ways: by using the pseudo-effective strain $\gamma'_{eff} = 0.4 \times PGV/AVS30$ [14], and by following a high-pressure gas conduit aseismic design guideline [15].

As noted above, the AVS30 was calculated by integration with bore data. Because the scale of the numerical data differed greatly, the data was standardized to have an average of 0 and a variance of 1. For categorical variables, one-hot encoding was performed by converting each category value to 0 or 1 as a parameter.

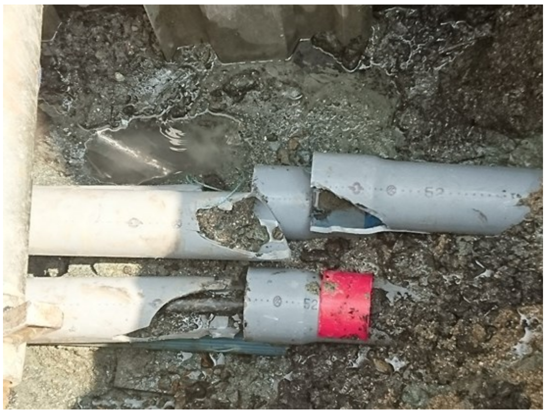


Fig. 2 An example of damage

Table 1 Overview of parameters

Division	Item	Description	Type of value
Conduit information	Conduit type	Type (joint type) of representative conduit in span	Category
	Length	Length of span	Value (m)
	Number of conduits	Number of conduits in span	Integer
	Age	Years since construction	Integer (y)
	Presence of shallowly buried conduit	Whether span includes shallow buried part	Boolean value
	Protected length	Length of conduit protected with concrete	Value (m)
Land and ground conditions	Basic natural period (integrated / method of Senna et al.)	Natural period of ground (derived two ways)	Value (s)
	Average elevation	Values in 250-m mesh obtained from National Land Numerical Information	Value (m)
	Highest elevation		
	Lowest elevation		
	Average angle		Value (°)
	Maximum angle		
	Minimum angle		
	Engineering geomorphic classification (artificial flat terrain)	Classification from J-SHIS data for artificially embanked sites	Category
	AVS30 (integrated / J-SHIS)	Average shear-wave velocity 30 m underground (derived two ways)	Value (m/s)
Seismic-wave parameters	PGA	Peak ground acceleration	Value (cm/s^2)
	PGV	Peak ground velocity	Value (cm/s)
	SI	Spectral intensity	Value (cm/s)
	Converted displacement	Displacement estimated from PGV and PGA	Value (cm)
	Ground strain (method of Yamazaki et al. / gas method)	Strain in surface layer	Value
	Equivalent predominant frequency	Earthquake frequency	Value (Hz)
	Japanese seismic intensity	Seismic intensity scale used in Japan	Value
Maintenance hole information	Age	Years since construction	Integer (y)
	Unique design	Whether special construction method was used	Category
	Size number	Size of maintenance hole (minimum 1, maximum 8)	Integer
	Maintenance hole type	Materials constituting maintenance hole	Category
	Number of covers	Number of covers on maintenance hole	Integer
	Maintenance hole shape	Whether shape diverges	Category
	Pooled water	Amount of water at maintenance hole	Category (3 steps)
	Spring water	Existence of spring at maintenance hole	Boolean value
	Stored gas	Detection of harmful gas	Boolean value
	H_2S	Detection of hydrogen sulfide	Boolean value
Cables	Number of optical cables	Number of optical cables in span	Integer
	Number of metal cables	Number of metal cables in span	Integer
	Total number of cables	Total number of all cables in span	Integer

2.2. Machine learning approach

In this study, we performed regression analysis using gradient boosting and constructed a prediction model. Boosting is a method of calculating a final predicted value by combining predictions from a series of weak classifiers. Here, a decision tree was used as the weak classifier. Let $\boldsymbol{x}_i \in \mathbb{R}^m$ be the feature values for data item i among n inputs, and let $y_i \in \mathbb{R}$ be its label. Here, m represents the number of features, as listed in Table 1, and $n = 25651$ is the total number of records. The decision rule $f(\boldsymbol{x}_i)$ can be expressed as the following:

$$f(\boldsymbol{x}_i) = \sum_{j=1}^{T} w_j I\left(\boldsymbol{x}_i \in R_j\right), \tag{1}$$

given a region R_j that is represented by the terminal vertex of the tree and does not overlap with X, and a value w_j assigned to each region. Here, T is the number of leaves in the tree, and I is the indicator function, which returns 1 if $\boldsymbol{x}_i \in R_j$ and 0 otherwise. By combining such decision trees, the predicted value $\hat{y}_i$ for i is expressed as the following:

$$\hat{y}_i = \phi(\boldsymbol{x}_i) = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i)\,. \tag{2}$$

In this case, K is the number of additive functions to predict the output. Boosting determines the next decision tree according to the one created in the previous step. Therefore, when determining the t-th decision tree, it is necessary to minimize an objective function $\mathcal{L}$:

$$\mathcal{L}^{(t)}(f_t) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(\boldsymbol{x}_i)\right) + \Omega(f_t) \quad , \tag{3}$$

where $l(a, b)$ is a loss function between a and b, and $\Omega(f_t)$ is a penalty term for the complexity of the t-th tree structure:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|\boldsymbol{w}\|^2\,. \tag{4}$$

Here, the coefficient γ is a penalty for the tree size, and λ is a coefficient for the value returned by the decision tree. Gradient boosting thus refers to the method of minimizing this objective function by the steepest-descent method.

In this study, we used the XGBoost library, specifically a package in R, to calculate the probability for binary classification. The data was divided into portions of 80% for learning and adjustment and 20% for evaluation. Note here that, in the former portion, 3/4 was used for learning and 1/4 was used for adjustment. In other words, of the entire data, 60% was used for learning and 20% was used for adjustment. Through the learning and adjustment process, a receiver operating characteristic (ROC) curve was created. The model was then tuned to maximize the area under the curve (AUC). Finally, the tuned model was applied on the evaluation data to generate predictions. Figure 1 shows an overview of the data division, learning, adjustment, and evaluation. As a result of the adjustment, the maximum tree depth was set to 7, γ was set to 0.77, and λ was set to 4.5.

In addition, note that learning and adjustment were performed only for the 2011 Tohoku earthquake data. Then, the 2016 Kumamoto earthquake data was used as test data for prediction through the same process described above, except for the division of the data.

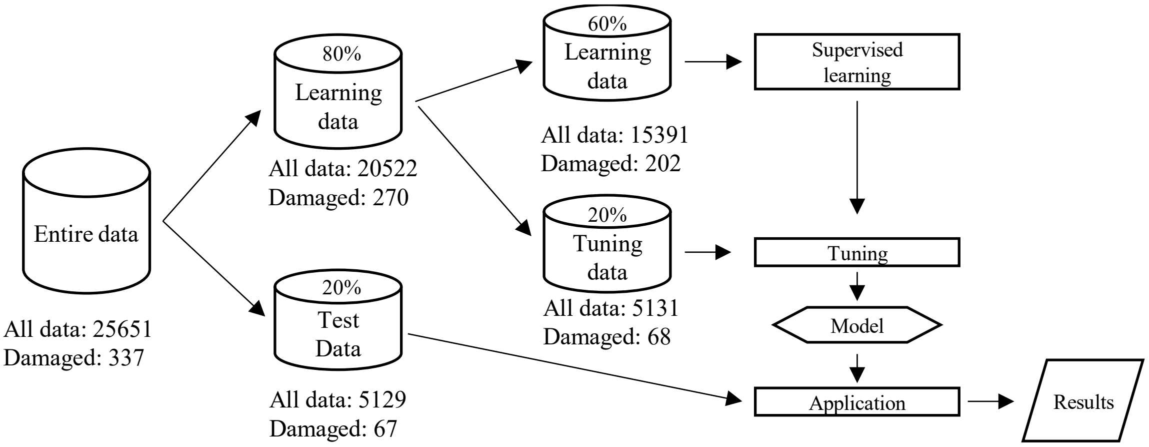


Fig. 3 Data division and model creation flow (data listed in terms of spans)

3. **Results**

3.1. Model construction results for all data

This section describes the results of applying the created model on the evaluation data. Figure 2 shows the distribution of the estimated damage probabilities via a histogram, and Fig. 3 shows the ROC curve. The AUC was 0.82, with a false positive rate of about 15% and a true positive rate of about 70%. We can conclude that data items with higher susceptibility were actually affected.

Next, we consider the feature gain to examine the contributions of the variables. Here, the feature gain is the sum of the loss reduction values resulting from each tree's branching. It is an index indicating how important each variable was in the progress of learning. Fig. 6 shows the top variables in terms of feature gain. The top variables included many seismic indices such as the seismic intensity, PGA, and PGV, which does not differ much from existing knowledge. On the other hand, as a pipe type, vinyl pipe with an adhesive splice was also among the top variables, suggesting that vinyl pipe might be more susceptible to damage than other pipe types are. In addition, the state of the maintenance hole, the cable information, and other factors were not among the top variables, and their influence may be relatively small. Finally, note that there are many seismic motion indices, and the feature gain may have been dispersed. Selection of variables is thus likely to be important.

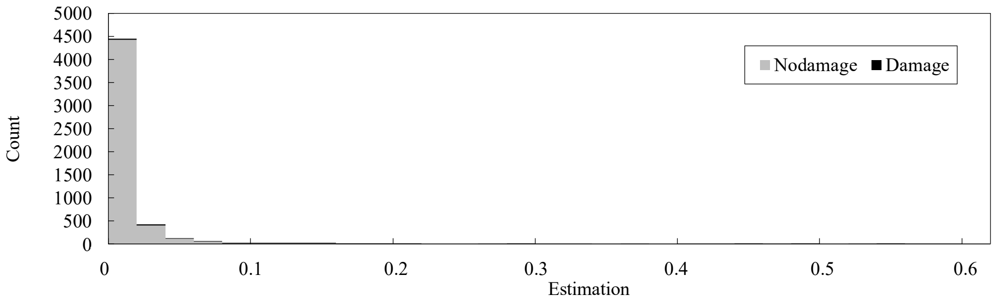


Fig. 4 Histogram

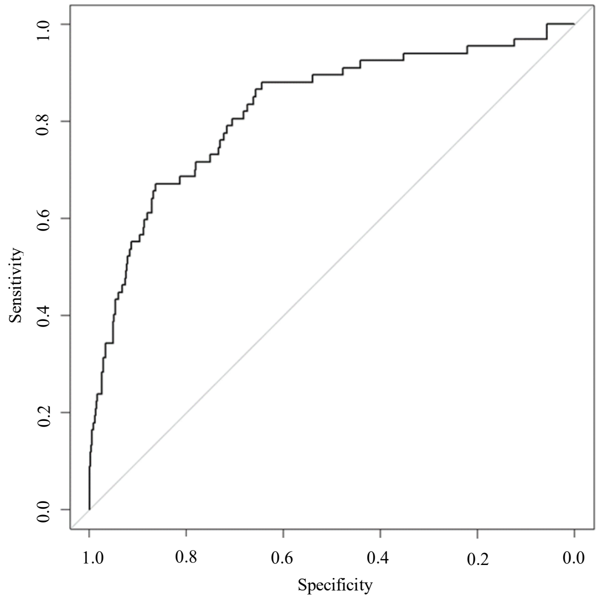


Fig. 5 ROC curve

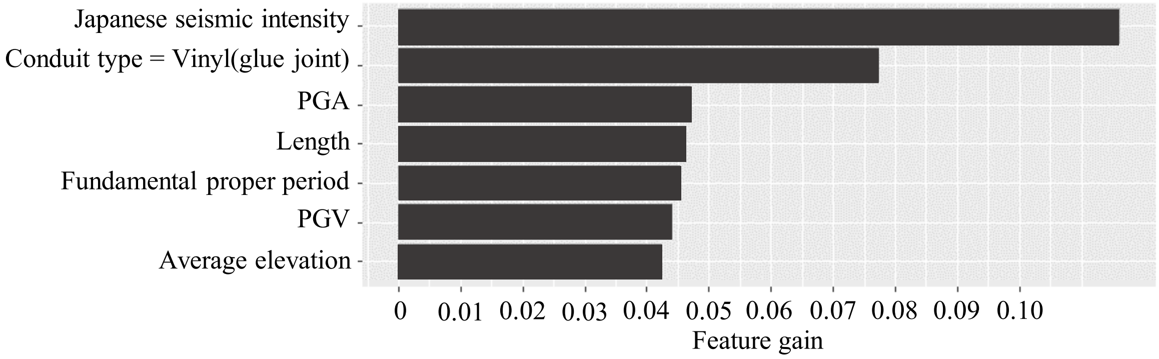
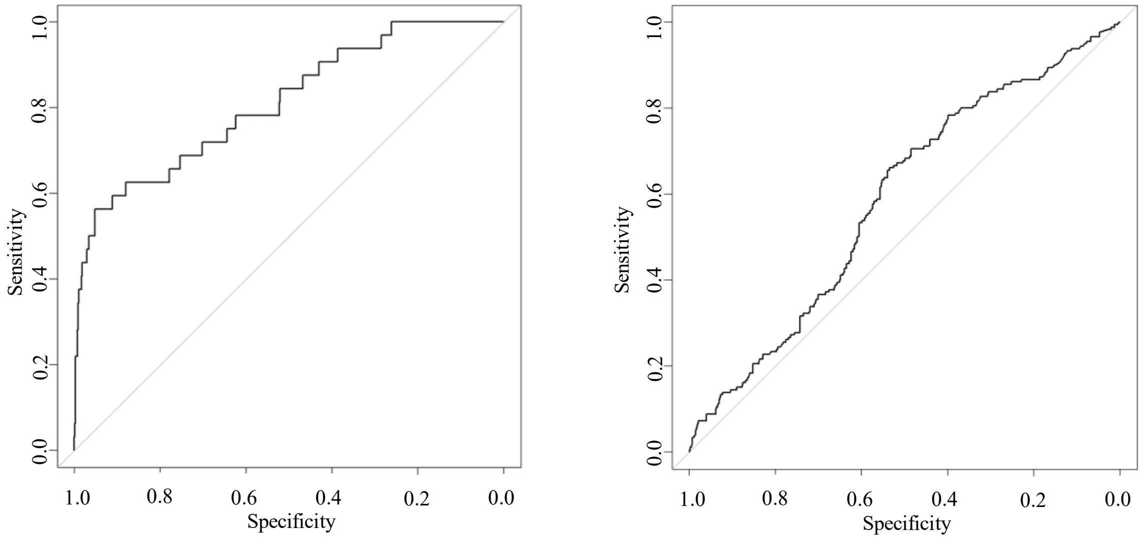


Fig. 6 Top variables for feature gain

3.2. Testing Kumamoto earthquake with model created from Tohoku earthquake

Next, we created a model by using the 2011 Tohoku earthquake data as training data. Figure 5 shows the ROC curve during learning and the predicted ROC curve for the 2016 Kumamoto earthquake. The AUC in (a) was 0.81, whereas that in (b) was significantly reduced to 0.59. This is probably because the Tohoku earthquake was a trench-type earthquake with short-period shaking of a long duration, while the Kumamoto earthquake was caused by an inland-type fault. Hence, the two earthquakes had ground motions with significantly different features, such as the ground motion having a peak period in the case of the Tohoku and Kumamoto earthquake. This suggests that it is difficult to apply a model learned from only a single earthquake directly to another earthquake.



(a) Learning result (2011 off the Pacific coast of Tohoku Earthquake)

(b) Test result (2016 Kumamoto Earthquake)

Fig. 7 ROC curves for model learned from 2011 Tohoku earthquake data

4. Conclusion

In this paper, a model was created using gradient boosting and evaluated on communication pipeline damage data from the 2011 Tohoku earthquake and 2016 Kumamoto earthquake. The following conclusions were obtained by examining the model.

1. When a damage prediction model was created by mixing data from both earthquakes and applied to test data, it had good prediction performance, with an AUC of 0.81.
2. The results suggested that the measured Japanese seismic intensity is the most important parameter in the mixed data, and that the span length, pipe type, and basic natural period are important except for seismic motion. This is because the parameters used in the current statistical method do not contradict the trends for those parameters, while conversely, information on connection maintenance holes and cables may not have a significant effect.
3. When a model was created using only the Tohoku earthquake data and applied to the Kumamoto earthquake, the prediction accuracy for the Kumamoto earthquake was greatly reduced, with an AUC of only about 0.59. This suggests that it is difficult to predict other earthquakes with a model based on a single earthquake.

In the future, we will continue studies using more data and studies for each parameter, such as the pipe type, to improve the damage prediction accuracy. In addition, by investigating the contributions of different variables, we want to construct a physical calculation model and clarify the characteristics of pipelines that are easily damaged.

5. References

[1] Fumihide Sugino, Hiroshi Masakura (2014): Maintenance and Management Technology for Safe, Scure, and Economical Operation of Telecommunication Infrastructure Facilities, *NTT Technical Review*, vol. 12, No. 10.

[2] Katsumi Sakaki, Katsuya Takeshita, Koji Tanaka, Akira Koizumi, Yoshihiko Tashiro, Ryo Seta, Katsuhiro Tanabe, Daiki Kobayashi (2014): Effective Repair and Reinforcement Technology for Conduit Facilities, *NTT Technical Review*, vol. 12, No. 10.

[3] Gaku Shoji, Fumihito Miyazaki, Masato Wakatake Akira Ito, Takanobu Suzuki (2016): Screening Techniques for Clarification of Seismic Vulnerability of Buried Telecommunication Pipes and Development of Seismic Damage Functions, *Journal of Japan Society of Civil Engineers, Ser. A1 (Structural Engineering & Earthquake Engineering (SE/EE))*, vol. 72, issue 4, pp. I_523-I_541 (Japanese).

[4] Akira Ito, Masaru Okutsu, Takanobu Suzuki, Gaku Shoji (2019): Development of Estimation Method of Seismic Damage for Telecommunication Conduits, *Seventh International Conference of Earthquake Geotechnical Engineering,* Rome, Italy, pp. 3062-3069.

[5] Adam Bagriacik, Rachel A. Davidson, Matthew W. Huges, Brendon A. Bradley, Misko Cubrinovski (2018): Comparison of Statistical and Machine Learning Approaches to Modeling Earthquake Damage to Water Pipeline, *Soil Dynamics and Earthquake Engineering*, 112, pp. 76-88 (Japanese).

[6] Yuta Taki, Wataru Inomata, Yoshihisa Maruyama (2018): A Study on Pipeline Damage Estimation Using Machine Learning Techniques, *Proceedings of the Ninth Symposium on Disaster Mitigation and Resilience of Infrastructures and Lifeline Systems,* pp. 71-75 (Japanese).

[7] Kyohei Arita, Yoshihisa Maruyama (2019): Use of machine learning of earthquake damage location in water supply pipeline, *Proceedings of the Ninth Symposium on Disaster Mitigation and Resilience of Infrastructures and Lifeline Systems,* pp.63-72. (Japanese)

[8] National Land Information Division, National Spatial Planning and Regional Policy Bureau, Ministry of Land, Infrastructure, Transport and Tourism (MLIT), Japan (2009): National Land Numerical Information download service, Retrieved January 16, 2020, http://nlftp.mlit.go.jp/ksj/other/faq.html.

[9] National Research Institute for Earth Science and Disaster Resilience (2019): *J-SHIS Map*, Retrieved January 16, 2020, http://www.j-shis.bosai.go.jp/map/.

[10] Shigeki Senna, Saburoh Midorikawa, Kazue Wakamatsu (2008): Estimation of Spectral Amplification of Ground Using H/V Spectral Ratio of Microtremors and Geomorphological Land Classification, *Journal of Japan Association for Earthquake Engineering*, vol. 8, issue 4, pp. 1-15.

[11] Shigeki Senna, Saburoh Midorikawa (2009): Estimation of Spectral Amplification of Ground Motion Based on Geomorphological Land Classification, *Journal of Japan Association for Earthquake Engineering*, vol. 9, issue 4, pp. 4_11-4_25.

[12] Iwao Suetomi, Eisuke Ishida, Yasuhro Fukushima, Ryuji Isoyama, Sumio Sawada (2007): Mixing Method of Geomorphologic Classification and Borehole Data for Estimation of Average Shear-Wave Velocity and Distribution of Peak Ground Motion During the 2004 Niigata-Chuetsu Earthquake, *Journal of Japan Association for Earthquake Engineering*, vol. 7, issue 3, pp. 1-12 (Japanese).

[13] Akira Ito, Masaru Okutsu, Iwao Suetomi, Hiroyuki Tsukamoto, Takanobu Suzuki (2019): Damage Analysis of Rigid PVC Pipe and Steel Pipe for Communication, *Proceedings of the 39th JSCE Earthquake Engineering Symposium,* Osaka, Japan, C21-1482 (Japanese).

[14] Makoto Yamaguchi, Saburoh Midorikawa (2014): Empirical Models for Nonlinear Site Amplification Evaluated from Observed Strong Motion Records, *Journal of Japan Association for Earthquake Engineering*, vol. 14, issue 1, pp. 56-70 (Japanese).

[15] Japan Gas Association, Technical Committee on Technical Standards for Gas Works (2004): Guidelines for Seismic Design of High Pressure Gas Pipes, JGA-206-03.

[16] Trevor Hastie, Robert Tibshirani, Jerome Fredman (2014): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Edition.

[17] Tianqi Chen, Carlos Guestrin (2016): XGBoost: A Scalable Tree Boosting System, *KDD'16, San Francisco*, CA, USA, pp. 785-794.