Desalination and Water Treatment
www.deswater.com

doi:10.5004/dwt.2018.22714

120 (2018) 190–197
July

Statistical approach to developing screening models for pipe failure events in water network systems

Bumjo Kim[a], Seo Jin Ki[b], Dong Jin Jeon[a], Joon Ha Kim[a,*]

[a]*School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Korea, Tel. +82-62-715-3391, email: joonkim@gist.ac.kr (J.H. Kim)*
[b]*Department of Environmental Engineering, Gyeongnam National University of Science and Technology, 33 Dongjin-ro, Jinju-si, Gyeongsangnam-do 52725, Korea*

Received 9 April 2018; Accepted 20 June 2018

ABSTRACT

Accurate assessment of piping systems' risk to damage reduces annual operation and maintenance costs. Recently, extreme climate events (e.g. cold snaps or heavy snow) due to global climate change have increased pipe system failure. The objective of this study is to establish a framework for developing screening models of pipe failure events, due to water network systems freezing using two statistical approaches. More specifically, logistic regression was used to estimate the probability of failure at a household level, whereas the customized model developed to predict the frequency of community-wide failure events. The data recorded at least one failure event in Korea from 2008 to 2015, which was provided to the logistic regression model. The customized model, however, only used the data set compiled from three areas of concern with the highest frequency of the failures. Results showed that the logistic model showed the best performance out of the 11 constructed models, in terms of *R* and the variance inflation factor (of lower than two). The logistic model incorporated three variables: the minimum temperature on the day of failure, the natural logarithm of the total water usage in the previous month and the mean minimum temperature over the previous 10 days. The selected model had an overall prediction accuracy of 66.4%. When the customized model at the community level was examined the three models not only yielded moderate R^2 values ranging from 0.53 to 0.66, but also helped identify water network systems at risk of failures. Overall, this study demonstrated that the proposed methodology can be used to highlight areas of concern at different geographic scales, along with refining existing statistical models with new variables updated in real time.

Keywords: Pipe failure event; Logistic regression; Customized model; Minimum temperature; Total water usage; Variance inflation factor

1. Introduction

Insurance companies revealed that pipe failures due to freezing or bursting are the leading source of residential water losses worldwide [1]. In addition, 80% of the total cost of a water network is used to operate or maintain the water supply system. Currently, in Korea, the probability of pipe failure is classified into four stages (very high, high, normal, and low) based simply on the daily minimum temperature. Pipe failure is dependent upon geographic location and temperature [2], but it is difficult to predict with only these two pieces of information. The risk of failure can actually be attributed to combined effects of various causes.

To identify and predict pipe failure due to freezing, many studies were performed using physical deterministic and statistical models dependent on the available data [3]. Physically-based models used data such as soil properties, pipe properties, hydraulic conditions, and environmental conditions [4–6]. These studies were focused on the mechanical behavior of materials by analyzing the stress on the pipes.

*Corresponding author.

While the failure of a pipe could be experimentally verified, there still remain restrictions due to insufficient data available. Therefore, statistical models have generally been used to predict pipe failures, by attempting to define the relation of historical failure data with environmental conditions [7–13].

This study developed a logistic regression model to estimate the probability of a failure at the household level. In addition, a customized model was developed to predict the frequency of community-wide pipe failure events when considering site-specific characteristics. This study revealed that using the proposed methods and models identified the factors related to freeze events, and the water networks at risk of failure. Further, the proposed methodology could be utilized for forecasting, by updating the current statistical model with additional monitoring data in real time.

2. Materials and methods

2.1. Study area

The study area was located throughout the Republic of Korea. The Korea peninsula is geographically located in mid-latitude and approximately 70% of the country is composed of mountainous terrain. Climatologically, Korea belongs to a tropical climate zone with four distinct seasons. Winters are long, cold and dry due to the expanding Siberian high-pressure air mass. Over the past 30 years (1981 ~ 2010), the average annual precipitation of the Korean Peninsula was 1,162.2 mm, the annual average temperature was 11.0°C, the average winter temperature was –1.7°C, and the average daily temperature range was 10.4°C [18]. Fig 1. Shows the spatial distribution of pipe failure events in Korea.

2.2. Data acquisition

In this study, historical pipe failure event data from 2008 to 2015 of 21 cities were acquired from the Korea Water Resources Corporation. Pipe failure event data contained the number of failure events as well as the time of the event and the properties of pipe (e.g. diameter of pipe, installation year of pipe and pipe material). In addition, the water use data were provided by Korea Water Resources Corporation. Meteorological data, which contained daily mean temperature, maximum temperature and minimum temperature were acquired from Automatic Weather Stations (AWS) adjacent to individual households. To analyze the influence of persistent low temperatures, the average mean temperature 10 d antecedent the pipe failure events, the average

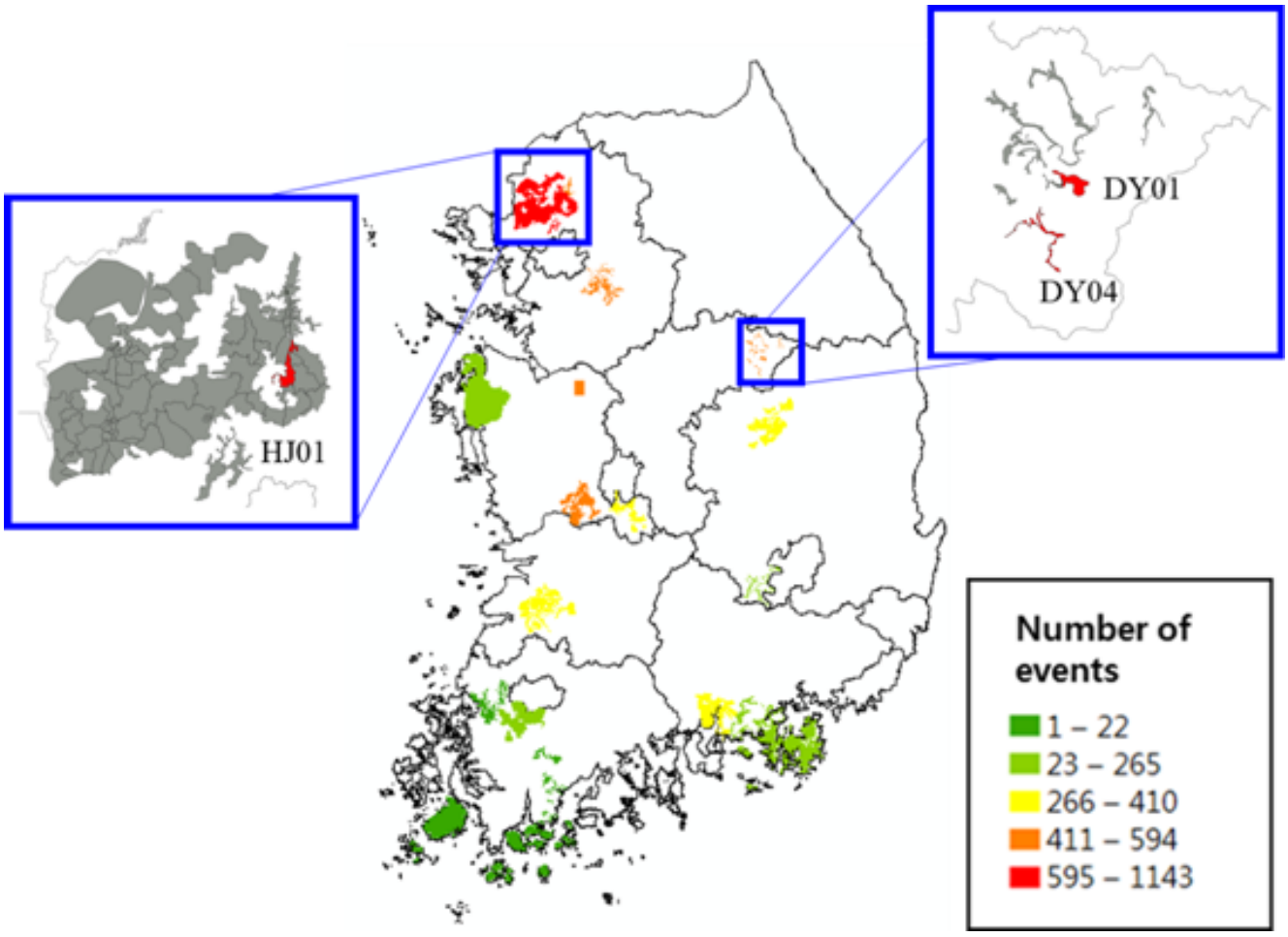

Fig. 1. A map showing pipe failure events in Korea during the monitoring period from 2008 to 2015.

maximum temperature 10 d antecedent the pipe failure events the average minimum temperature 10 d antecedent the pipe failures were calculated for further analyses. The same number of individual households within the range of temperature and water use conditions for frozen pipe failures could not be matched, and instead were randomly extracted to set the control data for the experiment.

2.3. Logistic regression model

Logistic regression analysis proposed by Cox (1958) predicted which groups of individual observations can be classified when the subjects of an analysis are categorized into two or more groups [19]. When the dependent variable is binary, such as the occurrence of an event or not, it is called a binomial logistic regression, which is a specific case of linear regression with S-shape curve function. The binomial logistic regression is used to predict the occurrence of an event from a number of categorical or continuous explanatory variables. The probability of occurrence for the independent variables X_i can be expressed by the following equation [20]:

$$p = \frac{1}{1 + e^{\beta_i X_i}} \quad (1)$$

where β_i is coefficient of the logistic regression model.

With logistic regression analysis, the possibility of pipe failure events was analyzed using a combination of specific explanatory variables. For this purpose, the influence of individual explanatory variables on the occurrence of failure was analyzed and the probability of failure was calculated from the combination of selected explanatory variables. The sensitivity and predictive accuracy of the model constructed using the cut-off value of the logistic regression was also analyzed.

2.3.1. Selection of major factors

For the logistic regression analysis, the dependent variable was set as the categorical type, which represented occurrence and non-occurrence of pipe failure. The rest of the data were used as explanatory variables in model construction, as summarized in Table 1. The logarithm of water usage during the month prior to the pipe failure event (Log_Q_b) and the average minimum temperature during the antecedent 10 days (T_{bmin10}) were forcibly entered as explanatory variables. The model was then determined by selectively adding variables in descending order of explanatory power.

2.3.2. Model selection

To determine the optimal explanatory variables, collinear variables were excluded based on the average variance inflation factor (VIF). Collinear variables are those which can be predicted from other variables with a considerable degree of accuracy. The VIF indicates correlation between explanatory variables, and should remain less than five, which is the maximum value to avoid multicollinearity

Table 1
Summary of variables used to develop logistic and customized models

Types	Groups of variables	Raw and derived variables	Units	Description
Dependent variable	Failure events	E	–	The frequency of pipe failure events
Explanatory (Independent) variables	Pipe properties	Pipe_install	year	The installation year of a pipe
		P_{type}	–	The types of a pipe
		P_{width}	mm	The diameter of a pipe
	The amount of water usage	Q_t	tons	The total water usage in the month at failure
		Q_b	tons	The antecedent total water usage in the previous month prior to failure event occurrence
		Log_Q_b	–	The logarithm of Q_b
	Air temperature	$T_{max0–10}$	°C	The maximum temperature the day of failure (T_{max0}); The antecedent maximum temperature 10 days prior to failure event occurrence ($T_{max1} – T_{max10}$)
		$T_{min0–10}$	°C	The minimum temperature the day of failure (T_{min0}); The antecedent minimum temperature 10 days prior to failure event occurrence ($T_{min1} – T_{min10}$)
		$T_{mean0–10}$	°C	The mean temperature the day of a pipe failure event (T_{mean0}); The antecedent mean temperature 10 days prior to failure event occurrence ($T_{mean1} – T_{mean10}$)
		$T_{bmin1–10}$	°C	The mean antecedent minimum temperature (e.g., T_{bmin10} indicates the mean minimum temperature during the day of failure and 10 days prior to failure event occurrence)

[21]. The selected model had an average VIF value of less than two and a relatively high coefficient of determination *R*, which indicated explanatory power.

2.4. Development of customized model

The equation for the customized model explaining the frequency of pipe failure events was designed using the explanatory variables from the previously derived logistic regression model. T_{bmin10}, Q_b, and T_{min0} were chosen as explanatory variables for further analyses. Since the frequency of a pipe failure event and the temperature were normally distributed and the water usage was exponentially distributed, the relationship between each explanatory variable and the frequency of pipe failure were expressed by the following equations.

$$f\left(T_{\min 0}\right) = a_1 e^{-b_1\left(T_{\min 0} - c_1\right)^2} \tag{2}$$

$$f\left(T_{bmin10}\right) = a_2 e^{-b_2\left(T_{bmin10} - c_2\right)^2} \tag{3}$$

$$f\left(Q_b\right) = a_3 e^{-b_3 Q_b} + c_3 \tag{4}$$

where T_{min0} is the minimum temperature on pipe failure day, T_{bmin10} is the average of minimum temperature during the 10 days antecedent to pipe failure, Q_b is total water usage for the previous month, and *a*, *b*, and *c*are constants.

Eqs. (2), (3) and (4) are expressed as follows through natural log transformations.

$$\ln f\left(T_{\min 0}\right) = -b_1\left(T_{\min 0} - c_1\right)^2 + \ln a_1 \tag{5}$$

$$\ln f\left(T_{bmin10}\right) = -b_2\left(T_{bmin10} - c_2\right)^2 + \ln a_2 \tag{6}$$

$$\ln f\left(Q_b\right) = -b_3 Q_b + \ln c_3^{*} \tag{7}$$

The rearranged formulae for the frequency of a pipe failure event are expressed as follows when combining Eq. (5), (6) and (7).

$$\mathrm{n}f\left(T_{\min 0}, T_{bmin10}, Q_b\right) = -b_1\left(T_{\min 0} - c_1\right)^2 - b_2\left(T_{bmin10} - c_2\right)^2 - b_3 Q_b + R \tag{8}$$

where *R* is constant. Finally, rearranging by Eq. (8), the generalized model equation for each community-level area used for further analysis was derived.

$$\ln f\left(T_{\min 0}, T_{bmin10}, Q_b\right) = m_0 + m_1 {T_{\min 0}}^2 + m_2 T_{\min 0} + m_3 {T_{bmin10}}^2 + m_4 T_{bmin10} + m_5 Q_b \tag{9}$$

where m_0, m_1, m_2, m_3, m_4, and m_5 arecharacteristic coefficients of specific sites. As a constraint condition for coefficient estimation, m_0 was assigned to a value higher than zero.

A methodology for intuitive and quick risk classification based on historical statistics was additionally investigated. Hence, a simple method of categorizing frequency of pipe failure events into a fewer number of warning levels was proposed. The warning levels of pipe failure were classified into Danger, Warning, and Caution by using the average and half standard deviation of the natural logarithm of observed frequency of failure events. The criteria and procedure for classifying the warning levels are shown in Fig. 2.

3. Results and discussion

3.1. Logistic regression model at the household level

Fig.3 compares the performance of different logistic regression models, which were developed at the household level, with respect to VIF (see solid bars) and *R* values (see a solid line with maker). Note that the variables incorporated into individual logistic regression models are provided in Table 2. From the figure, it was observed that the average VIF values generally increased with an increasing number of variables included in the model (Models 1–9), though Model 6 recorded a sudden drop in the average VIF value. The decrease in the average VIF value for Model 6 was likely attributed to P_{width} which was added to Model 5. Models 10 and 11 were found to have average VIF values significantly lower than the other models that included more than six variables (i.e., Models 5–9), due to adopted new variables (e.g., Q_b) as well as dropped old variables (T_{bmin10}) from the previous models (i.e., Models 5 to 9). When the model quality in terms of *R* was investigated, there was no significant difference in performance among Models 2–11. This result indicated that even if more than three variables were incorporated (Models 3–11), the model performance was not improved considerably from Model 2. Therefore, when considering VIF and *R* values for, Model 2, which simultaneously satisfied both conditions, was selected as the optimal model for this study. Note that low VIF values indicate a weak correlation among variables, such that the model with an average VIF value lower than two is free from multicollinearity issues. The coefficients of individual variables in Model 2 are summarized in Table 3. Within the table, the interpretation of exp(B) indicates the odds ratio of the independent variable. If the value of exp(B) is less than one, the independent variable becomes smaller as the explanatory variable becomes larger. However, if exp(B) is larger than one, the opposite is true. In the case of exp(B) approaching one, changes to incorporated variables were not associated with that of dependent variable. Therefore, Model 2 implied that as Log_Q_b, T_{min0}, and T_{bmin10} increased, the probability of a failure event decreased.

3.2. Accuracy of the selected logistic regression model

Table 4 shows the prediction accuracy of the selected logistic regression model (Model 2) by comparing the number of pipe failure events predicted by the model with the corresponding (field) observations. The cut-off value was 0.4 was used for accuracy assessment, which determined the sensitivity of the model. The table showed that the predicted accuracy of the model for failures that actually occurred was 59.0%, and for failures that did not occur was

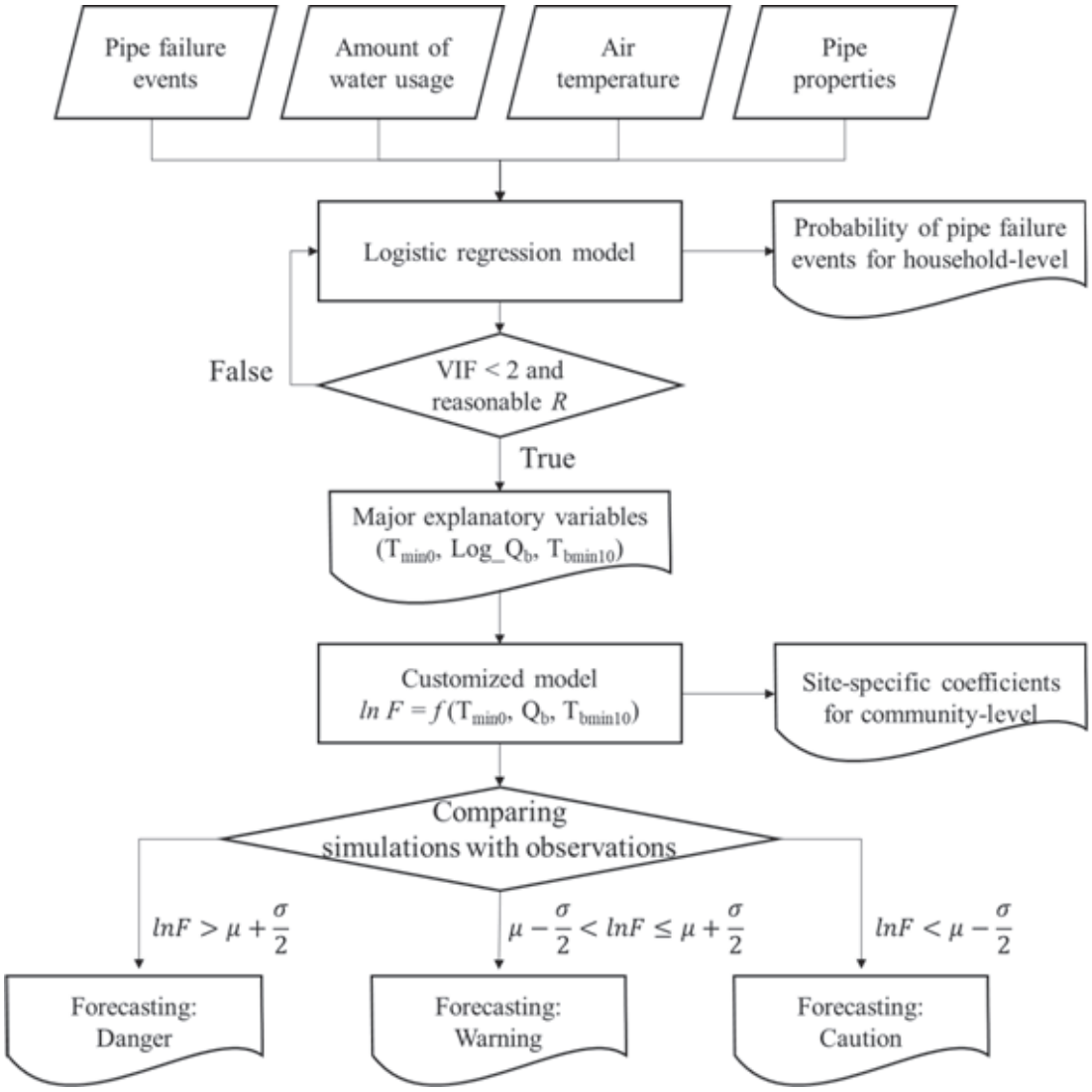

Fig. 2. Schematic diagram of developing screening models at household and community levels based on the monitoring data set of pipe failure events. Note that the logistic model uses both raw and derived variables, whereas only uncorrelated variables are used in the customized model.

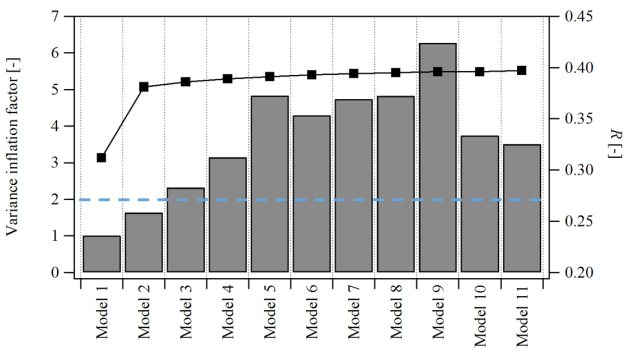

Fig. 3. Performance assessment of different logistic regression models in terms of VIF (indicated by solid bars) and R (indicated by a solid line with a marker). The dotted line (indicated by light blue color) represents the VIF cut-off value to select the best logistic regression models out of all models.

75.5%. The overall prediction accuracy of the selected logistic regression model was 66.4%, indicating that the model was relatively good at predicting a failure event at the household level.

3.3. Customized model at the community level

Nonlinear regression was used to develop the customized model at the community level, examining in three target areas (DY01, DY04, and HJ01), recording the highest frequency of failures. Note that independent and dependent variables incorporated into the logistic regression model at the household level were identically applied to the customized model, except for two variables employed in a different form: Q_b instead of Log_Q_b, and ln F instead of F. Based on the relationship between dependent and independent variables, the customized

Table 2
Logistic regression models developed from a series of variables

Models	*R*	Std. error of the estimate	Variables incorporated in each model
1	0.312	0.473	T_{min0}, Log_Q_b
2	0.381	0.460	T_{min0}, Log_Q_b, T_{bmin10}
3	0.386	0.459	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}
4	0.389	0.458	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}, T_{mean1}
5	0.391	0.458	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}, T_{mean1}, T_{min1}
6	0.393	0.458	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}, T_{mean1}, T_{min1}, P_{width}
7	0.394	0.457	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}, T_{mean1}, T_{min1}, P_{width}, T_{mean10}
8	0.395	0.457	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}, T_{mean1}, T_{min1}, P_{width}, T_{mean10}, T_{max0}
9	0.396	0.457	T_{min0}, Log_Q_b, T_{bmin10}, T_{mean4}, T_{mean1}, T_{min1}, P_{width}, T_{mean10}, T_{max0}, T_{min7}
10	0.396	0.457	T_{min0}, Log_Q_b, T_{mean4}, T_{mean1}, T_{min1}, P_{width}, T_{mean10}, T_{max0}, T_{min7}
11	0.397	0.457	T_{min0}, Log_Q_b, T_{mean4}, T_{mean1}, T_{min1}, P_{width}, T_{mean10}, T_{max0}, T_{min7}, Q_b

Table 3
Coefficients of the selected logistic regression model (Model 2)

Variables	Coefficients (B)	Standard error	*p*-value	exp(B)	95% confidence intervals for individual coefficients	
					Lower bound	Upper bound
Log_Q_b	–0.701	0.042	0.000	0.496	0.457	0.538
T_{min0}	–0.021	0.006	0.000	0.980	0.969	0.990
T_{bmin10}	–0.151	0.008	0.000	0.860	0.847	0.873
Constant	–0.760	0.055	0.000	0.468		

Table 4
Prediction accuracy of the selected logistic regression model (Model 2)

Observation	Prediction			Prediction accuracy [%]	Overall prediction accuracy [%]
	Pipe failure event				
	Occurrence	Non-occurrence			
Pipe failure event	Occurrence	2469	1716	59.0	66.4
	Non-occurrence	833	2563	75.5	

model included two additional variables (expressed in quadratic form) derived from the raw variables T_{min0} and T_{bmin10} (see Eq. (9)). The coefficients, standard errors, and 95% confidence intervals are summarized in Table 5. The coefficients of five individual variables varied considerably depending on the study area. This indicated that the coefficients could be used as site characteristic properties that reflected failures at the community level. Fig. 4 illustrates the performance of the individual customized models developed for three test areas using those characteristic properties. It was determined that the customized models efficiently captured the failure events at the community level, as most of observations were within the 95% confidence band. The customized models yielded coefficients of determination (R^2) of 0.658 (for DY01), 0.529 (for DY04) and 0.593 (for HJ01).

Fig. 5 exhibits the example alert system that classifies the status of pipe networks for three test areas based on the outputs predicted by the customized models. Three variables contributed to the failure event occurrence for the three test areas. According to the example alert system, warning or caution were assigned when T_{min0} and T_{bmin10} were between –10°C and 0°C and/or when Q_b was over 20 tons. The system assigned dangerous when both T_{min0} and T_{bmin10} were under –10°C as well as when Q_b was below 20 tons. This implied that there was a high risk for failure due to freezing and bursting if the air temperature was low, the influence of the low temperature was persistent and the water usage was relatively low.

4. Conclusions

This study developed a methodology constructing screening models for pipe failure events at household and community levels, to which the logistic and customized models were applied, respectively. Forty-nine independent variables were provided to the logistic regression model,

Table 5
Coefficients of the customized models developed for three areas of concern with the highest frequency of pipe failure events

Management areas	Coefficients	Estimates of coefficients	Standard error	95% confidence intervals for coefficients	
				Lower bound	Upper bound
DY01	m_0	0.306	0.298	−0.350	0.916
	m_1	0.011	0.003	0.005	0.017
	m_2	0.211	0.067	0.074	0.347
	m_3	0.007	0.004	−0.002	0.016
	m_4	−0.060	0.092	−0.248	0.128
	m_5	−0.047	0.027	−0.103	0.009
DY04	m_0	0.476	0.207	0.049	0.903
	m_1	0.000	0.002	−0.004	0.004
	m_2	0.007	0.039	−0.073	0.086
	m_3	0.011	0.004	0.002	0.020
	m_4	0.091	0.082	−0.078	0.261
	m_5	−0.019	0.020	−0.059	0.022
HJ01	m_0	1.447	1.150	−.922	3.815
	m_1	0.013	0.003	0.007	0.019
	m_2	0.211	0.063	0.081	0.341
	m_3	0.009	0.010	−0.012	0.030
	m_4	0.142	0.222	−0.314	0.598
	m_5	−0.006	0.007	−0.021	0.009

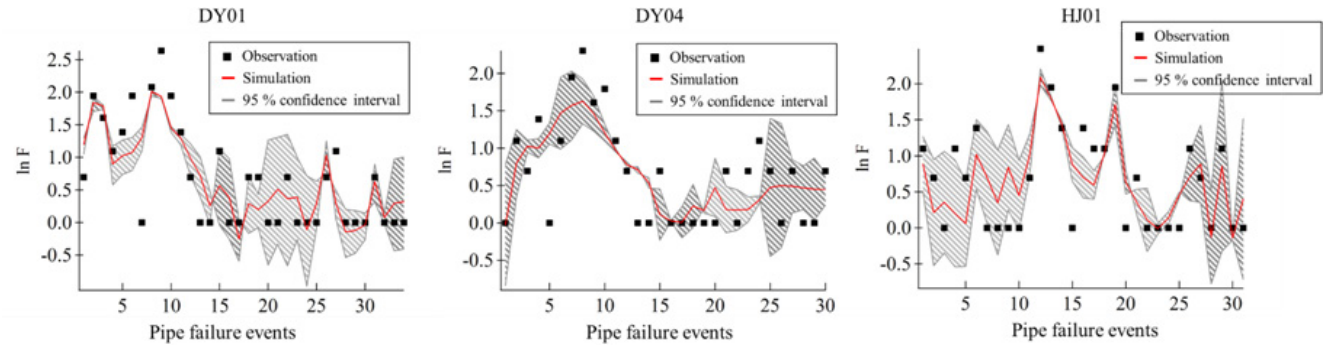

Fig. 4. Performance assessment of customized models developed at three areas of concern: DY01, DY04, and HJ01. Note that the shaded areas display the 95% confidence intervals.

including both antecedent and transformed variables,in addition to seven raw variables. Note that the occurrence of failures was used as a dependent variable and the customized model was designed to adopt variables only selected from the logistic model. The following conclusions were obtained from this study.

The logistic regression identifying the number of the failures at the household level recommended 11 models that contained different numbers of predictors ranging from two (as a minimum) to ten variables (as a maximum). Only two models had a VIF value lower than two, where the model including three variables performed better than that of two, with respect to *R*. The overall classification accuracy of the model developed with three variables was 66.4%.

The customized model developed at the community level showed good prediction accuracy for three test areas (DY01, DY04, and HJ01), which recorded the highest frequency of the failures. The coefficient of determination for the three models were estimated to be 0.53 for DY04, 0.59 for HJ01 and 0.66 for DY01. The alert system that classified the frequency of the failures into discrete risk levels (i.e., danger, warning and caution) successfully demonstrated in those areas using the outputs of the customized model.

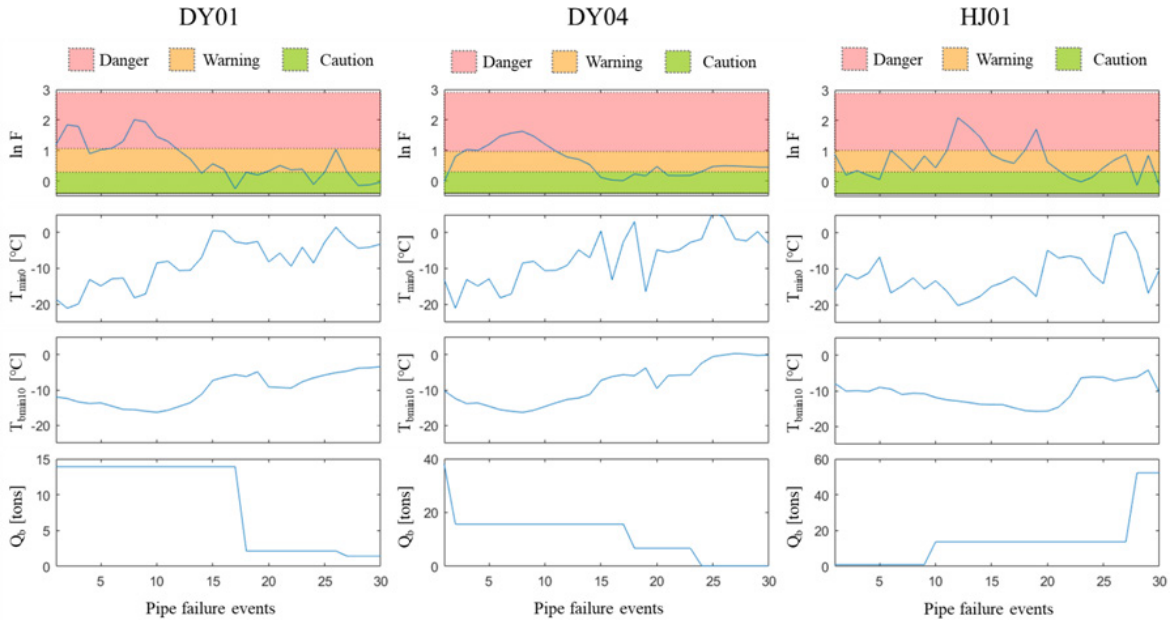

Fig. 5. The status of pipe network systems reported as three alert levels: danger, warning, and caution.

Therefore, the proposed screening model identified regions that were prone to pipe failures at different geographic scales. The model was also successfully used to modify existing models or indicators that assessed the risk of water network system failure using additional variables.

Acknowledgement

This work was supported by a GIST Research Institute (GRI) grant funded by the GIST in 2018. We would also like to thank K-water for providing the field data.

References

[1] Water Damage Studies, Insurance Institute for Business & Home Safety's (IBHS), Retrieved from https://disastersafety.org/ibhs/water-damage-studies/
[2] Freezing and Bursting Pipes, Insurance Institute for Business & Home Safety's (IBHS) guidance, Retrieved from https://disastersafety.org/freezing-weather/frozen-pipes/
[3] B. Rajani, Y. Kleiner, Comprehensive review of structural deterioration of water mains-physically based models, Urban Water, 3 (2001) 151–164.
[4] B. Rajani, C. Zhan, On the estimation of frost load, Canadian Geotech. J., 33 (1996) 629–641.
[5] C. Zhan, B. Rajani, Estimation of frost load in a trench: theory and experiment, Canadian Geotech. J., 34 (1997) 568–579.
[6] H. Rezaei, B. Ryan, I. Stoianov, Pipe failure analysis and impact of dynamic hydraulic conditions in water supply networks, Procedia Eng., 119 (2015) 253–262.
[7] M.J. Fadaee, R. Tabatabaei, Estimation of failure probability in water pipes network using statistical model, World Applied Sci. J., 11 (2010) 1157–1163.
[8] P. Gómez-Martínez, F. Cubillo, F. Martín-Carrasco, L. Garrote, Statistical dependence of pipe breaks on explanatory variables, Water, 9 (2017) 158.
[9] F.H. K, G.Y. Sagar, Statistical analysis of pipe breaks in water distribution systems in Ethiopia, the case of Hawassa, IOSR J. Math. (IOSR-JM), 12 (2016) 127–136.
[10] G. Kabir, S. Tesfamariam, A. Francisque, R. Sadiq, Evaluating risk of water mains failure using a Bayesian belief network model, Eur. J. Oper. Res., 240 (2015) 220–234.
[11] B.K. Riisnes, R. Ugarelli, Statistical models for structural reliability analysis of water mains, VannForeningen, 49 (2014) 483–491.
[12] E. Kimutai, Modelling Pipe Failure using Statistical models, in: Department of Civil and Construction Engineering, University of Nairobi, 2015.
[13] R.R. Fullwood, Review of Pipe-Break Probability Assessment Methods and Data for Applicability to the Advanced Neutron Source Project for Oak Ridge National Laboratory, in, Oak Ridge National Laboratory, 1989.
[14] U. Shamir, C.D.D. Howard, Analytical approach to scheduling pipe replacement, J. AWWA, 71 (1979) 248–258.
[15] T.M. Walski, A. Pelliccia, Economic analysis of water main breaks, J. AWWA, 74 (1982) 140–147.
[16] P. Jacobs, B. Karney, GIS development with application to cast iron water main breakage rate The 2nd international conference on water pipeline systems, Edinburgh, Scotland, 1994.
[17] Marks H.D. et al., Predicting urban water distribution maintenance strategies: A case study of New Haven Connecticut., in, US Environmental Protection Agency, 1985.
[18] Report on perspective of climate change in Korea, in, Korea Meteorological Administration.
[19] D.R. Cox, The regression analysis of binary sequences, J. Royal Stat. Soc. Ser. B (Methodological), 20 (1958) 215–242.
[20] H.M. Ramos, J. Ollero, A. Suárez-Llorens, A new explanatory index for evaluating the binary logistic regression based on the sensitivity of the estimated model, Stat. Probab. Lett., 120 (2017) 135–140.
[21] L. Murray, H. Nguyen, Y.-F. Lee, M.D. Remmenga, D.W. Smith, Variance inflation factors in regression models with dummy variables, in: 24th Annual Conference Proceedings Agriculture, 2012.