공학석사 학위논문

Susceptibility Model for Sinkholes Caused by Damaged Sewer Pipes Based on Logistic Regression

로지스틱 회귀 분석 기반 손상된 하수관에 의해 발생하는 지반함몰 위험도 예측 모델

2018년 2월

서울대학교 대학원

건설환경 공학부

김 기 연

Susceptibility Model for Sinkholes Caused by Damaged Sewer Pipes Based on Logistic Regression

지도 교수 정 충 기

이 논문을 공학석사 학위논문으로 제출함
2018 년 2 월

서울대학교 대학원
건설환경공학부
김 기 연

김기연의 석사 학위논문을 인준함
2018 년 2 월

위 원 장	박 준 범	(인)
부위원장	정 충 기	(인)
위 원	김 성 렬	(인)

Abstract

Susceptibility Model for Sinkholes Caused by Damaged Sewer Pipes Based on Logistic Regression

Kim, Ki Yeon

Department of Civil and Environmental Engineering

Master course

Seoul National University

The occurrence of anthropogenic sinkholes in urban area can cause serious social losses. A damaged and aged sewer pipes beneath the road contribute to occur such a phenomenon. This study used the best subsets regression method to develop a logistic regression model that calculate the susceptibility for sinkholes induced by damaged sewer pipes. The model was developed by analyzing both the sewer pipe network and cases of sinkholes in Seoul. Among numerous sewer pipe characteristics were analyzed as explanatory variables, the length, age, elevation, burial depth, size, slope, and materials of the sewer pipe were found to influence the occurrence of sinkhole. The proposed model reasonably estimated the sinkhole susceptibility in the area studied, with an area value under the receiver operating characteristics curve of 0.753. The proposed methodology will serve as a useful tool that can help local governments choose a cavity inspection regime and prevent sinkholes induced by damaged sewer pipes.

Keywords: sinkhole; susceptibility; damaged sewer pipe; logistic regression

Student Number: 2016-21242

Contents

List of Tables

List of Figures

Chapter.1 Introduction

1.1 General

A sinkhole is kind of ground failure defined by vertical deformation or the downward sinking of the ground caused by various natural phenomena and anthropogenic activities. The main cause of natural sinkholes is usually associated with dissolution of subsurface layer which mainly consists of limestone. The anthropogenic activities that cause subsurface failure and ground subsidence include sewer pipe damages, ground excavations, and groundwater extractions. Such anthropogenic sinkholes, which are irrelevant to the karst phenomena (e.g., solution sinkholes, collapse sinkholes), have been reported from many urban areas in the U.S., Italy, Japan, and South Korea (Galloway et al. 1999; Guarino and Nisio 2012; Yokota et al. 2012; Bae et al. 2016).

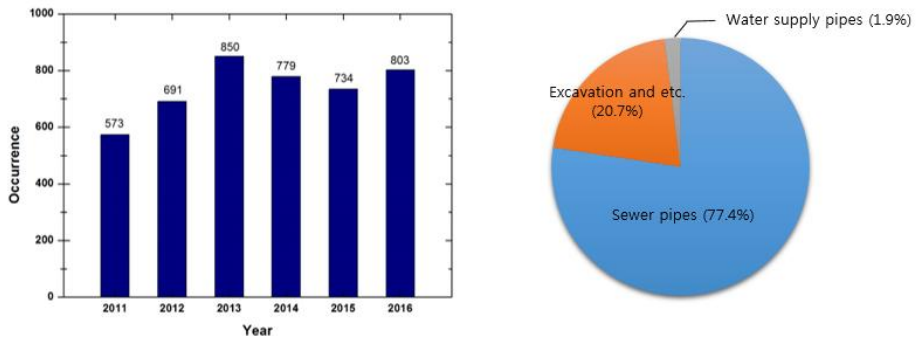

Figure 1-1-1 The number and causes of ground cave-ins in Seoul (Seoul Metropolitan Government, 2017)

In Seoul, 3,328 anthropogenic sinkholes occurred between 2011 and 2014, causing social anxiety. The primary cause of 81.4% of the aforementioned sinkholes was identified as sewer system failure (Bae et al. 2016). Yokota et al. (2012) also reported that about 17,178 sinkholes in Japan between 2006 and 2009 occurred due to sewer systems

1.2 Background

Rogers (1986) outlined mechanism of sewerage-system-induced sinkholes. The sewage infiltration to the ground occurs through a crack when the sewer pipe is full, especially during heavy rain season. As water level in the sewer pipe decreases, infiltration toward sewer pipe occur through the cracks of the pipe, that causes discharge of soil particles. This makes the ground loose, which results in underground cavities, and ground collapse and subsidence. Experiments on such sinkhole formation mechanism have been simulated by several researchers (Kwak et al. 2017; Kuwano et al. 2010; Mukunoki et al. 2009).

Seoul metropolitan government have installed tremendous number of underground sewer pipes up to 370,000 sewer pipes segments, which add up to approximately 10,000 km. The huge number of sewer pipes operating under the road makes sewer pipe inspection practically unrealizable because of limited budget. Identifying sewer pipes that are prone to sinkhole induced by damaged sewer pipe so that government can preferentially investigate high susceptible sewer pipes. Therefore, the government could perform sewer pipe inspection with minimized cost.

Logistic regression explains the relationship between a binary variable (i.e., presence or absence) and a set of predictor variables. Logistic regression could evaluate both susceptibility of hazard and the relative importance of each independent variable. Many researchers have calculated susceptibility for natural hazards such as landslides and sinkholes in karst terrain using this stochastic method. Ohlmacher and Davis (2003), Ayalew and Yamagishi (2005), and Yilmaz (2009) have applied logistic regression to propose a relationship between various influential factors and the occurrence of landslides, and to produce a landslide susceptibility map based on GIS (geographical information system). The sinkhole susceptibility in karst terrain has also been estimated based on logistic regression and GIS by some researchers (Ozdemir 2016; Ciotoli et al. 2016).

1.3 Aim and Scope of Study

The aim of this study is to propose a probabilistic model that can calculate the susceptibility of sinkholes induced by damaged sewer pipes and to apply developed model to make susceptibility map. The dataset in the model consists of sewer pipe network database in Seoul and sewer-pipe-induced sinkhole cases between 2010 January and 2014 July.

1.4 Outline

The present dissertation lays out the algorithm to calculate susceptibility of sinkhole induced by damaged sewer pipe and application of the model. In CHAPTER 2 presents the study area and study materials. And CHAPTER 3 explains the stochastic method and variables to develop probabilistic model. CHAPTER 4 discusses the developed model and the application of the model. CHAPTER 2, 3, 4 are closely related to the paper by Kim and Kim (2017) submitted in Natural Hazard and is under review at the time this dissertation was submitted. CHAPTER 5 summarizes the article and proposes conclusions and discusses future work related to the present dissertation.

Chapter.2 STUDY AREA AND MATERIALS

2.1 Study area

Seoul is located in the west-central part of the Korean Peninsula, lying between longitudes 126.8E and 127.2E, and latitudes 37.4N and 37.7N. Seoul consists of 25 administrative districts (Gus), which are subdivided into 522 sub-districts (Dongs). This metropolis, spanning over 605.2 km2, is populated by 10 million inhabitants. The mean elevation is 40 m above sea level. Although the mean annual precipitation is 1450.5 mm, much of the precipitation is concentrated during summer which causes soil discharges through damaged sewer pipes, leading to sinkhole occurrences (Korea Meteorological Administration 2011).

Seoul Metropolitan Government administers sewage pipes adding up to 10,570.5 km in length and 408.9 km2 in coverage. The sewage distribution rate, which is the ratio of the sewage-serviced population to the total population, reaches 100.0 % (Seoul Metropolitan Government, 2016). Table 2.1 illustrates the status quo of the sewage pipes and extensions in Seoul. It shows that most of the sewer pipes are unclassified pipes, meaning that the sewer and the storm water flow unseparated. It can also be seen that most of the pipes fall under the category of culvert, which is a conduit topped with covers to prevent the release of the odor. The pipes that are responsible for every sinkhole case are culverts, which this study focused on.

Table 2.1 Classification and extension of the sewer pipe line in Seoul (Seoul Metropolitan Government, 2016)

	Classified Pipe (km)		Unclassified Pipe (km)
	Sanitary sewer	Storm sewer	
Culvert	533.9	248.7	9671.9
Open ditch	-	50.6	0.1
Gutter	-	63.9	1.4

2.2 Data Source

Seoul runs 374,612 sewer pipes, whose information are organized and provided by Seoul Metropolitan Government (2015). The database provides 58 attributes for each sewer pipe, among which only 8 were appropriate for logistical regression analysis considering the data availability. Subsequently, 8 attributes — length, age, elevation, burial depth, equivalent radius, slope, cross-sectional shape, and material — were used as independent variables for the logistic regression model. Then the dataset underwent a screening process for the exclusion of the pipes containing null data and outliers. For example, some pipes had missing installation dates or pipe materials, and others had negative burial depths or zero lengths. A total of 158,424 sewer pipes were found to have null values or outliers in their attributes and were thus excluded from the statistical analysis.

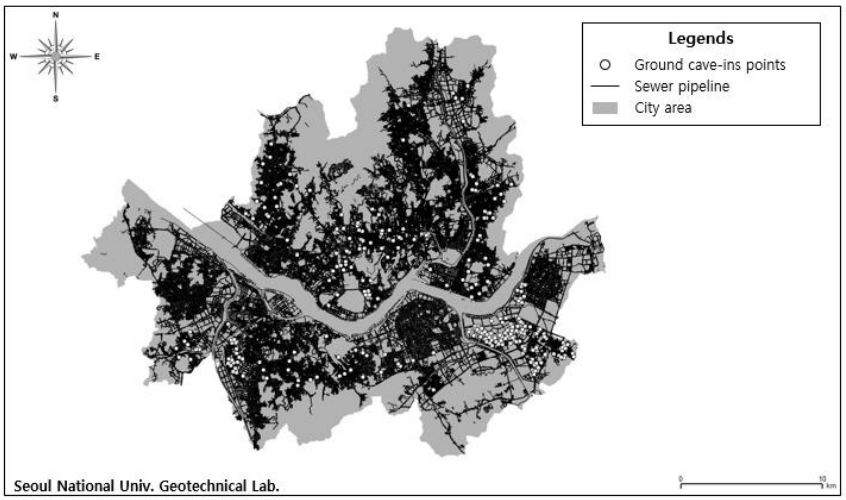

Figure 2-0-1 Seoul metropolitan sewer pipeline network and ground cave-

ins occurrences

Of the 3,119 sinkhole cases between January 2010 and July 2014, it was confirmed by the investigations that 1,173 cases were caused by damaged sewer pipes. For the 1,173 cases of damaged-sewer-pipe-induced sinkhole, the information on each case, including the location, size, and date, were documented by Seoul Metropolitan Government (unpublished data, 2015).

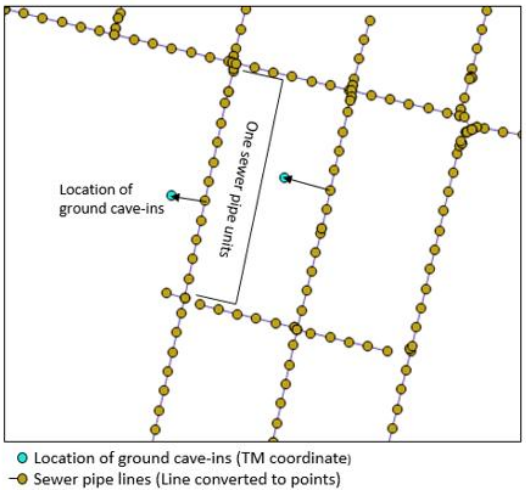

Figure 2-0-2 Finding process sewer pipe that contribute to occur ground cave-ins

The locations of the sinkholes, however, were recorded in the street address system whereas the locations of the sewer pipes were given in the TM (Transverse and Mercator) coordinate system. Therefore, the location of each sinkhole was converted to the TM coordinate system using Google Maps so that all the geographical information could be presented in an identical format. Then, among the multiple pipes which

traverse nearby land around the sinkhole, the nearest sewer pipe identified using QGIS — an open-source GIS software (QGIS Development Team 2014) — and was assumed to have induced that sinkhole. In some cases, however, it was not possible to specify a single sewer pipe responsible for a given sinkhole. In this situation, the sinkhole occurrence case and the nearby sewer pipes were excluded from the dataset altogether.

Finally, the sewer pipes responsible for the sinkholes were assigned to the sinkhole occurrence group (442 pipes), with a value of 1 for the binary dependent variable. The remaining sewer pipes were assigned to the non-occurrence group (215,955 pipes), with a value of 0 for the variable. Figure 1 illustrates the locations of sinkhole that were used in this study along with the sewer pipe network in Seoul.

Chapter.3 STATISTICAL ANALYSIS

3.1 Logistic Regression

Statistical analysis is one of the methods most applicable for regional-scale susceptibility assessment due to the simplicity of its implementation, updating, and quantitative results provision. In particular, multivariate statistical techniques such as logistic regression and discriminant analysis are the most frequently used (He and Beighley 2008). The advantage of logistic regression over discriminant analysis is that in the former, the independent variables can be either continuous or categorical, or any combination of both. Moreover, the assumption of multivariate normality is rarely satisfied in reality, even approximately; logistic regression calls for such assumption, while discriminant analysis does not. (Truett et al. 1967). As such, logistic regression has been one of the most important models for binary response data, and has been used for a wide variety of applications (Agresti 2003).

The threshold model for binary generalized linear models (GLMs) is generally expressed as equation (1),

$$\pi_i = P(\mathrm{Y}_i = 1) = F\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) \text{ and } F^{-1}(\pi_i) = \sum_{j=1}^{p} \beta_j x_{ij} \quad (1)$$

where Y_i is the binary dependent variable, x_{ij} are the independent variables, β_j are the estimated regression coefficients, and π_i is the probability of occurrence for i_{th} observation. Logistic regression uses the following

standard cumulative distribution function as a link function:

$$F(Z)=\frac{e^{Z}}{1+e^{Z}} \quad (2)$$

Therefore, logistic regression model formulas can be expressed according to the following equation:

$$\tilde{\mathrm{Y}}=\frac{\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)} \quad (3)$$

where $\hat{\mathrm{Y}}$ is the predicted probability of being in one particular category of Y.

The coefficients β_j are determined through maximum likelihood estimation (MLE). Likelihood function, which represents the objective information gained from the observations, is proportional to the conditional probability of the given Y, as expressed in equation (4).

$$L(\boldsymbol{\beta}|\mathbf{Y})=\prod_{i=1}^{n}\left[\pi_i^{Y_i}\cdot\left(1-\pi_i\right)^{(1-Y_i)}\right] \quad (4)$$

Therefore, the likelihood function of binary logistic regression can be expressed by

$$L(\boldsymbol{\beta}|\mathbf{Y})=\prod_{i=1}^{n}\left[\left(\frac{\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right)^{Y_i}\left(1-\frac{\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)}\right)^{(1-Y_i)}\right] \quad (5)$$

and the likelihood function is transformed to log-likelihood function to make a linear combination.

$$\log L(\boldsymbol{\beta}|\mathbf{Y}) = \sum_{i=1}^{n} \log \left[\left(\frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)} \right)^{Y_i} \left(1 - \frac{\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)}{1+\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)} \right)^{(1-Y_i)} \right] \quad (6)$$

MLE method estimates the coefficients β_j, by maximizing the log-likelihood function for a given set of variables.

In the model building process, this study adopted the best subsets logistic regression method, which considers all the possible combinations of independent variables and selects the fittest model based on some goodness-of-fit criteria (Hosmer et al. 1989). As a measure of goodness-of-fit, AIC (Akaike information criteria) was used, which can be calculated from equation (7), in which k is the number of independent variables in a model and $\hat{L}$ is the maximized value of the likelihood function for the model (Akaike 1974).

$$AIC = 2k - 2\ln(\hat{L}) \quad (7)$$

Equation (7) infers that AIC includes a penalty on an increasing number of independent variables in a model to avoid overfitting, and that the optimal model is the one with the minimum AIC value for a given dataset.

3.2 Variables in the Model

This study focused on the prediction of sinkhole occurrence (dependent binary variable) as a function of multiple sewer pipe characteristics (independent variables). Due to the dearth of research on the explanatory power of specific sewer pipe characteristics, all the available information in the aforementioned dataset were considered in evaluating a logistic regression model.

As for the continuous variables, six sewer pipe parameters — length, elevation, burial depth, slope, size (i.e., radius for circular shape, and width B and height H for rectangular shape), and age — were available. In order to numerically account for the different cross-sectional shapes, the equivalent radius re was computed from the width and height of the rectangular sewer pipe, as follows:

$$r_e = \sqrt{\frac{BH}{\pi}} \quad (8)$$

The elevation refers to the distance from the sea level whereas the burial depth refers to the distance from the upper pavement. Also, the slope, which equals the ratio of elevation difference between the ends of the pipe to the pipe length, is expressed in parts per thousand. Table 3.1 and Figure 3-1 show the descriptive statistics and histogram of the continuous variables included in the logistic regression model, respectively. It can be seen that some of the sewer pipes reach 200 meters because the sewer network database assigns a single identification when

sewer segments with identical characteristics are continuously connected to one another.

Table 3.1 Descriptive statistics of the continuous variables

Variables	Mean	Median	Standard Deviation	Min	Max
Equivalent radius (m)	0.3328	0.3000	0.2430	0.10	2.4463
Length (m)	30.50	28.08	21.9605	0.11	199.75
Elevation (m)	24.54	20.43	16.3483	0.14	226.66
Slope (‰)	32.49	14.42	48.3906	0.00	399.61
Burial depth (m)	0.9935	0.8500	0.6190	0.01	17.76
Age (years)	24.16	24.00	12.6763	0.00	82

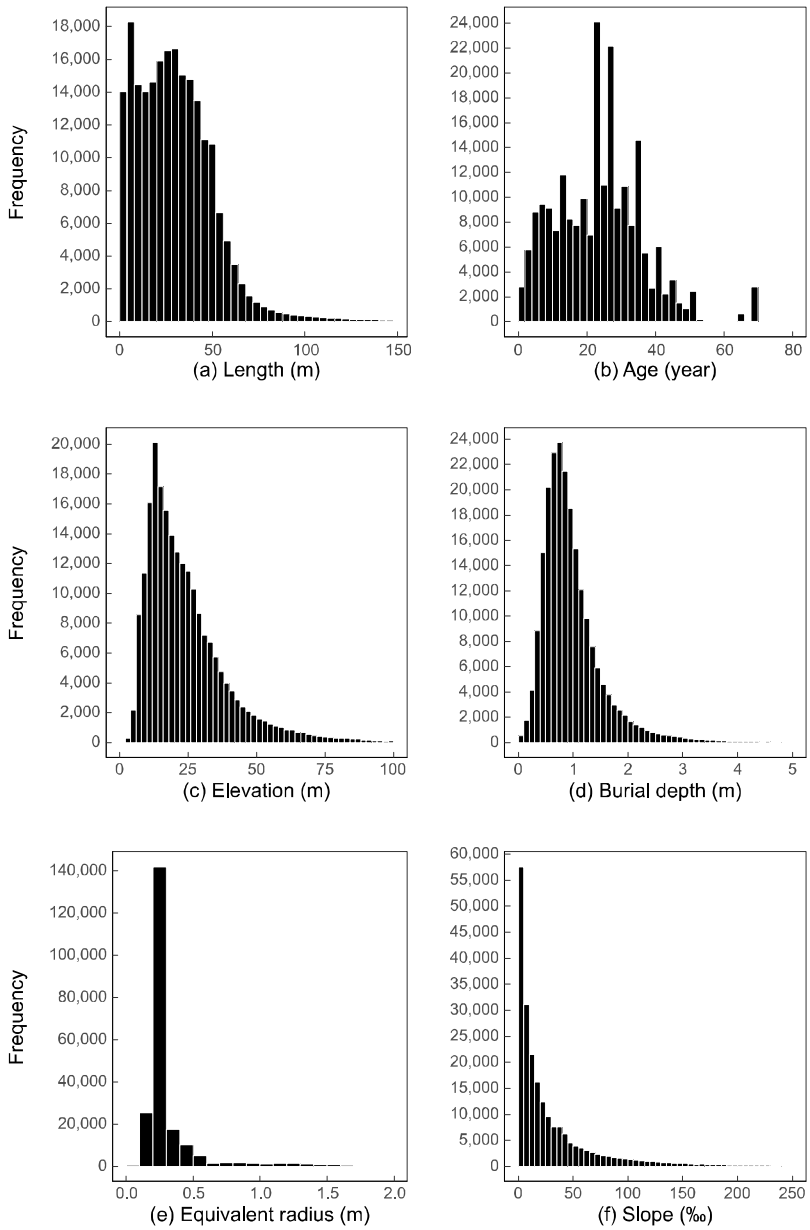

Figure 3-1 Histograms of the continuous

For the categorical variables, the pipe material and cross-sectional shape were available. As shown in the frequency table of categorical variables (Table 3.2), a great majority (87%) of the pipes were Hume pipes. The materials of the remaining sewer pipes, with a frequency of less than a thousand each, including polyethylene (PE), polyvinyl (PVC), fiberglass, etc., were grouped as "others." As for the cross-sectional shape of the sewer pipe, most of the pipes (92.6%) had a circular shape.

Table 3.2 Frequency table of categorical variables

Variable		Frequency
Material	Hume	187,976
	Concrete	18,351
	Wrinkle	4,283
	Others	5,578
Cross-sectional shape	Circular	200,111
	Rectangular	16,077

When performing regression analysis, multicollinearity can be a serious problem. Multicollinearity occurs when a variable can be explained by the other variables in the analysis, leading to unreliable and unstable estimates of the regression coefficients. Moreover, as multicollinearity increases, it becomes more difficult to ascertain the effect of any single variable due to the variables' interrelationships (Hair et al.

2006). Multicollinearity can be detected with the help of the variance inflation factor (VIF), which is the most widely used diagnostic index. Notwithstanding the differing opinions, a VIF larger than 5 is generally a cause for concern, and a VIF larger than 10 certainly indicates a serious collinearity problem (Menard, 1995). In this study, the cross-sectional shapes with the largest VIF and with a VIF of more than 5 were omitted from consideration. After dropping the cross-sectional shape from among the independent variables, the VIFs of the remaining variables ranged from 1.01 to 2.30, as shown in Table 3.3 indicating very low dependence between the variables. Consequently, the length, age, elevation, burial depth, equivalent radius, and slope, together with the pipe materials (i.e., Hume, concrete, wrinkle, and “others”) were used for the logistic regression modeling.

Table 3.3 Variance inflation factor (VIF) for the independent variables

Input Variables	Before Removal		After Removal	
	VIF	Multicollinearity	VIF	Multicollinearity
Cross-sectional shape	8.0879	Serious	-	-
Material: Wrinkle pipe	1.0110	Lack	1.0099	Lack
Material: Concrete pipe	6.6517	Serious	2.2000	Lack
Material: Others	1.0403	Lack	1.0363	Lack
Length	1.0742	Lack	1.0740	Lack
Age	1.0497	Lack	1.0495	Lack
Elevation	1.1963	Lack	1.1958	Lack
Burial depth	1.0439	Lack	1.0434	Lack
Equivalent radius	2.7887	Lack	2.2946	Lack
Slope	1.1940	Lack	1.1940	Lack

Chapter.4 RESULTS AND DISCUSSION

4.1 Model Development

This study applied the best subsets logistic regression method to select the fittest model for assessing sinkhole susceptibility. R, a language and environment for statistical computing (R core team 2017), was used in this procedure. As for categorical variables, the Hume pipe was fixed as the reference category while the other three were set as indicator (dummy) variables. Thus, the best subsets logistic regression method considers 512 models from the 9 independent variables.

Table 4.1 summarizes the list of independent variables selected by the top three candidate models, which were ranked based on the AIC value. These models included the length, age, elevation, burial depth, equivalent radius, and pipe materials (with the exception of the “others” group) in common. While Model II and Model III, respectively, excluded “others” in the material and slope of the sewer pipe, the fittest model (Model I) included all the independent variables that were considered in this study. The AIC relatively presents how far the model in question deviates from the true model. so it can prove nothing about the quality of the model in an absolute sense. Thus, the statistical significance of each estimated coefficient in the candidate models should be evaluated using the Wald test, where the coefficient is divided by its standard error (SE) (Tabachnick and Fidell 2001):

$$z_j = \frac{\beta_j}{SE_{\beta_j}}$$

The null hypothesis of the Wald test is that the coefficient is zero; as such, the variables with estimated coefficients whose significance probability (p) is more than 0.05 were not found to be significantly different from zero in the 95% confidence interval, and can thus be excluded from the model. Otherwise, the variables should be accepted in the model as influential predictors.

Table 4.1 Independent variables of the top three candidate models based on AIC.

Independent Variables	Model I	Model II	Model III
Material: Hume	O	O	O
Material: Concrete pipe	O	O	O
Material: Wrinkle pipe	O	O	O
Material: Others	O	X	O
Length	O	O	O
Age	O	O	O
Elevation	O	O	O
Burial depth	O	O	O
Equivalent radius	O	O	O
Slope	O	O	X
AIC value	6026.2	6028.2	6031.9

In candidate Model I, the “others” category in the pipe material variable has a significance probability of more than 0.05, whereas the other independent variables have a significance probability of less than 0.05, as shown in Table 4.2

Table 4.2 Independent variables retained in Model I with their coefficient

Variables	β	Standard Error	Wald test (z)	p
(Intercept)	7.0057	0.2163	-32.393	$< 10^{-16}$
Material : Concrete pipe	2.4522	0.4085	-6.002	1.94×10^{-9}
Material : Wrinkle pipe	0.8552	0.2527	3.384	7.13×10^{-4}
Material : Others	0.9756	0.5819	-1.677	9.36×10^{-2}
Length	0.0239	0.0015	15.710	$< 10^{-16}$
Age	0.0199	0.0035	5.689	1.28×10^{-8}
Elevation	0.0208	0.0047	-4.390	1.14×10^{-5}
Burial depth	0.2520	0.0929	-2.713	6.66×10^{-3}
Equivalent radius	1.0110	0.3529	2.865	4.17×10^{-3}
Slope	0.0047	0.0018	-2.541	1.10×10^{-2}

Table 4.3 shows the independent variables retained in candidate Model II with their coefficients. The significance probability of each estimated coefficient was found to be less than 0.05, which demonstrates that all the predictors are statistically significant in the given dataset. Consequently, candidate Model II, which has the second lowest AIC value and whose logistic coefficients are all statistically significant with respect to the Wald test, is the optimal choice. The logistic regression equation of Model II can be expressed as equation (10).

$$\text{logit } \hat{Y}_i = -7.0299 - (2.4600 \text{ Material: Concrete pipe}) + (0.8710 \text{ Material: Wrinkle pipe}) + (0.0239 \text{ Length}) + (0.0205 \text{ Age}) - (0.0208 \text{ Elevation}) - (0.2651 \text{ Burial depth}) + (1.0306 \text{ Equivalent radius}) - (0.0046 \text{ Slope}) \quad (10)$$

Table 4.3 Independent variables retained in Model II with their coefficients

Variables	β	Standard Error	Wald test (z)	p	β*
(Intercept)	-7.0299	0.2153	-32.646	$< 10^{-16}$	
Material : Concrete pipe	-2.4600	0.4079	-6.031	1.63×10^{-9}	-0.0327
Material :Wrinkle pipe	0.8710	0.2526	3.449	5.64×10^{-4}	0.0058
Length	0.0239	0.0015	15.751	$< 10^{-16}$	0.0250
Age	0.0205	0.0035	5.897	3.69×10^{-9}	0.0124
Elevation	-0.0208	0.0047	-4.391	1.13×10^{-5}	-0.0161
Burial depth	-0.2651	0.0923	-2.872	4.08×10^{-3}	-0.0078
Equivalent radius	1.0306	0.3502	2.943	3.25×10^{-3}	0.0119
Slope	-0.0046	0.0018	-2.540	1.11×10^{-2}	-0.0107

4.2 Interpretation of Estimated Coefficients

According to Table 4.3 estimated coefficients (β) of length, age, equivalent radius, and wrinkle pipe in Model II are positive, whereas the coefficients of elevation, burial depth, slope and concrete pipe material are negative. A positive regression coefficient increases the probability of failure as its predictor variable increases, whereas a negative coefficient decreases the probability as its predictor increases. Thus, the probability of sinkhole occurrences increases with the rise in each of the following factors: length, age, and equivalent radius. On the other hand, an increase in any one among elevation, slope, and burial depth decreases the probability of failure. As for the categorical variables, the wrinkle pipe increases whereas the concrete pipe decreases the probability of sinkhole occurrences compared to the Hume pipe, the reference category. The following paragraphs discuss the relationship between the probability of sinkhole occurrences and each of the sewer pipe parameters retained in the predictive model.

4.2.1 Length

As previously mentioned, the sewer network database of Seoul assigns a single identification number when sewer segments with identical characteristics are continuously connected to one another. It is obvious that a long sewer pipe with multiple segments contains more defects and cavities than that with few segments. Moreover, based on the fact that the

length of a sewer segment is restricted in practice, a long sewer pipe has numerous pipe joints through which groundwater infiltration with movement of soil into a sewer is most likely (Fenner 1990).

4.2.2 Age

As discussed in the review paper of Davies et al. (2001), the UK sewerage system had more structural defects in the older sewers, with a break point in the trend at the end of World War II (O'reilly et al. 1989), while those sewers that had been constructed during the war period (1918 – 1939) had the highest defect rate (Lester and Farrar 1979). The age of the sewerage system in Japan and the sinkhole occurrence in the same country showed similar trends. The number of sinkhole cases per 100 kilometers linearly increased with the age of the sewer system until the 1950s, and no tendencies were found thereafter (Yokota et al. 2012). The sewer pipes, which were used in the present analysis, were laid after the Korean War. Thus, as expected, a positive relationship between the age of the sewer pipes in Seoul and the sinkhole probability was demonstrated. There are a number of possible reasons for this, but the most likely is that the older sewers have had greater exposure to the corrosive action of hydrogen sulphide and the external surface load due to the traffic. The advances in the pipe materials and the improved pipe installation techniques can also reduce the deficiencies in the recently installed pipes.

4.2.3 Equivalent radius

A number of surveys conducted in the UK have provided contradictory results in the relationship between sewer size and structural soundness (Davies et al. 2001). Several sewer deterioration models based on logistic regression analysis resulted in a negative coefficient of sewer size, implying that the larger sewers are more structurally sound (Ariaratnam et al. 2001, Davies et al. 2001b).

For sinkholes to occur, however, the cracks and apertures must reach a certain size for the soil particles to be discharged through the defects. When cracks and apertures are small, soil resistance (e.g., soil arching) can override driving forces, disallowing soil discharge. Furthermore, sewer pipes with a large radius are likely to have large-enough cracks and fractures needed for soil infiltration. It is believed that additional data on the defect size and equivalent radius of sewer pipes, which were not available for use in this study, will provide evidences for this finding.

4.2.4 Elevation

Few studies on pipe deterioration models have considered elevation as an input variable. Elevation, however, remained in the proposed logistic regression model, presenting a negative relationship with sinkhole occurrence. This is related to the fact that flooding during heavy rainfall usually occurs at lower altitudes. Storm water gathers at lower altitudes,

resulting in overflows of the sewers and a rise in the ground water level, which are both favorable conditions for soil infiltration according to Rogers (1986).

4.2.5 Slope

The higher the slope is, the higher the velocity, which leads to the greater subsequent abrasion of the sewer walls. Conversely, the milder the slope is, the lower the velocity, which allows for possible the sedimentation of the solid material and obstructions (Giudice et al. 2016). As a result, the sewer pipes with a high slope are more likely to become deficient. In addition, a high flow rate can accelerate the soil infiltration, especially during and after a heavy rain; thus, it is expected that sinkholes are prone to occur under a high-slope environment. The proposed model, however, predicted a decreasing failure probability with the increasing slope of the sewer pipe. It is expected that the slope has an influence on the soil-pipe interaction, and subsequently, the soil infiltration through an aperture, although further investigations are required.

4.2.6 Burial depth

In order to protect pipes from external loads, most of the codes and regulations regarding the construction of sewers stipulate the minimum burial depth of sewer pipes, which vary depending on the pipe type, pipe size, and use of the land above the pipe (BS EN 1610; CFR 195.248). Ariaratnam et al. (2001) showed that the average depth of the cover is not

significant to pipe deficiency, whereas Davies et al. (2001b) found that the burial depth conveyed less information after accounting for the sewer size. Both studies excluded the burial depth from their sewer deterioration model. For the sewer pipes that had been laid in Seoul, however, there is no meaningful collinear relationship between the sewer size and the burial depth according to the VIF values presented in Table 3.3

Kwak et al. (2017) examined the effect of the burial depth on sinkhole formation and found that a larger amount of soil infiltrated with a horizontally larger cavity, given that the burial depth was shallow. Moreover, grounds with shallow burial depth easily collapse, even with minor cavities. The proposed model also inferred that the sinkhole probability increases with decreasing burial depth. This finding supports the necessity of abiding by the minimum burial depth standards.

4.2.7 Significance

The relative importance of predictors in a logistic regression model can be evaluated using fully standardized logistic regression coefficients calculated using Equation (11) (Menard 2011):

$$\beta_i^* = \beta_i \, \sigma_{xi} \, R/\sigma_{\text{logit}(\hat{Y})} \qquad (11)$$

where β_i^* is the standardized logistic regression coefficient, β_i is the unstandardized logistic regression coefficient, σ_x is the standard deviation of independent variable x, $\sigma_{\text{logit}(\hat{Y})}$ is the standard deviation of logit($\hat{Y}$), [i.e.,

the standard deviation of the predicted values of logit(Y)], and R is the correlation between the observed values of Y (either 0 or 1) and the predicted values of Y.

As shown in Table 4.3, the most influential variable was concrete pipe under the categorical variable. Although the Hume pipe, the reference category, is the traditional reinforced-concrete cement pipe, the failure probability of concrete pipe was greatly reduced compared to that of Hume pipe. The statistical analysis that was performed in this study could not adequately explain the mechanism behind this result.

Following concrete pipe under the categorical variable, other influential factors were revealed to be length, elevation, age, equivalent radius, slope, and burial depth, in descending order of their influential power. It should be noted that there were no large differences in importance between age and the other variables, and contrary to the expectations, elevation was even more influential than age. This implies that the current cavity exploration scheme, which is mainly based on the age of the sewer pipes, needs modification considering the other influential parameters.

4.3 Model validation

The accuracy of a binary logistic regression model is assessed in terms of the probability that the model correctly classifies a non-occurrence case as negative, namely the true negative rate (specificity), and the probability that the model correctly classifies an occurrence case as positive, namely the true positive rate (sensitivity) (Zou et al. 2007). Determining a threshold (cut-off) value as the observed probability, the proposed regression model classified the sewer pipes in Seoul as shown in Table 4.4. The validation showed that the model correctly classified 65.1% as having a value of 0 for the dependent variable (sinkhole absence), and 73.1% as having a 1 value (sinkhole presence), with an overall correct classification percentage of 65.1%.

Table 4.4 Contingency table of the Model 2

Observed		**Predicted**		
		Sinkhole		Correct percentage
		Non-occurrence 0	Occurrence 1	
Non-occurrence	0	140,47	75,29	65.1
Occurrence	1	119	323	73.1
Overall percentage				65.1

*Threshold probability: 2.045×10^{-3}

The receiver-operating characteristics (ROC) curve plots the true positive rate against the false positive rate for different cut-off values. ROC analysis is a useful tool for evaluating the accuracy of a statistical model (e.g., logistic regression, discriminant analysis) that classifies the dependent variable into several categories. It has been widely used by numerous researchers to validate their proposed logistic regression model predicting landslides or sinkholes (Dou et al. 2015; Ciotoli et al. 2016; Ozdemir 2016). Figure 4-1 Receiver-operating characteristic (ROC) curve of the proposed Model 2, which is the ROC curve corresponding to random choice. The area under the ROC curve (AUC), which is an overall summary of a predictive model's accuracy, was 0.753, indicating that the proposed model reasonably estimated the sewer-induced sinkhole susceptibility in Seoul Metropolitan City.

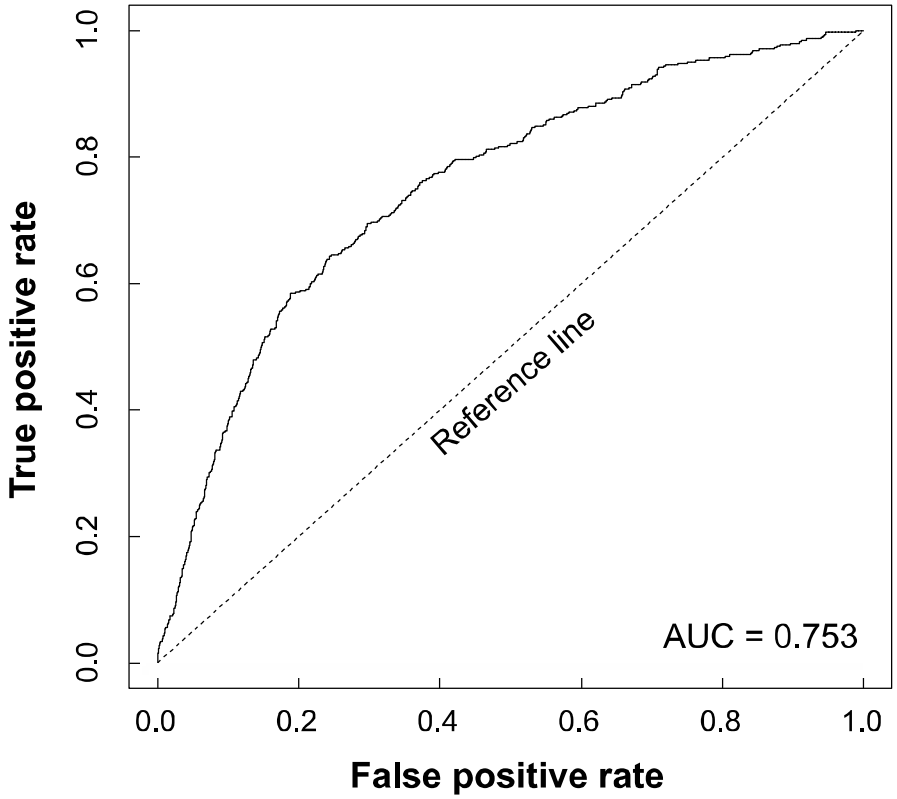

Figure 4-1 Receiver-operating characteristic (ROC) curve of the proposed Model 2

4.4 Model application

The susceptibility to hazards estimated by a statistical model is usually divided into several classes to help the decision maker schedule inspections and come up with failure prevention measures (Ayalw and Yamagishi 2005; Van Den Eeckhaut et al. 2006; Pourghasemi et al. 2013). In this study, the susceptibility to anthropogenic sinkholes was classified into five different categories using the Fisher-Jenks natural-breaks algorithm (Slocum 2009) with respect to the predicted probability: extremely low, low, moderate, high, and extremely high susceptibility. The Fisher-Jenks method divides classes through an iterative algorithm so that variances within classes are minimized and variances between classes are maximized.

As anthropogenic sinkholes are essentially rare phenomena, most of the sewer pipes (96.1 %) were classified as having extremely low or low susceptibility, as shown in Table 4.5. Those locations with moderate to extremely high susceptibilities to anthropogenic sinkhole, to which special attention should be paid, accounted for only 3.9 % of the total study area.

Table 4.5 Classification of anthropogenic sinkhole susceptibility

Probability	Susceptibility	Percentage of the Study Area
$- 2.209 \times 10^{-3}$	Extremely low	65.9
$2.209 \times 10^{-3} - 5.261 \times 10^{-3}$	Low	30.2
$5.261 \times 10^{-3} - 1.461 \times 10^{-2}$	Moderate	3.47
$1.461 \times 10^{-2} - 3.716 \times 10^{-2}$	High	0.36
$3.715 \times 10^{-2} -$	Extremely high	0.07

The resulting susceptibility class of each sewer pipe and the surrounding subsurface can be used in planning the inspection strategies. Because surveying projects are performed region by region, it is necessary to prioritize certain regions with high risks of sinkhole. Although a number of strategies are available for this, setting the priorities based on the ratio of critical sewers to the total sewers located in a district can be a reasonable option. And we made the susceptibility map for sinkhole that can induced by damaged sewer pipes

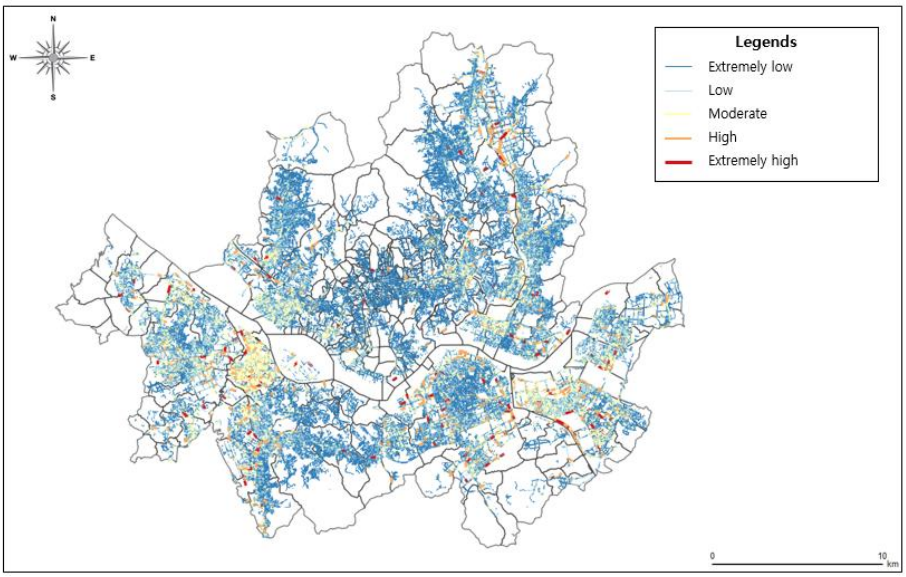

Figure 4-2 susceptibility map of ground cave-ins in Seoul

Chapter.5 Conclusions

In Seoul, sinkhole due to damaged sewer pipe frequently occur and have been reported. This study used best subset regression method to find the fittest model estimating susceptibility of the presence or absence of sinkholes. The stochastic method was based on the data set that consists of 216,188 sewer pipes and 442 sinkhole cases in Seoul. The dataset includes 11 sewer pipe characteristics which were used to develop the probabilistic model, among which 9 variables eventually were in the fittest model (length, elevation, age, equivalent radius, slope, burial depth, and pipe materials: Hume, concrete, and wrinkle pipe). The length, age, and equivalent radius of sewers were proven to have significantly positive relationship. On the other hands, the elevation, slope, and burial depth had negative relationships. The

fully standardized regression coefficients imply that other factors than age are also influential. The proposed model was also validated to reasonably predict sinkhole occurrence and provides useful criteria that could help Seoul governments identify a sewer pipe that is prone to sinkhole based on the calculated susceptibility.

The primary advantage of the proposed model and framework is that it relies only on the sewer pipe parameters, of which there are relatively well-established databases in almost big cities. Conversely, it excludes the geological and geotechnical properties of the ground where sewer is buried, which serve as key factors contributing to soil infiltration through the damaged pipes. Further studies incorporating the geological and geotechnical properties into the logistic regression model are expected to improve the predictability of sinkhole occurrence.

References

Akaike H (1974) A new look at the statistical model identification. IEEE transactions on Automatic Control 19:716-23. doi: 10.1109/TAC.1974.1100705

Agresti A, Maria K (2003) Categorical data analysis. 2nd edn. John Wiley & Sons, Hoboken, New Jersey, pp 206-2008

Ariaratnam ST, El-Assaly A, Yang Y (2001) Assessment of Infrastructure Inspection Needs Using Logistic Models. Journal of Infrastructure Systems 7:160-165

Ayalew L, Yamagishi H (2005) The application of GIS-based logistic regression for landslide susceptibility mapping in the Kakuda-Yahiko Moutains, Central Japan. Geomorphology 65:15-31

Bae Y, Shin S, Won J, and Lee D (2016) The road subsidence conditions and safety improvement plans in Seoul. The Seoul Institute, Seoul, South Korea, Report 2016-PR-09 (in Korean)

Baur R, Herz R (2012) Selective inspection planning with ageing forecast for sewer types. Water science & Technology 46:389-96

Beradi L, Giustollisi O, Kapelan Z, Savic DA (2008) Development of pipe deterioration models for water distribution systems using EPR. Journal of Hydroinformatics 10:113-126

BSI (1998) EN 993-2:1996. Construction and testing of drains and sewers. BSI, London, UK.

Ciotoli G, Loreto ED, Finoia MG, Liperi L, Meloni F, Nisio S, Sericola A (2016) Sinkhole susceptibility, Lazio Region, central Italy. Journal of Maps 12:287-294. doi:10.1080/17445647.2015.1014939

Code of Federal Regulations 49 CFR 195.248. Cover over buried pipe line

Davies JP, Clarke BA, Whiter JT, Cunningham RJ (2001a) Factor influencing the structural deterioration and collapse of rigid sewer pipes. Urban Water 3:73-89

Davies JP, Clarke BA, Whiter JT, Cunningham, RJ, Leidi A (2001b) The structural condition of rigid sewer pipes: a statistical investigation. Urban Water 3:277-286

Del Giudice G, Padulano R, Siciliano D, (2016) Multivariate probability distribution for sewer system vulnerability assessment under data-limited conditions. Water Science and Technology 73: 751-760

Dou J, Bui DT, Yunus AP, Jia K, Song X, Revhaug I, Zhu Z (2015) Optimization of causative factors for landslide susceptibility evaluation using remote sensing and GIS data in parts of Niigata, Japan. PloS one 10:e0133262

Egger C, Scheidegger A, Reichert P, Maurer M (2013) Sewer deterioration modeling with condition data lacking historical records. Water Research 47:6762-6779

Fenner RA (1990) Excluding groundwater infiltration into new sewers. Water and Environment Journal 4:544-551

Galloway D, Jones DR, Ingebritsen SE (1999) Land sunbsidence in the United States. U.S. Geological Survey. Circular 1182

Guarino PM, Nisio S (2012) Anthropogenic sinkholes in the territory of the city of Naples (Southern Italy). Physics and Chemistry of the Earth 49:92-102

Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2006) Multivariate data analysis. Upper Saddle River, 6th edn. NJ: Prentice hall, pp 557

He Y, Beighley RE (2008) GIS-based regional landslide susceptibility mapping: a case study in southern California. Earth Surface Processes and Landforms 33:380-393. doi: 10.102/sep1562

Hosmer D, Jovanovic B, Lemeshow S (1989) Best Subsets Logistic Regression. Biometrics 45:1265-1270. doi:10.2307/2531779

Korea Meteorological Administration (2011) Regional climate change report on Seoul, Seoul, South Korea (in Korean)

Kwak TY, Kim KY, Lee MH, Chung CK, Kim J (2017) Evaluation of the effect of burial depth and rainfall intensity on ground cave-in induced by a damaged sewer pipe. Proc., the 70th Canadian Geotechnical Conference and the 12th Joint CGS/IAH-CNC Groundwater Conference, Ottawa (in press)

Kuwano R, Horii T, Yamaguchi K, Kohashi H (2010) Formation of subsurface cavity and loosening due to defected sewer pipe. Japanese Geotechnical Journal 5:349-361

Lester J, Farrar DM (1979) An examination of the defects observed in 6 km of sewers. TRRL Supplmentary Report 531

Menard S (1995) Applied Logistic Regression Analysis. Thousand Oaks, California

Menard S (2011) Standards for standardized logistic regression coefficients. Social Forces 89:1409-1428

Mukunoki T, Kumano N, Otani J, Kuwano R (2009) Visualization of three dimensional failure in sand due to water inflow and soil drainage from defected underground pipe using X-ray CT. Soils and Foundations 49:959-968

Ohlmacher GC, Davis JC (2003) Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. Engineering Geology 69:331-343

O'reilly MP, Rosbrook RB, Cox GC, McCloskey A (1989) Analysis of defects in 180 km of pipe sewers in southern water authority. TRRL Research Report 172

Pourghasemi HR, Moradi HR, Fatemi Aghda SM (2013) Landslide susceptibility mapping by binary logistic regression, analytical hierarchy process, and statistical index models and assessment of their

performances. Natural Hazard 69:749-779. Doi 10.1007/s11069-013-0728-5

Ozdemir A (2016) Sinkhole susceptibility mapping using logistic regression in Karapınar (Konya, Turkey). Bulletin of Engineering Geology and the Environment 75:681-707

QGIS Development Team (2014) QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://www.qgis.org/

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/

Rogers CJ (1986) Sewer deterioration studies: the background to the structural assessment procedure in the sewage rehabilitation manual. 2nd edn. WRc report ER199E

Seoul Metropolitan Government (2016) Seoul statistical yearbook. Data and Statistics Division of Seoul Metropolitan Government, Seoul, South Korea

Seoul Metropolitan Government (2015) Maintenance and Enhancement of Sewage Information Systems and Increase in precision of GIS DB. Sewerage Treatment Planning Division of Seoul Metropolitan Government, Seoul, South Korea, D0000021501950

Slocum TA, McMaster RB, Kessler FC, Howard HH (2009) Thematic cartography and geovisualization. 3rd edn. Upper Saddle River, NJ: Pearson Prentice hall

Tabachnick BG, Fidell LS (2001) Using multivariate statistics. 4th edn. Allyn & Bacon, inc Boston, U.S., pp 524-525

Tran DH, Ng AWM, Perera BJC, Burn S, Davis P (2006) Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. Urban Water Journal 3:3 175-184. doi: 10.1080/15730620600961684

Truett J, Cornfield J, Kannel W (1967) A multivariate analysis of the risk of coronary heart disease in Framingham. Journal of chronic diseases 20:511-524

Van Den Eeckhaut M, Vanwalleghem T, Poesen J, Govers G, Verstraeten G, Vandekerckhove L (2006) Prediction of landslide susceptibility using rare events logistic regression: a case-study in the Flemish Ardennes (Belgium). Geomorphology 76:392-410

Wilson SS, Gurung L, Paaso EA, Wallace J (2009) Creation of robot for subsurface void detection. In IEEE Conference on Technologies for Homeland Security HST'09. pp 669-676. doi: 10.1109/THS.2009.5168102

Yamijala S, Guikema SD, Brumbelow K (2009) Statistical models for the analysis of water distribution system pipe break data. Reliability Engineering and System Safety 94:282– 293

Yilmaz I (2009) Landslide susceptibility mapping using frequency ratio, logistic regression, artificial neural networks and their comparison: A case study from Kat landslides (Tokat-Turkey). Computers & Geosciences 35:1125-1138

Yokota T, Fukatani W, Miyamoto T (2012) The present situation of the road cave in sinkholes caused by sewer systems (FY2006~FY2009). National institute for land and infrastructure management, Ministry of Land, Infrastructure, Transport and Tourism, Japan, Report No.668. (in Japanese)

Zisman ED, Wightman MJ, Taylor C (2005) The effectiveness of GPR in Sinkhole Investigations. The 10th Multidisciplinary Conference on Sinkholes and the Engineering and Environmental Impacts of Karst. pp 608-616. doi: 10.10614/40796(177)65

Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115:654-657