International Journal of Civil Engineering and Technology (IJCIET)
Volume 10, Issue 1, January 2019, pp.93–104, Article ID: IJCIET_10_01_010
Available online at http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=10&IType=1
ISSN Print: 0976-6308 and ISSN Online: 0976-6316

CONDITION PREDICTION MODELS OF DETERIORATED TRUNK SEWER USING MULTINOMIAL LOGISTIC REGRESSION AND ARTIFICIAL NEURAL NETWORK

Dr. Basim Hussein Khudair
Assistant professor, University of Baghdad/ Engineering College, Iraq

Dr Ghassan Khalaf Khalid
Assistant Professor, Al Manhal Academy of Science/Civil Engineering Department, Iraq

Rehab Karim Jbbar
Assistant Lecturer, University of Baghdad/ Engineering College, Iraq

ABSTRACT

Sewer systems are used to convey sewage and/or storm water to sewage treatment plants for disposal by a network of buried sewer pipes, gutters, manholes and pits. Unfortunately, the sewer pipe deteriorates with time leading to the collapsing of the pipe with traffic disruption or clogging of the pipe causing flooding and environmental pollution. Thus, the management and maintenance of the buried pipes are important tasks that require information about the changes of the current and future sewer pipes conditions. In this research, the study was carried on in Baghdad, Iraq and two deteriorations model's multinomial logistic regression and neural network deterioration model NNDM are used to predict sewers future conditions. The results of the deterioration models' application showed that NNDM gave the highest overall prediction efficiency of 93.6% by adapting the confusion matrix test, while multinomial logistic regression was inconsistent with the data. The error in prediction of related model was due to its inability to reflect the dependent variable (condition classes) ordered nature.

Keywords: Condition Prediction, Trunk Sewer Deterioration, Multinomial Logistic Regression, Artificial Neural Network.

Cite this Article: Dr. Basim Hussein Khudair, Dr Ghassan Khalaf Khalid and Rehab Karim Jbbar, Condition Prediction Models of Deteriorated Trunk Sewer Using Multinomial Logistic Regression and Artificial Neural Network, International Journal of Civil Engineering and Technology (IJCIET), 10 (1), 2018, pp. 93–104.
http://www.iaeme.com/IJCIET/issues.asp?JType=IJCIET&VType=10&IType=1

1. INTRODUCTION

Sewer networks are subsurface infrastructure systems, in which sewers collect sewage and/or storm water to sewage treatment plants or other places for disposal. Buried sewer pipes deteriorate with time due to several deterioration factors (e.g. environmental and operational factors) leading to structural deterioration (e.g. breakage or deformation of pipes) and hydraulic deterioration (e.g. blockages or tree intrusion that reduces pipe's hydraulic efficiency) (Micevski et al., 2002). Thus, deterioration causes cities to be susceptible to its effect of collapse and flooding due to huge length of the sewers making it more difficult to monitor them. Alternatively, sewers future condition can be predicted using deterioration models. Then, the predicted information is used by utilities to make optimal decisions on repairing, overhauling or replacing pipes in poor condition (Tran, 2007).

Many researchers wrote in various fields related to the issue of this research, as Salman (2010) applied several deterioration models (binary logistic regression, ordinal regression and multinomial logistic regression) on inspection data of Cincinnati city (USA). The binary logistic regression analysis showed the best performance in predicting sewer deterioration; the total model efficiency was 66%. Prediction efficiency for good condition was 78% and for bad condition 46%.

Khan et al., (2010) and Kadhim Naief Kadhim, 2018 developed deterioration models using data from Pierrefonds, Canada. They used neural network modeling with back propagation (BPNN) and probabilistic (PNN) approaches. Taking about 20% of the available dataset to test the model, the coefficient of determination (R^2) ranged within 71 and 86 % depending on the deterioration factors considered.

The aim of this study is the developing of two deteriorations model's multinomial logistic regression and NNDM using available data. Then, the developed models are used to identify the most important factors that influence the deterioration of sewers.

2. MATERIAL AND METHOD

2.1. Data Acquisition

The effectiveness of condition prediction models depends upon the quality and quantity of the collected data and selection of predictors. Zublin trunk sewer with data from Al-Rusafa side in Baghdad, Iraq was used as a case study to illustrate the applicability of the developed models. For this study, data are collected from site investigation, information from different departments of Baghdad Mayoralty (design, implementation, planning, operation and maintenance and geographic information systems) and also from questionnaire distributed on different sections in the different municipalities of Al-Rusafa that Zublin sewer serves them. The data included: sewer condition, age, diameter, depth, length, slope, traffic intensity and material. A condition rating system with five classes, from 1 (excellent) to4 (poor) and 5 (very poor) was used to describe sewers conditions. There were no sewers in condition 1 (excellent) and few sewers in condition 2 (very good) that have been combined with sewers in condition 3 (good). The dataset available for this trunk sewer contained 97 records corresponding to individual manhole-to-manhole sewer length. Out of the 97 useful sewer data, 77 were set aside for calibration and 20 were for validation (These data numbers were chosen after several attempts through the SPSS programs to obtain the best results with high coefficient of correlation). The statistical analysis software SPSS was selected to perform calibrations for the selected deterioration models. A part of the calibration sample used in developing these models is shown in Table 1.

Table 1. Dataset portion for the calibration of the deterioration models

NAME	M.H	CONDITION	AGE (YEAR)	DIAMETER (M)	LENGTH (M)	DEPTH (M)	SLOPE (M/M)	TRAFFIC	MATERIAL
TH	49A	4	35	1.8	19.09	3.84	0.0005	3	2
TH	50	3	28	1.8	43.19	3.96	0.0007	3	2
TH	34	4	36	2.4	165.57	5.81	0.0005	1	1
TH	35	5	37	2.4	100.36	5.82	0.0005	1	1
NT	39	4	34	3.0	193.35	7.16	0.0005	2	1
NT	40	3	33	3.0	197.23	6.99	0.0006	1	1

Note: Traffic 1 = low, 2 = medium, 3 = high, Material 1 = concrete, 2 = PVC, M.H: Manhole, TH: Al-Thawra Trunk, NT:North Trunk

2.2. Sewer Deterioration Models

2.2.1. Multinomial logistic regression

Multinomial logistic regression is used to model categorical dependent variable with more than two categories. As there are m categories of the dependent variable, one of them is selected as the reference category. Then, (m – 1) log its are generated using the remaining (m – 1) categories as in Eq.1 (Salman, 2010):

$$\ln \frac{(P(Y=j|x_1,x_2,\ldots,x_n))}{(P(Y=m|x_1,x_2,\ldots,x_n))} = \alpha_j + \beta_{j1}\, x_1 + \beta_{j2}\, x_2 + \cdots + \beta_{jn}\, x_n \quad (1)$$

Where: j = 1, 2, …, m – 1 is the dependent variable categories; α_j is the intercept for category j; x_1, x_2, …, x_n are independent variables and β_{j1}, β_{j2}, … , β_{jn} are the regression coefficients that correspond to n-number of independent variables defined for each dependent category j.

2.2.1.1. Model calibration

Maximum Likelihood Estimation method is used to estimate the parameters α_j and β_{jn}in multinomial logistic regression due to this method handles (m – 1) equations simultaneously and determines the parameters that maximize the likelihood function (Agresti, 2002). For i^{th} observation with independent variables $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$, the Log Likelihood function is as follows:

$$\log\left[\prod_{j=1}^{m} \pi_j(x_i)^{y_{ij}}\right] = \sum_{j=1}^{m-1} y_{ij} \log \pi_j(x_i) + \left(1 - \sum_{j=1}^{m-1} y_{ij}\right) \log\left[1 - \sum_{j=1}^{m-1} \pi_j(x_i)\right] \quad (2)$$

Where: for i = 1, 2, …, p; $y_i = (y_{i1}, y_{i2}, \ldots, y_{im})$ = multinomial trial for subject i; $y_{ij} = 1$ if the response is in category j, and otherwise 0. Eq. 2 may be shortly written, if all observations are considered as follows (Agresti, 2002):

$$\log Likelihood = \log \prod_{i=1}^{p} \left[\prod_{j=1}^{m} \pi_j(x_j)^{y_{ij}}\right] \quad (3)$$

2.2.1.2. Model significance

To evaluate this model significance, a comparison is made between the values of log likelihood of the base model (defined by eq. 4) and final model (defined as a model that approved after different statistical significant using chi-square test). After multiplying these log likelihood values by -2, a chi-square distribution is used to evaluate the statistical significance of the difference between these values. The critical chi-square value is compared with the test value with DOF (i.e. degree of freedom) equals to the model estimated number of the coefficients of regression. The log likelihood equation of the base model is as follows (Menard, 2002):

$$\log Likelihood(base\ model) = [n_1 \times \ln(p_1) + n_2 \times \ln(p_2) + \ldots + n_m \times \ln(p_m)] \quad (4)$$

Where, n_1, n_2, …, n_m is the observations number in each category level and p_1, p_2, …, p_m is the observations proportion corresponding to the respective category.

2.2.1.3. Significance of regression coefficients

The likelihood ratio method can be used to determine the variables significance. In this method, the difference in the values of -2 Log likelihood between a final model and a reduced model (is formed by omitting the variable of interest from the final model) should be calculated (Menard, 2002). The critical chi-square value is compared with the resultant value with DOF equals to the regression coefficients number that are omitted from the model. The DOF would be equal to m – 1, if the omitted variable is categorical of values equal to m.

2.2.1.4. Assumptions of multinomial logistic regression

If an ordinal relationship has been found between the dependent variable categories, this model cannot reflect the dependent variable ordered nature. When using an ordinal regression model, independent variables coefficients are required to be constant for each dependent variable level, which represents the proportional odds assumption (McCullagh, 1980).This assumption is restrictive in this model application compared to the multinomial logistic regression that provides more flexibility to the independent variables coefficients.

2.2.2. NNDM

Neural networks can be used to predict outcome data from input data in a manner that simulates the operation of the human nervous system. Unlike statistical models, NNs have no assumptions related with the model structure because it is determined by data (Tran, 2007). Generally, the model can simulate nonlinear relationships within the deterioration process and can handle ordinal outputs such as condition classes.

2.2.2.1. Model structure

Generally, a neural network consists of several layers containing artificial neurons connected together (Al-Barqawi and Zayed, 2008) as shown in Fig. 1. The connection weights, which attach the connections between the neurons are calculated using the observed data when the difference (i.e. error) between the values of the actual output and predicted output is small (Salman, 2010). The NNs have always a special input signal values equal to 1, with a bias weight. They are not shown in Fig. 1 in order to reduce the complexity. The function of bias weight is to allow or stop the input signals going through by (being non-zero) or (being zero) value respectively. The general form of a neural network function is as follows:

$$Y = f(\sum_{i=1}^{k} Xi\ Wi) \quad (5)$$

Where, Y the output signal, Xi the input signal, K the number of input signals, Wi the connection weights, f the activation function.

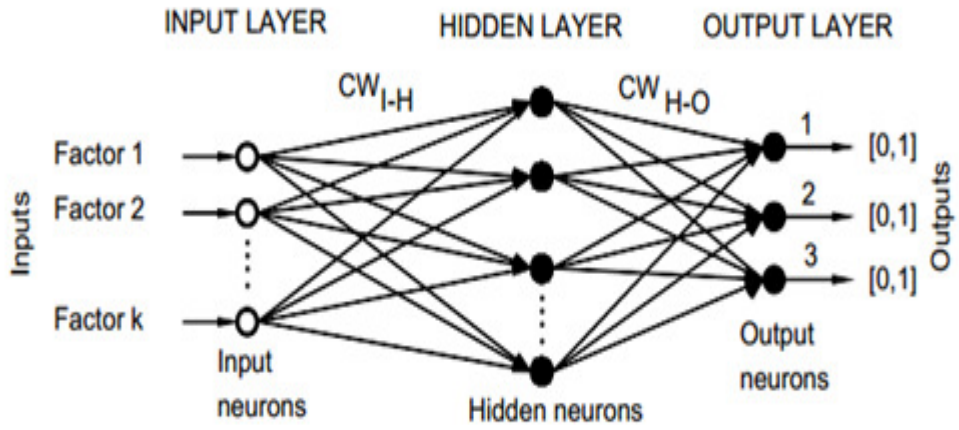

Figure 1 Structure of the NNs (Al-Barqawi and Zayed, 2008)

2.2.2.2. Activation functions

The values of units in the succeeding layer are linked to the weighted sums of units in a preceding layer by the activation function. The hyperbolic tangent function was used for the neurons of hidden layer and the soft max function was used for the neurons of output layer in this study, since using automatic architecture and the output is categorical (IBM® SPSS® Statistics 20 User Guide).

3. RESULTS AND DISCUSSION

3.1. Development of Multinomial Logistic Regression

3.1.1. Model Parameters

The generated equations of multinomial logistic regression are two, since the outcome variable has three possible states (the procedure of predicting three outcome variables from two equations is written in section 3.1.4). Mathematically, multinomial logistic regression equations can be written as below:

$$\ln\left(\frac{P(CR=j)}{P(CR=5)}\right) = \alpha_j + \beta_{j1} \times Age + \beta_{j2} \times Diameter + \beta_{j3} \times Length + \beta_{j4} \times Depth + \beta_{j5} \times Slope + \beta_{j6} \times Z_{Traffic=1} + \beta_{j7} \times Z_{Traffic=2} + \beta_{j8} \times Z_{Material=1} \quad (6)$$

Where: CR= Condition Rating, j = 3 and 4 indicates the condition level, and j=5 was selected as reference category; αj and *βj1, βj2,…, βj8* are the intercept and regression coefficients for condition level *j* respectively, Age, Diameter, Length, Depth and Slope are the numerical independent variables and the different values of the categorical variables are defined by dummy variables *Zi* (Ariaratnam et al., 2001), for which, a value of either 1 or 0 are assigned as follows:

- $Z_{Traffic=1} = 1$, if traffic intensity = 1 (low), otherwise 0
- $Z_{Traffic=2} = 1$, if traffic intensity = 2 (medium), otherwise 0
- $Z_{Material=1} = 1$, if material type = 1 (concrete), otherwise 0

Table 2 provides the intercept term and regression coefficients for the first and second equations. In addition, this table gives Exp (β), which explains how much the odds of Y change for a unit increase of an independent variable.

Table 2. Parameter estimates of multinomial logistic regression for condition ratings 3 and 4.

Condition[a]		β	Std. Error	Wald	df	Sig.	Exp(β)	95% Confidence Interval for Exp(β)	
								Lower Bound	Upper Bound
3.0	Intercept	95.890	28.769	11.109	1	.001			
	age	-3.074-	.868	12.538	1	.000	.046	.008	.253
	diameter(m)	3.314	5.737	.334	1	.563	27.504	.000	2101282.970
	length	.014	.031	.213	1	.644	1.014	.955	1.078
	depth	1.672	1.626	1.058	1	.304	5.323	.220	128.891
	slope	-519.433-	2186.123	.056	1	.812	2.588E-226	.000	.[b]
	[traffic=1]	-10.570-	3.806	7.714	1	.005	2.566E-5	1.478E-8	.045
	[traffic=2]	-6.585-	4.542	2.102	1	.147	.001	1.881E-7	10.143
	[traffic=3]	0[c]	.	.	0	.	.	.	.
	[material=1]	2.794	7.709	.131	1	.717	16.342	4.480E-6	59604937.147
	[material=2]	0[c]	.	.	0	.	.	.	.
4.0	Intercept	67.426	24.661	7.475	1	.006			
	age	-2.042-	.755	7.323	1	.007	.130	.030	.570
	diameter(m)	4.795	5.014	.914	1	.339	120.863	.007	2241546.931
	length	.031	.027	1.308	1	.253	1.032	.978	1.089
	depth	-.450-	1.389	.105	1	.746	.637	.042	9.706
	slope	1371.297	1364.346	1.010	1	.315	.[b]	.000	.[b]
	[traffic=1]	-2.685-	2.586	1.078	1	.299	.068	.000	10.836
	[traffic=2]	1.717	3.020	.323	1	.570	5.569	.015	2072.682
	[traffic=3]	0[c]	.	.	0	.	.	.	.
	[material=1]	-4.596-	4.558	1.017	1	.313	.010	1.331E-6	76.488
	[material=2]	0[c]	.	.	0	.	.	.	.

a. The reference category is: 5.0.

b. Floating point overflow occurred while computing this statistic. Its value is therefore set to system missing.

c. This parameter is set to zero because it is redundant.

3.1.2. Significance of the Model

The model significance was evaluated using the likelihood ratio of the final to the intercept only models. Table 3 provides the model significance and values of likelihood ratio. Based on the results, the– 2 log likelihood values difference between the final and intercept only models is 130.808 that corresponds to a significant result as p-value was 0.00 which is less than significance level of 0.05.

Table 3 Significance test results for multinomial logistic regression analysis.

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	168.763			
Final	37.955	130.808	16	.000

3.1.3. Significance of Model Parameters

Table 4 shows the significance of each model parameter and differences between -2 Log Likelihood values. Based on the results, excluding diameter variable from the overall model results in the lowest difference in -2 log likelihood value (Chi-Square that is 1.096); while, the exclusion of age from the model corresponds the highest difference in the -2 log likelihood value (i.e. 90.921). According to the significance values shown in Table 4, the parameters of age, traffic and depth are significant as they have p-values less than 0.05.

Table 4 Significance test results of multinomial logistic regression for independent variables.

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	37.955	.000	0	.
Age	128.876	90.921	2	.000
diameter(m)	39.051	1.096	2	.578
Length	40.407	2.452	2	.293
Depth	44.654	6.699	2	.035
Slope	40.576	2.621	2	.270
traffic	65.562	27.607	4	.000
material	41.787	3.832	2	.147

3.1.4. Prediction Rate

The prediction procedure involves two steps. The first step involves entering the parameter estimates and independent variables into the equations of the multinomial logistic regression and computing odds ratios. Second step includes using the values of the odds ratio to compute the probabilities associated with each class. By inserting the predicted parameters of the model, the equations of multinomial logistic regression can be rewritten as follows (Salman, 2010):

$$\ln\left(\frac{P(CR=3)}{P(CR=5)}\right) = 95.89 - 3.074 \times Age + 3.314 \times Diameter + 0.014 \times Length + 1.672 \times Depth - 519.433 \times Slope - 10.57 \times Z_{Traffic=1} - 6.585 \times Z_{Traffic=2} + 2.794 \times Z_{Material=1} = G_1(7)$$

$$\ln\left(\frac{P(CR=4)}{P(CR=5)}\right) = 67.426 - 2.042 \times Age + 4.795 \times Diameter + 0.031 \times Length - 0.45 \times Depth + 1371.297 \times Slope - 2.685 \times Z_{Traffic=1} + 1.717 \times Z_{Traffic=2} - 4.596 \times Z_{Material=1} = G_2(8)$$

G1 and G2 are the odds ratios and once they are determined, the following equations are used to calculate the probabilities associated with each condition level:

$$P(CR=3) = \frac{exp(G_1)}{[1 + exp(G_1) + exp(G_2)]}$$

$$P(CR=4) = \frac{exp(G_2)}{[1+exp(G_1)+exp(G_2)]} \qquad (9)$$

$$P(CR = 5) = \frac{1}{[1 + exp(G_1) + exp(G_2)]}$$

To make a classification, the observed values of the predictors are inserted into the Eq. 7 and Eq. 8. Then, classification probabilities are calculated from Eq. 9. The observation is assigned to the class with the highest classification probability.

3.2. Development of NNDM

In this model, approximately 61% of the data were assigned for training, 19.5% for testing and 19.5% to a holdout sample as this configuration gave high overall prediction efficiency. Furthermore, values of all the scale input factors are rescaled using normalized method according to Eq. 10 to improve network training (IBM® SPSS® Statistics 20 User Guide).

$$X_i = \frac{x-min}{max-min} \quad (10)$$

3.2.1. Training of NNDM

NNDM training in this study is used to calculate the model structure (i.e. the network weights and the hidden neurons numbers). Three neurons in the hidden layer have been chosen by automatic architecture selection.

3.2.2. Sample prediction

The model architecture is listed in Table 5, where the condition of a sewer with a particular characteristic can then be predicted.

The non-linear relationship between the input and output data can be written as follows:

$$\mathrm{C} = \sum_{i=1}^{n} X_i W_i + Wo \quad (11)$$

$$H_i = \tanh(C) \quad (12)$$

$$Y_j = \frac{e^{(H_k)}}{\sum_j e^{(H_j)}} \quad (13)$$

Where: n the number of the predictors, Wo bias weight, Hi the output of the hidden neurons, $Yj$the output of the output neuron, H_k is the input for Yj.

To make a classification, the observed values of the predictors are inserted into the equations above to calculate probabilities, which are values between 0-1. The observation is assigned to the class with the highest probability.

Table 5. Estimation of hidden and output parameters

Predictor		Predicted					
		Hidden Layer 1			Output Layer		
		H(1:1)	H(1:2)	H(1:3)	[Condition=3.0]	[Condition=4.0]	[Condition=5.0]
Input Layer	(Bias)	.207	.499	-1.453-			
	[traffic=1]	-1.196-	.067	-.207-			
	[traffic=2]	1.021	.092	-.739-			
	[traffic=3]	-.432-	-1.219-	-.620-			
	[material=1]	-.536-	-.768-	-1.396-			
	[material=2]	1.287	-.089-	-.985-			
	Age	-.447-	3.834	5.020			
	Diameter m	-.815-	-.607-	-.208-			
	Length	-.275-	-.444-	-.554-			
	Depth	-.758-	-.507-	.064			
	Slope	.465	.931	-.325-			
Hidden Layer 1	(Bias)				-1.196-	.844	.273
	H(1:1)				.206	1.456	-1.949-
	H(1:2)				-3.780-	1.746	1.411
	H(1:3)				-2.760-	-1.088-	3.987

3.2.3. Independent Variable Importance

The importance of each independent variable in determining the neural network is computed based on both training and testing samples (IBM® SPSS® Statistics 20 User Guide). It appears from Fig. 2 that the variable age has the greatest effect on how the network classifies sewers followed by traffic, material, slope, length, diameter and depth respectively. To determine the variation in the model-predicted value for various independent variable values, the importance of the independent variable is used. In addition, when dividing the importance values on the largest importance values, normalized importance is obtained, which is expressed as percentages.

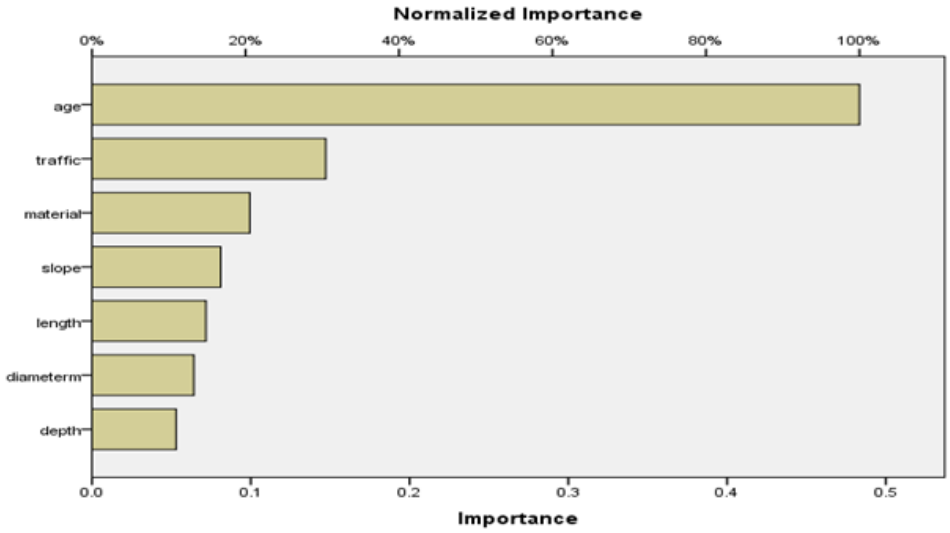

Figure 2 Independent variable importance chart

3.3. Model Performance Evaluation

For evaluating model performance, the model error (i.e. the difference between predicted and observed values) must be quantified (Wright et al., 2006). With high model error, the performance of the model is low. The confusion matrix is often used for ordinal and categorical outputs. The validation dataset should be used to effectively test the model (Baik et al., 2006). When comparing the observed values with model prediction, four possible situations can be observed: (1) true positive (TP) when the model correctly predicts the sewer in good condition, (2) true negative (TN) when the model correctly predicts the sewer in poor condition, (3) false positive (FP) when the model incorrectly predicts the sewer in good class as being in bad class, and (4) false negative (FN) when the model incorrectly predicts the sewer in poor condition as being in good condition as shown in Table 6 (Tran, 2007).

In Table 6, the TP_{11} means the pipes number which are observed and correctly predicted in condition 1. In addition, O_1, O_2 and O_3 represent the total pipes numbers that are Titiladunayo, I. F, Akinnuli, B.O, Ibikunle, R. A, Agboola, O.O, Ogunsemi, B.T Titiladunayo, I. F, Akinnuli, B.O, Ibikunle, R. A, Agboola, O.O, Ogunsemi, B.T C. O. Osueke, T. M. A. Olayanju, C. A. Ezugwu, A. O. Onokwai, I. Ikpotokin, D. C. Uguru-Okorie nd F.C. Nnaji observed in condition 1, 2 and 3 respectively and P_1, P_2 and P_3 represent the total pipes numbers which are predicted in condition 1, 2 and 3 respectively.

Table 6 Confusion matrix (Tran, 2007)

		Predicted condition			**Total**
		1 (good)	**2 (fair)**	**3 (poor)**	
Observed condition	1 (good)	TP_{11}	FP_{12}	FP_{13}	O_1
	2 (fair)	FP_{21}	TP_{22}	FP_{23}	O_2
	3 (poor)	FN_{31}	FN_{32}	TN_{33}	O_3
Total		P_1	P_2	P_3	

The overall predicted efficiency (OPE) was used to evaluate the performance prediction of multinomial logistic regression and NNDM Salman (2010), which were developed in this research to predict sewers future conditions. From the confusion matrix, the OPE can be computed using Eq. 14.

$$OPE = \frac{TP11 + TP22 + TN33}{O1 + O2 + O3} \qquad (14)$$

Tables 7 and 8are the confusion matrices for multinomial logistic regression and NNDM. According to Table7, the overall predicted efficiency of multinomial regression model was 90.9% for the calibration sample and for all condition classes (3, 4 and 5), prediction percentages were satisfactory. While, for the validation sample, the overall percentage of correct estimations was 55%, which is lower than the calibration sample. For condition classes 3 and 5, the overall predicted efficiencies were satisfactory, but the prediction rate remained low for condition class 4. A higher threat to the model validity is caused by the low prediction rate for condition class 4, which confirms the finding of Salman (2010).

Table 7. Prediction efficiencies for the calibration and validation of the multinomial logistic regression

Condition			Predicted Group Membership			Percent Correct
			3	4	5	
(a) Calibration	Count	3	23	4	0	85.2%
		4	2	25	0	92.6%
		5	1	0	22	95.7%
(b) Validation	Count	3	5	2	0	71.4%
		4	6	1	0	14.3%
		5	0	1	5	83.3%

(a) 90.9% of original grouped cases are correctly classified.

(b) 55% of original grouped cases are correctly classified.

Table 8 shows that the deterioration model based on the NNDM provides the highest overall prediction efficiency in the calibration and validation samples as compared with the multinomial logistic regression. The high overall prediction efficiency by the NNDM could be attributed to its inherent ability to model complex processes. The NNDM was used by Tran (2007), which gave the highest prediction efficiency of 82%.

Table 8. Prediction efficiencies for the calibration and validation of the NNDM

Condition			Predicted Group Membership			Percent Correct
			3	4	5	
(a) Training	Count	3	13	0	1	92.9%
		4	0	16	1	94.1%
		5	1	0	15	93.8%
(b) Testing	Count	3	6	1	1	75.0%
		4	0	4	0	100.0%
		5	0	0	3	100.0%
(c) Holdout	Count	3	4	1	0	80.0%
		4	1	5	0	83.3%
		5	0	1	3	75.0%
(d) Validation	Count	3	6	1	0	85.7%
		4	5	2	0	28.6%
		5	0	0	6	100%

(a) Prediction efficiency is 93.6%

(b) Prediction efficiency is 86.7%

(c) Prediction efficiency is 80.0%

(d) Prediction efficiency is 70.0%

4. CONCLUSIONS AND RECOMMENDATIONS

In order to achieve the research objectives (sewers future condition prediction to make optimal decisions on repairing, overhauling or replacing pipes in poor condition), two deteriorations models' multinomial logistic regression and NNDM were developed, tested and evaluated using the available data for Zublin trunk sewer for predicting the sewer's future conditions. This type of information is extremely important in projecting budgetary requirements for sewer system maintenance and rehabilitation, and the success of proactive pipe maintenance strategies.

The confusion matrix test showed that multinomial logistic regression was inconsistent with the data and the error in prediction of this model was due to its inability to reflect the dependent variable ordered nature. While, NNDM was found to have high overall prediction efficiency, which could be attributed to its inherent ability to model complex processes. In addition, based on the application of these two models, pipe age was found to be highly significant in the deterioration of the Zublin trunk sewer.

In addition, other factors like type of soil backfill, presence of H_2S and groundwater level should be investigated and gathered to understand further the sewer deterioration mechanism and, consequently, to develop a more effective model.

REFERENCES

[1] Agresti, A., 2002. Categorical data analysis, 2nd Ed., Wiley, Hoboken, New Jersey.

[2] Al-Barqawi, H. and Zayed, T., 2008. Infrastructure Management: Integrated AHP/ANN Model to Evaluate Municipal Water Mains' Performance, J. Infrastructure Systems, 14 (4): 305-318.

[3] Ariaratnam, S.T., El-Assaly, A. & Tang, Y., 2001. Assessment of infrastructure inspection needs using logistic models. Journal of Infrastructure Systems, 7(4), 160-165.

[4] Baik, H. S., Jeong, H. S. and Abraham, D. M., 2006. Estimating Transition Probabilities in Markov Chain-Based Deterioration Models for Management of Wastewater Systems, Journal of Water Resources Planning and Management, ASCE, 132 (1): 15-24.

[5] IBM® SPSS® Statistics 20 User Guide.

[6] Kadhim Naief Kadhim (Estimating of Consumptive Use of Water in Babylon Governorate-Iraq by Using Different Methods). (IJCIET), Volume 9, Issue 2, (Feb 2018)

[7] Khan, Z., Zayed, T. and Moselhi, O., 2010. Structural Condition Assessment of Sewer Pipelines. Journal of Performance of Constructed Facilities 24: 170-179.

[8] McCullagh, P., 1980. Regression models for ordinal data. Journal of the Royal Statistics Society, Series B (Methodological), 42(2), 109 – 142.

[9] Menard, S., 2002. Applied logistic regression analysis. *Sage University Papers Series onQuantitative Applications in the Social Sciences, 07-106*, Thousand Oaks, CA.

[10] Micevski, T., Kuszera, G. and Coombes, P., 2002. Markov model for storm water pipe deterioration. Journal of Infrastructure Systems 8(2): 49-56.

[11] Salman, B., 2010. Infrastructure Management and Deterioration Risk Assessment of Wastewater Collection Systems, Ph.D. thesis, University of Cincinnati. Ohio.

[12] Tran, D.H., 2007. Investigation of deterioration models for stormwater pipe systems. Doctoral dissertation, School of Architectural, Civil and Mechanical Engineering, Faculty of Health, Engineering and Science, Victoria University, Australia.

[13] Wright, L.T., Heaney, J.P. and Dent, S., 2006. Prioritizing Sanitary Sewers for Rehabilitation Using Least-Cost Classifiers. Journal of Infrastructure Systems, ASCE, 12 (3): 174-183.