IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 3 Ver. IV (May. - Jun. 2016), PP 127-136
www.iosrjournals.org

Statistical Analysis of Pipe Breaks in Water Distribution Systems in Ethiopia, the Case of Hawassa

Faris Hamdala K[1], G Y Sagar [2]
[1, 2] *School of Mathematical & Statistical Sciences, Hawassa University, Hawassa, Ethiopia*

Abstract: *The aim of this paper is to investigate the high influential factors of pipe breaks in water distribution systems by using various statistical models such as multiple linear regression model (MLRM), Time exponential model (TEM), Poisson generalized linear model (PGLM), Negative Binomial generalized linear model (NBGLM) and Proportional hazard model (PHM). This paper discusses the effect of different types of covariates with the assumption that each of the covariates has linear effect on the number of pipe breaks. The study focuses on comparing statistical models, parameter estimate of pipe breaks and to determine the covariates that most affect the number of pipe breaks. We developed statistical analysis to compare the usefulness of different statistical methods and obtained the most predictive models for the number of pipe breaks data in water distribution system. The covariates used in the analysis are pipe diameter, length of pipe, year of installation, average rainfall, average temperature, concrete steel cage, asbestos cement, Polyvinyl chloride, types of pipe and time to event data. We compared the results with different other studies on the problem of pipe breaks in water distribution systems.*

Keywords: *Pipe break, Water distribution system, Infrastructure, and Generalized linear models.*

I. Introduction

The Millennium Development Goal (MDG) drinking water targets to halve the proportion of the population without sustainable access to safe drinking water (an increase in coverage from 76% to 88%) between 1990 and 2015 [13]. Between1990 and 2014, 2.3 billion people gained access to an improved drinking water source, raising global coverage to 89% in 2014. In a further 35 countries, 26 of which are in sub-Saharan Africa, coverage of improved drinking water supply was between 50% and 75% [13]. All sources confirmed that water supply coverage in Ethiopia is on a strong upward trajectory. According to official government data, water supply coverage has risen from 19% in 1998 (11% rural, 70% urban) to 66% in 2015 (62% rural, 89% urban) [13]. As official government data, Ethiopia has already met the MDG target of 60%. Estimates of current coverage from the international Joint Monitoring Programme (JMP) are significantly more cautious, due to a range of factors. Nevertheless, the JMP data still portray a remarkable increase in coverage of over 1 million people per year (1990-2007) [12]. The deterioration of pipes in water distribution systems is of concern to water utilities throughout the world. This deterioration generally leads to pipe breaks, which may result in reduction in the water carrying capacity of the pipes from tuberculation of interior walls of the pipe. Deterioration can also lead to contamination of water in the distribution systems [2]. All these systems constitute the infrastructure of urban centers and water distribution systems play a critical role in the successful functioning of a life. Community public health standards and the drive for future growth and economic development are heavily dependent upon the condition of water mains and the services they provide [5].

The main objectives of the study is to examine the high influential factors of pipe breaks in water distribution system at Hawassa city water supply and sewerage service enterprise, Ethiopia. To identify the most determinant factors associated with number of pipe breaks of water distribution system. To comparing the usefulness of different statistical models on the number of pipe breaks. To develop an understanding about the pipe break mechanisms and factors contributing to pipe breaks.

Causes of Pipe Breaks

[2] Some causes of common pipe failures are: 1. pipe manufacturing defects:- inclusions, discontinuities problems, dimensional irregularities particularly in jointing areas. 2. Storage and handling:- stress deformation due to poor stacking or storage, cuts or scratches on pipe walls, impact cracks due to dropping or striking. UV degradation over weathering contamination inadvertent mixing of pipe class or jointing material. 3. During construction:- poor laying, jointing or tapping techniques, excessive soil slips causing distortion, construction traffic, groundwater flooding. 4. Subsequent works:- superimposed loadings, impacts or cover reduction, side slips, service pulling and loss of support or bedding. 5. Soil movement:- subsidence due to mining, filled land, differential consolidation or geological changes, changes in water table or soil moisture content, extremes of climate such as frost heave or clay shrinkage, loss of anchorage (horizontal or vertical), shock waves such as seismic, blasting or vibration.

Background

The need to obtain quantitative estimates of the likelihood of pipe breaks on individual pipes for making repair versus replacement decisions for deteriorating water pipes led to the derivation of predictive models for pipe breaks [2]. Three basic categories of such models are, Aggregate type models, where the expected number of breaks is a function of time t, since a reference time period and a set of constant model parameters. Regression type models where the expected number of pipe breaks are predicted as a function of independent variables reflecting environmental condition and pipe characteristics. Probabilistic or choice models, where discriminate analysis is applied on the data, and break time models, where a survivor function is estimated for each individual pipe, which provides the probability that a pipe will survive without breaks beyond time t, as a function again of a number of independent covariates related to environmental conditions and other pipe characteristics [2]. Two equations [9] were developed (linear and exponential) to describe break rate as a function of time in the range of 0.01-0.15 proposed a value of 0.086 and reported values of 0.021 and 0.014 for pit cast iron and sands pun cast iron pipes respectively, when they employed a similar modeling approach on other data sets. [11] Developed the multiple linear regression model the R^2 value 0.23 was obtained. This shows that the models do not fit the data satisfactorily well and also it is not known how statistically significant the estimated coefficients are. [6] Related pipe break linearly to its age. They obtained the data based on a relatively constant sample of pipes installed within a 10 year period in Winnipeg, Manitoba. For asbestos cement and cast iron pipes, they obtained a moderate correlation of 0.563 and 0.103 respectively between annual breakage rate and pipe age. [8] Introduced semi parametric Proportional Hazards Model (PHM) for analyzing pipe breaks in water distribution. [2] Developed a model for predicting failure probability for each individual pipes in the network for two large water utilities in the Northeastern U.S. In this model, the life span of a pipe is divided into a slow break-stage and a fast break-stage. The fast break-stage starts after three breaks [8]. A Proportional Hazard Model (PHM) is used to describe the break rate h(x) for each pipe as a function of time. A bathtub shaped curve gives the best description of the baseline hazard function. The case study finds the following variables to be most significant in analyzing breaks are pressure in the pipe, number of previous break, age of pipe at the time of the second break, installation period, land used and pipe of length. The break rate does not increase after the third break. The pipes used in the analysis vary greatly in length. Some of the pipes are too long, i.e. more than 1000m to be analyzed as one component, as conditions affecting break rate could vary along the pipe length [2]. The developed estimation procedures for accelerated failure model with random effects to allow for possible correlations among the survival time. Under this model, we can assess the strength of association between event times [14]. Accelerated failure model with random effects provide an attractive alternative to multivariate frailty proportional hazards models based on frailties with a multivariate log normal joint distribution. He proposed an estimation procedure for the Accelerated failure model with random effects that is computationally feasible for large data sets. It is based on a Laplace approximation of the marginal likelihood function, which is similar to generalized linear mixed models [14].

Deterioration models, based on break data of water mains, have been developed in order to predict the annual break rate considering pipe diameter, length, and age. The developed models have been proved to be statistically significant and sound. According to the analysis of the developed models, pipe length has a great impact on the annual break rate [1]. Pipes of shorter lengths that have higher annual break rates do not necessarily have more breaks compared to pipes of longer lengths. The developed models have relatively satisfactory R-squared: 68.9, 65.0, 71.5, 78.4, and 81.3 percent for gray cast iron, ductile iron without lining, ductile iron with lining, PVC and hyprescon pipes, respectively [1].

Statistical model for predicting failures for each pipe in a water distribution network has been presented. The proposed method is able to model the power law. The model has several parameters which include both the effects of repair and the types of failure. Thus, this model uses the failure intensity function is capable and convenient for prediction of failures [7].

II. Statistical Methods And Data

The study will be conducted in Hawassa city water supply and sewerage service enterprise. Hawassa city is the capital of SNNPR, Ethiopia. The service sector is leading the investment following by the industry and agriculture. It is located on the main road Kenya Moyale at a distance 275*km* from the center south direction of Addis Ababa. According to Hawassa city water supply and sewerage service enterprise, the geographic coordinateness of the city is approximately $7^0$09' latitude North and $38^0$29' longitude east and average altitude of 1700*m* the town has got plain topography. Hawassa city climate is warm temperature which varies 10^0c in winter and 30^0c in summer.

Pipe break data of water distribution systems are typically count (0, 1, 2 ...) because there is a number of pipe breaks. The oldest break records in the pipe break data set used in this study date back to 2001. That is, the pipe break observation period is of 10 years only. According to Hawassa city water supply and sewerage service enterprise, the size of the pipe break data set is relatively large, about 63, 262 pipes and 4, 549 breaks,

with such an observation period it will be difficult to assess the deterioration process. The tight observation window causes a great amount of pipes with no recorded breaks; more than 30 percent of pipes have broken during the 10 years of observation. Besides that, needs to be taken into account is the scarcity of the collected component characteristics. The majority of pipe break models developed to estimate and to fit pipe breaks models in water distribution systems and use the following variables: pipe diameter, installation year, length of pipe, PVC pipe, CSC pipe, AC pipe, pipe type. In some models even environmental characteristics, such as temperature and rain fall are used.

Nevertheless, the available data used in this research work with 12 variables. The study instruments that will be employed under this study are secondary data. The climatic data of temperature and rainfall were obtained from National Oceanographic and Atmospheric Administration's National Climatic Data Center. Another issue survival data is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs, alternatively.

2.1 Multiple linear regression models

Multiple linear regression analysis can be used as statistical tool to discover relationships between variables, which are used in prediction models. The goal of multiple linear regression analysis is to develop a statistical model that can be used to predict the values of dependent variable based on the values of more than one explanatory variable. The relationship between the variables based on the coefficient sign show an indication for the nature of the relation if increasing or decreasing relationship.

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon_i \quad 1$$

Where y_i is the number of pipe breaks, while $x_1, x_2 \ldots x_p$ stand for the input indicators $n \times p$ matrix. The regression parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the coefficients of multiple linear regression $p \times 1$ vectors, $x_1, x_2 \ldots x_p$ are independent identically distributed $N(0, \sigma^2)$. We can write this matrix form $Y = X\beta + \epsilon$ we denote the column of matrix X by $x_1, x_2 \ldots x_p$. So we can solve for β to get MLE. $\beta = (X'X)^{-1}X'Y$.

2.2 Time exponential model

Non-linear regression model extends linear regression model to a much larger and more general class of functions. Use non-linear regression analysis to relate pipe's breakage to the exponent of its time to break and installation time. This type of model is often transformed to yield a model that is linear in the regression parameters, and then the transformed model is fit as a linear regression model. The biggest advantage with non-linear models is that it can fit a broad range of functions. The only advantage of this model is their simplicity and the smaller amount of data required as compared to other regression models. For instance, the strengthening of concrete as it cures is a non linear process.

$$N(t) = N(t_0)e^{A(t-t_0)} \quad 2$$

where $N(t)$ is the number of pipe breaks , $N(t_0)$ is the number of breaks per unit length at the year of installation of the pipe, the value is between 0.10-0.25 estimated by [4], time is the time to break of a given break in the past in months and A is a breakage growth rate between 0.05-0.15 estimated by [10].

2.3 Poisson Generalized linear models

A Poisson Generalized linear model is a model commonly used for regression analysis of count data such as pipe breaks in an infrastructure system. Let $x_i = [x_{1i}, x_{2i} \ldots, x_{mi}]$ are the vector of n covariates for system segments i, $i = 1, 2, \ldots, m$ and the number of pipe breaks on segment i be given by Y_i. A regression model based on the Poisson distribution for the counts conditional on the observed values of the covariates specifies that the conditional mean of the counts is given by a continuous function $\mu = (\beta, x_i)$ of the covariate values as given, where β is the nx1 vector of regression parameters. Conditional on, x_i the probability density function assumed for y_i in a Poisson generalized linear model is given, for positive integers y_i

$$f(y_i / x_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \quad 3$$

The log link function will use in this paper to specify the conditional mean, i.e., it is assume by $E(y_i / x_i) = \exp(x_i'\beta)$. GLMs assume that the explanatory variables are independent, but they do not assume that the errors are normally distributed or homoscedastic.

2.4 Negative Binomial Generalized Linear Model

The major assumption of the Poisson model is

$$E(y_i/x_i) = \mu_i = e^{x_i'\beta_i} = var(y_i/x_i)$$

Implying that the conditional mean function equate the condition variance function. This is very restrictive. If $E(y_i/x_i) > var(y_i/x_i)$ then we speak about over dispersion, and when $E(y_i/x_i) < var(y_i/x_i)$ we say, we have under dispersion. The generalized Poisson model does not allow for over or under dispersion. A richer model is obtained by using the negative binomial distribution instead of the Poisson distribution. We then use

$$p(y_i, x_i) = \frac{\Gamma(\theta+y_i)}{\Gamma(y_i+1)\Gamma(\theta)}\left(\frac{y_i}{\mu_i+\theta}\right)^{y_i}\left(1-\frac{\mu_i}{\mu_i+\theta}\right)^{\theta} \qquad 4$$

The negative binomial distribution can be shown to have conditional mean μ_i and conditional variance $\mu_i(1+\eta^2\mu_i)$ with $\eta^2 = 1/\theta$, $\eta>0$ it shows dispersion parameter. Note that the parameter η^2 is not allowed to vary over the observations the conditional mean $E(y_i/x_i) = \mu_i = e^{x_i'\beta_i}$ and conditional variance $var(y_i/x_i) = e^{x_i'\beta_i}\left(1+\eta^2 e^{x_i'\beta_i}\right)$. Using maximum likelihood estimate, we can then estimate the regression parameter β and also the extra parameter η. The parameter η measures the degree of over or under dispersion. The limit case $\eta = 0$ corresponds to the Poisson mode. Fitting negative binomial regression is very similar to fitting of Poisson regression, assuming that the model is the same as the one described in Poisson generalized linear model, that is, the log of the mean μ, is a linear function of independent variables, log $log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ which implies that μ is the exponential function of independent variables, $\mu = e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$.

2.5 Proportional Hazards Model

Survival models simulate repairable system, some change in notation and terminology are required. This method can be thought of survival data analysis in which first break and repair, second break, second repair and subsequent breaks. The basic quantity employed to describe time-to-event phenomena is the survival function. Censoring in lifetime analysis only the first break is considered. Essentially data are said to be censored when there are individuals in the sample where only a lower or upper bound on lifetime is available [1]. The response variable in survival analysis is survival time and is no longer limited to only time to break. It is a non-negative random variable used loosely for the time period from a starting time point to the occurrence of any event [1]. The Cox model or proportional hazard model is semi-parametric, since its hazard function is the product of an unspecified baseline hazard function, and a parametric function relating the hazard function and the covariates [1].

$$h(t/x) = h_0(t)c(x'\beta) \qquad 5$$

Where $h_0(t)$ is the baseline hazard function, $\beta = \beta_1, \beta_2 \ldots \beta_p$ is a parameter vector, x' is a column vector of independent variables $[x = [x_1, x_2, \ldots, x_p]$ and $c(x'\beta)$ is a known function. If $c(x'\beta) = \exp(x'\beta) = \exp(\sum_{i=1}^{p}\beta_i x_i)$ then $h(t/x) = h_0(t)\exp(\sum_{i=1}^{p}\beta_i x_i)$. The Cox model is often called a proportional hazard model (PHM) because if we look at two pipes with covariate values x and x^* the ratio of the hazard functions is:

$$\frac{h(t/x)}{h(t/x^*)} = \frac{h_0(t)\exp(\beta x)}{h_0(t)\exp(\beta x^*)} = \frac{h_0(t)\exp(\sum_{i=1}^{p}\beta_i x_i)}{h_0(t)\exp(\sum_{i=1}^{p}\beta_i x_i^*)} = \exp(\sum_{i=1}^{p}\beta_i(x_i - x_i^*))$$

When that there is no tied time assumed the partial likelihood is defined over all breaks time t_i that $i = 1, 2, \ldots, m$ and given as

$$Lp = \prod_{i=1}^{m}\frac{\exp(\beta x_i)}{\sum_{j\epsilon R_{ti}}\exp(\beta x_j)}$$

Where the product is over m distinct break times and x_i denotes the value of the covariate for the subject with ordered survival time t_i. The log partial likelihood function is

$$Lp = \sum[\beta x_i - (\sum_{j\epsilon R_{ti}}\exp(\beta x_j))]$$

We obtain the maximum partial likelihood estimator by differentiating the equation with respect to the component of β setting the derivative equal to zero and solving for the unknown parameters. However, this partial likelihood function methods are based on the assumption that there are no tied values among the observed survival times. But, in most real situations tied survival times are more likely to occur. Once a model has been developed, we would like to know how effective that model is in describing the outcome variable. This

is referred to as goodness of fit. Thus, the goodness of fit of a statistical model describes how well it fits a set of observations. The goodness of fit test measures the compatibility of a random sample with a theoretical probability distribution function.

2.6 Diagnostics in the generalized linear models

The two basic types of residuals are the so called Pearson residuals and deviance residuals. Pearson residuals are based on the difference between observed responses and the predicted values; deviance residuals are based on the contribution of the observed responses to the log likelihood statistic. In addition, leverage scores, studentized residuals, generalized Cook's D, and other observational statistics can be computed.

Goodness of Fit Test

Once a model has been developed, we would like to know how effectively that model is in describing the outcome variable. This is referred to as goodness of fit. Thus the goodness of fit of a statistical model describes how well it fits a set of observations and measures the compatibility of a random sample with a theoretical probability distribution function. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. The deviance or likelihood ratio test statistic, G^2 is the most useful summary of the adequacy of the fitted model. It represents the change in deviance between the fitted model and the model with a constant term and no covariates. The deviance goodness of fit test reflects the fit of the data to a Poisson distribution in the regression. This test is significant shown by the P value, and we should consider other covariates or other error distributions such as negative binomial.

$$Deviance = \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)$$

Where y is the number of events n is the number of observations and $\hat{\mu}_i$ is the fitted Poisson GLM mean. The first term is identical to the binomial deviance, representing twice a sum of observed times log of observed over fitted. The second term, a sum of differences between observed and fitted value is usually zero, because MLEs in Poisson models have the property of reproducing marginal totals, The log likelihood function is

$$LL = \sum_{i=1}^{n} y_i \log(\hat{\mu}_i) - \hat{\mu}_i - \log(y_i)$$

The maximum likelihood equation proceeds by iteratively reweighted least squares, using singular value decomposition to solve the linear system at each iteration until the change in deviance is within the specified accuracy.

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a way of selecting a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth.

$$AIC = -2\big(log(likelihood)\big) + 2k$$

Where likelihood is the probability of the data given a model and K is the number of free parameters in the model. Smallest AIC scores are often shown as best model.

III. Results And Discussion

This section provide the preliminary analysis of data on the number of pipe breaks and reviews various statistical methods for analyzing count data in water distribution systems. These methods also provide information on the value and significance of the covariates. R statistical software was used for analyzing the data.

3.1 Multiple linear regression models

	Estimate	Std. Error	t-value	Pr(>|t|)	Vif
Intercept	-0.0058	0.904	-64.401	0.000 ***	
AC	0.1293	0.0325	3.974	0.000 ***	1.08736
Di	0.92	0.779	1.179	0.2390	1.0825
Le	0.878	0.3928	2.23	0.0262 *	1.3967
Temp	-0.294	0.105	-2.832	0.0050 **	1.2537
Rain	0.0136	0.0026	5.455	0.000 ***	1.1412
Yi	2.946	0.0675	62.80	0.000 ***	2.1712
PVC	-0.00152	0.0068	-0.225	0.823	1.7445
CSC	-0.0536	0.0055	-9.607	0.000 ***	1.4821
Pt	-1.401&	5.35	-2.617	0.00947**	1.1019

Table 1: Results of Parametric Coefficients in Multiple Linear Regression Model

Table 1 shows the summary of coefficients. From the results we observed that there is no multicollinearity in covariates. It can be seen that majority of the coefficients are found significant at 5% level of significance. The coefficient denotes the change in the value of number of pipe breaks for each one unit change in the corresponding covariates. The covariate variables, year of installation, average rainfall and concrete steel cage AC pipe are highly significant. Type of pipe, length of pipe and average temperature are significant. The only covariate diameter of pipe and PVC pipe are non significant compare with 5% level of significant and P-value. Also we can see the relationship between independent variable and covariates in the multiple linear regression models have positive coefficients, Asbestos cement (0.1293), Diameter of pipe (0.92), Length of pipe (0.878), average rainfall (0.0136) and year of installation (2.946) are positive relationship with number of pipe breaks. This implies the value of covariates increase then the number of pipe break increase and the value of covariates decrease then the number of pipe break decrease, and average temperature, PVC pipe, concrete steel cage and type of pipe are negative relation with number of pipe breaks. That is the value of covariate increase, decrease in number of pipe breaks. The covariate variables influence the number of pipe breaks.

$$Number\ of\ pipe\ breaks = -0.0058 + 0.1263_{AC} + 0.3821_{Di} + \boldsymbol{\Delta} \quad \ldots\ldots\ 6$$

Where

$$\boldsymbol{\Delta} = 0.8556_{Le} - 0.331_{Temp} + 0.0137_{Rain} + 2.940_{YI} - 0.0025_{PVC} - 0.0546_{CSC} - 1.288_{Pt}$$

The model suggests the adjusted R square for the regression was 0.9746 and multiple R^2 = 0.9734, p-value = 0.000 significant at 5% level of significance. R^2 0.9734 indicating the regression model is valid and good fitted model and the 9 independent variables are explaining 97.44% of dependent variable Number of pipe breaks.

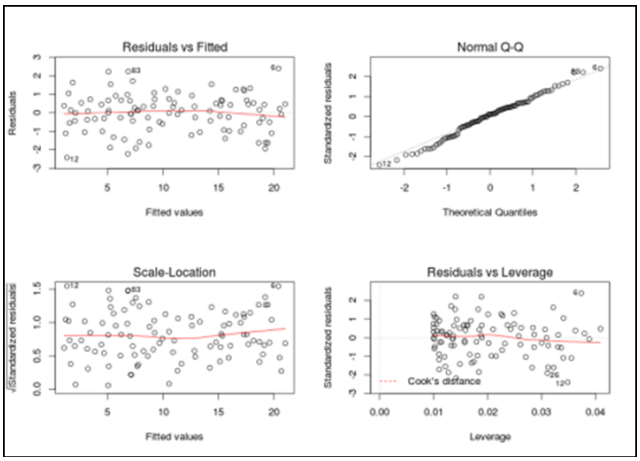

Figure 1: Diagnostic Plot of Multiple Linear Regression Models

Regression diagnostics are methods for determining whether a fitted regression model adequately represents the data. These are useful for detecting in homogeneity in the distribution of residuals. The results of the first diagnostic plot in figure 1 raise no warning flags or no pattern concerning our modeling assumptions. It is also useful to examine the residuals as a function of the values of each explanatory variable. The second graph check for the normal distribution of residuals, the point should fall on the line. A QQ plot graphs the quantiles of the residuals against the theoretical quantiles marked deviation from a straight line indicates that the actual error normally distributed and the plot we see is linear. The bottom left graph is similar to the top left one, the y-axis is changed. The residuals are square root standardized making it easier to heterogeneity of the variance. And there are no indications of failures of the assumptions underlying the linear model. The plot of residuals versus leverage discloses no problems with the fit.

3.2 Time exponential model

The time exponential model is a non linear model, it was fitted in R software using the 'nls' command. It requires initial values of the model parameters. Good initial values are those that are close to the true parameter values that minimize convergence difficulties. Their advantage lies in the fact that they are simple to use. However, they have a few drawbacks which include a. they do not consider other factors such as environmental characteristics, operating pressures, and previous break history of individual pipe segments, and b. the study do not provide any information about the statistical significance of the coefficients of their models. However these models do require smaller amounts of data as compared with other regression models.

Parameter	Estimate	Std. Error	t-value	Pr(> |t|)
$N(t_0)$	2.7847	1.542	1.806	0.0721
A	0.085	0.0176	4.823	0.000 ***

Table 2: Parametric Significance in Time Exponential Model

In fitting a non linear regression model take initial values of the model parameters 0.0185 and 0.086, number of pipe break per unit length at the year of installation of the pipe and average breakage growth rate respectively and residual error 22.85 on 238 degree of freedom. Average break growth rate is statistically significant at a lower p-value than number of pipe breaks at year of installation. The estimated equation is

$$N(t) = 2.7847\, e^{0.0849(t-t_0)} \quad \ldots\ldots 7$$

We take starting value number of pipe break per unit length at the year of installation of the pipe (N_0) =0.0186 and average breakage growth rate (A) = 0.086 these converge to 2.7847 and 0.08494 respectively. The underlining assumption is $N(t_0) \neq 0$, which means that on average a pipe is assumed to always have breakage frequency, though very small in the beginning of its life.

3.3 Poisson Generalize Linear Model

Poisson GLM is often used to fit rare occurrence using count data. The number of pipe break is a count data and Poisson GLM can be applied to the data. Hence, the following discussion shows how the Poisson GLM applied to this count data.

Table 3 shows that the coefficients denote the change in the expected value of Poisson GLM for each one unit change in the corresponding covariates. The covariate variables, year of installation, average rain fall and Polyvinyl chloride pipe highly significant covariates, type of pipe, CSC pipe and Asbestos cement pipe are significant in the output result with compare to 5% α-level and p-value.

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-0.0148	3.546	-41.723	0.000 ***
AC	0.00373	0.0014	2.671	0.00075**
Di	0.0241	0.033	0.730	0.465
Le	0.00081	0.0138	0.060	0.952
Temp	0.0056	0.0044	1.27	0.204
Rain	0.0004	0.0001	3.955	0.000 ***
Yi	0.0754	0.0018	41.38	0.000 ***
PVC	0.0035	0.0003	11.83	0.000***
CSC	0.0004	0.00023	1.99	0.046 *
Pt	-0.045	0.0208	-2.046	0.0407 *
		degree of freedom		
Null deviance	3908.55	239		
Residual deviance	220.94	230		
AIC	1478.4			

Table 3: Results of Parametric Coefficients in Poisson GLM

The covariates average temperature, diameter of the pipe and length of the pipe non significant covariates. The coefficient of covariates temperature, year of installation, diameter of the pipe, Polyvinyl chloride pipe, Asbestos cement, Rain fall, length of pipe and CSC are positive, this means the mean of the number of pipe break increase with the value of covariates increase. Covariate variables that influence the number of pipe break cases in Hawassa city is based on Poisson GLM which are generated based on the Year of installation, Polyvinyl chloride pipe, average monthly rainfall and the Asbestos cement (AC) pipe. The dispersion parameter was found to be 0.96 and P-value of 0.000 which indicates that the model is significant at 5% α-level. However, the assumption of equal variance to the mean in generalized Poisson model has been accept since the dispersion parameter is 0.96 approximately equal to 1 an indication of equi dispersion in the data.

$$\hat{\mu} = e^{-0.0148 + 0.0037_{AC} + 0.0241_{Di} + 0.0008_{Le} + 0.0056_{Temp} + 0.0004_{rain} + \boldsymbol{\Delta}} \quad \ldots\ldots 8$$

Where, $\boldsymbol{\Delta} = 0.0754_{Yi} + 0.0035_{PVC} + 0.0004_{CSC} - 0.0450_{Pt}$

The first two plots of Figure 2 display the pattern of residuals with respect to the fitted values. These are useful for detecting non homogeneity in the distribution of residuals. If the model is appropriate, we expect residuals that are un-patterned with no trends in magnitude. The results of the diagnostic plot raise no warning flags concerning our modeling assumptions. It is also useful to examine the residuals as a function of the values of each explanatory variable. A QQ plot graphs the quantiles of the residuals against the theoretical quantiles of a Poisson distribution. The plot we see is roughly linear. We see a plot of the square root of the magnitude of the residuals as a function of the fitted values. This diagnostic plot is used to homoscedasticity of the residuals. Again, there are no indications of failures of the assumptions underlying the linear model. The plot of residuals

versus leverage discloses no problems with the fit. Note that, certain points have been labeled by row number in the data frame. These have been flagged as potential outliers and the user might want to examine them in more detail.

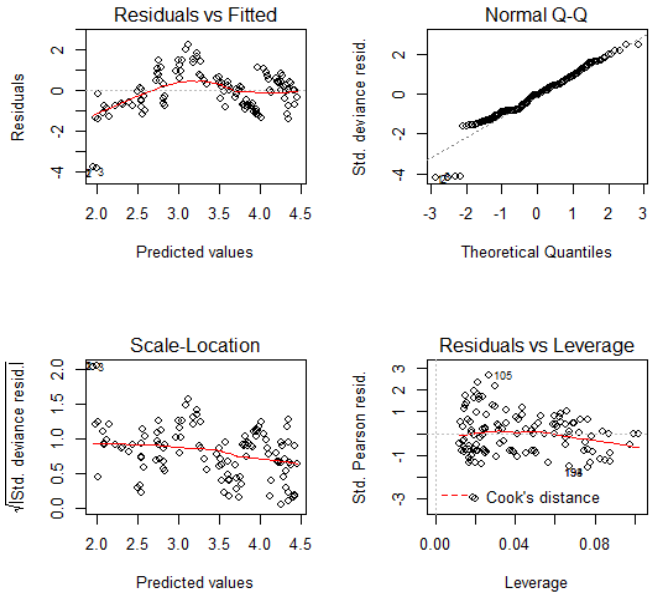

Figure 2: Diagnostic Plots for the Poisson Glm in Glm Object

3.4 Negative Binomial Generalize Linear Model

To do so, we need to estimate the regression parameters and the variance parameter. We use the *glm.nb* function in R to do the frequent approaches on Negative Binomial linear regression model.

	Estimate	Std. Error	z-value	Pr>|z|
Intercept	-0.0148	3.55	-41.72	0.000 ***
AC	0.0037	0.0014	2.67	0.00075**
Di	0.0241	0.0331	0.730	0.4653
Le	0.0008	0.0137	0.06	0.9525
Temp	0.0055	0.0044	1.27	0.2032
Rain	0.0004	0.0001	3.95	0.000 ***
Yi	0.0754	0.0018	41.37	0.000 ***
PVC	0.0035	0.0003	11.83	0.000***
CSC	0.0004	0.0002	1.99	0.046 *
Pt	-0.0451	0.0220	-2.04	0.040 *
		degree of freedom		
Null deviance	3908.26	239		
Residual deviance	220.93	230		
AIC	1480.4			
2 x log-likelihood	-1458.403			
Dispersion parameter	460923			

Table 4: Results of Parametric Coefficients in Negative Binomial GLM

Table 4 shows the parameter estimates of Negative Binomial GLM. The parameter estimates of both Poisson GLM and Negative Binomial GLM are quite similar, this is not unexpected since estimates from both the Poisson GLM and Negative binomial GLM are consistent. When fitting Negative Binomial GLM models to count data with log links, it is helpful to exponentiation the estimated coefficients. The resulting values represent ratios of expected counts on the dependent variable associated with a one-unit increase in a given predictor. In other words, the expected multiplicative increase in the mean of the count response that is associated with a one-unit increase in a given predictor. The parameter of θ is the dispersion parameter $\eta = \sqrt{1/\theta}$ which is 0.001458453 approximately equal to 0 that indicates there is equi-dispersion in Negative Binomial GLM.

3.5 Proportional hazards model

This is done in order to identify statistically significantly factors that influence the survival/break of water distribution system pipe break using Cox proportional hazard model. We use *survival* and $KMsurv$ package in R statistical software to estimate cox model or Proportional hazard model by $coxph$ function for estimating the parameter coefficient and $survfit$ function plot for the assumptions.

	Coef	exp(coef)	se(coef)	z-value	Pr(>|z|)
PVC	0.0069135	1.0069375	0.0020060	3.446	0.000568 ***
AC	0.0172653	1.0174152	0.0103439	1.669	0.095091
CSC	-0.0007549	0.9992454	0.0016735	-0.451	0.651937
	exp(coef)	exp(-coef)	lower .95	upper .95	
PVC	1.0069	0.9931	1.003	1.011	
AC	1.0174	0.9829	0.997	1.038	
CSC	0.9992	1.0008	0.996	1.003	

Table 5: Results of Parametric Coefficients in Proportional Hazard Model

Likelihood ratio test= 15.69 on 3 degree of freedom, p=0.00131, Wald test = 14.81 on 3 degree of freedom, p=0.001989, Score $(logrank)$ test = 15.13 on 3 degree of freedom, p=0.001709. The model will be constructed by first identifying three covariates which are significant at p-value 0.05 in the Cox proportional hazard analysis. Likelihood ratio test, Score $(logrank)$ test, Wald chi-square test is used to test the significance of the model. The result of those tests the model is statistically significant. Table 5 shows that the covariate PVC pipe is highly significant covariate, which has highly significant effect on pipe break. AC pipe also significant at 5% level of significance, it has significant effect on the pipe break, and only non significant covariate is CSC pipe. In table 4 e^{β} which is the hazard ratio and can be interpreted as the predictor change in a hazard for a unit constant in the predictors. There is one assumption for Cox's regression which is the proportional hazards assumption that the hazard ratio between two groups remains constant over time on this study, the assumption is not violated. All possible interactions of the variables of the model are formed to see if the interaction effects can increase or decrease the survival time of pipe break. The Wald test is used to assess the significance of reasonable and possible interactions. Let us begin with baseline PVC pipe of the water distribution system that is supposed to be significant statistically. In this study, PVC pipe has been found to have a significant impact on the pipe break. Furthermore, plotting the scaled Schoenfeld residuals of each covariate against log time will be used to check whether the assumption of proportional hazards is violated or not. Clearly, a close look of this plot indicates that the residuals are random and loss curve are smooth and horizontal with zero slope. On the other hand the curve is fairly flat no pattern or trend on the assumption, the proportionality is not violated. This also suggests that the plots support proportionality assumption to hold. In figure 3, the first left side plot that is beta (t) for PVC pipe versus survival time the line is fairly flat no patter or trend. The other two covariates beta(t) for AC pipe and beta(t) for CSC pipe versus survival time, the curve is approximately flat in the case of these the assumption is not violated. The p-value associated with in the box (PVC, AC, CSC and GLOBAL) is none significant suggesting the proportional hazards assumption be satisfied.

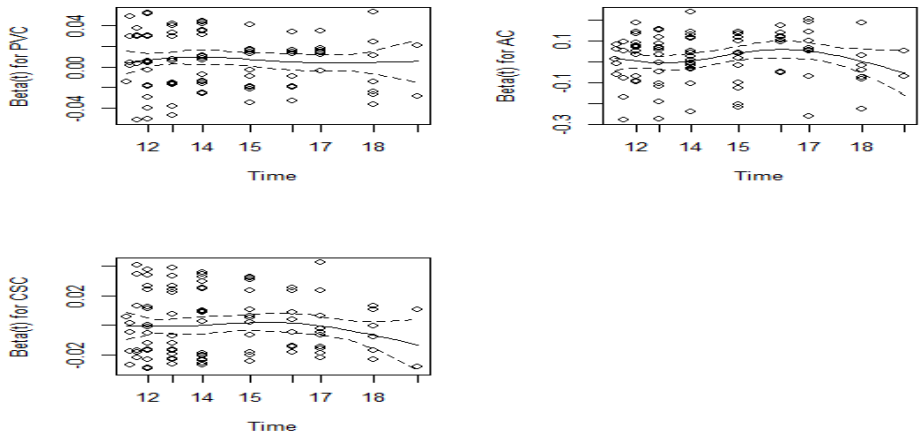

Figure 3: Scaled Schoenfeld Residuals of Proportional Hazard Model

3.6 Model comparison

We analyzed the count data in Hawassa water supply and sewerage service enterprise by using five different models. It is reasonable to ask which type of model is better or when is one to be preferred over the others. These five models are listed in the following:

Model 1 = MLRM (Multiple Linear Regression Model)
Model 2 = TEM (Time Exponential Model)
Model 3 = PGLM (Poisson Generalized Linear Model)
Model 4 =NBGLM (Negative Binomial Generalized Linear Model)
Model 5 = PHM (Proportional hazard model)

Assessment parameter	MLRM	TEM	PGLM	NBGLM	PHM
Deviance			218.4147	219.031	
AIC	1346.517	2187.059	1477.865	1480.501	1211.222
BIC					
LL	-661.2584		-727.9326	-728.2505	-602.6109

Table 6: Comparison of MLRM, TEM, PGLM, NBGLM, PHM

All of these refer to the size of the residual variance, and test whether the residuals are correlated or not. Large values indicate lack of fit, and can be tested against a chi-square distribution for the degree of freedom given. In addition the AIC, BIC, -2log L, R-square, adjusted R-square etc values are given to investigate model adequacy condition. We compare the first two models Model 1 and model 2 by using AIC, Multiple linear regression models has less value (1346.517), and it is chosen by the data analyst. The point is that multiple linear regression model is a better model than Time exponential model. As mentioned the Deviance, AIC, and log-likelihood values in the Table 6: Poisson GLM is 218.4147, 2187.059 and -727.9326 respectively, it is chosen by the data analysis. Poisson GLM is better model than Negative Binomial GLM model. Further, Proportional hazard model is the best model comparing with all other models because AIC and LL values are smaller than the others.

IV. Conclusions

This paper presented a statistical analysis on pipe breaks in water distribution systems in the case of Hawassa, Ethiopia. The analysis on pipe breaks have been explored using Multiple linear regression model, Poisson Generalized linear model, Negative Binomial Generalized linear model, Time exponential model and Proportional hazard model. The analysis shows that the covariates variables such as year of installation, Polyvinyl chloride pipe, Asbestos cement pipe, average temperature, concrete steel cage, and length of pipe and average rainfall are important influential factors which affect the number of pipe breaks. Among these Polyvinyl chloride pipe is the most influential factor. Comparing these models proportional hazard model is appropriate model. The criterion for selection of the best model used is AIC method. Based on our findings one should give due attention to the high risk factors identified as contributing high risk of pipe breaks in the water distribution, emphasis in improving the knowledge of persons who work in Hawassa city water supply and sewerage services enterprise on appropriate installation time of pipe practice and awareness about Polyvinyl chloride pipe and Asbestos cement pipe, to decrease the number pipe breaks. Finally, our recommendation is that water utility managers need more tools to help them make the right decisions about network rehabilitation.

References

[1]. Achim D., Ghotb F. and McManus, K. (2011) Prediction of water pipe asset life using neural networks. *Journal of Infrastructure System*, 131, 26-30.
[2]. Andreou, S.A., (1996). "Predictive models for pipe break failures and their implications on maintenance planning strategies for water distribution systems," PhD Thesis, Department of Civil Engineering, Massachusetts Institute of Technology, and Cambridge MA.
[3]. Ayunanda Melliana, Yeni Setyorini, Haris Eko, Sistya Rosi, Purhadi (2013), The Comparison Of Generalized Poisson Regression And Negative Binomial Regression Methods In Overcoming Over dispersion, *International journal of scientific and Technology research* 2, 8.
[4]. Clark, R. M. Stafford, C. L. and Goodrich, J. A. (2002), Water distribution systems: A spatial and cost evaluation., *Journal of Water Resources Planning and Management*, 108(3), 243-256.
[5]. Jon Rostum. Norway (2000), Statistical Modeling of pipe failure in water networks, Norwegian University of Science and Technology NTNU, Department of Hydraulic and Environmental Engineering.
[6]. Kettler, A.J., Goulter, I.C. (2005), An analysis of pipe breakage in urban water distribution networks, Canadian Journal of Civil Engineering, 12, 286-293.
[7]. M.J. Fadaee and R. Tabatabaei (2010), Estimation of Failure Probability in Water Pipes Network Using Statistical Models. Civil Engineering and Department of Islamic Azad University, World Applied Sciences Journal 11 (9), 1157-1163.
[8]. O'Day, D. K., Fox, C. M., and Huguet, G. M (2010) "Aging urban water systems, a computerized case study" Public Works
[9]. Rajani, B., Zhan, C. and Kuraoka, S., (2012). Pipe soil Interaction Analysis for Jointed Water Mains, *Canadian Geotechnical Journal*, 33(3), 393-404.
[10]. [Shamir.U. and Howard C.D.D. (1979). Analytic approaches to scheduling pipe replacement, Journal of American Water Works Association, 71(5), 248-258.
[11]. Walski, T. M., and Pellicia, A., (2002). Economic Analysis of Water Main Breaks, *Journal of American Water Works Association*, 74(3), 140-147.
[12]. Water Supply and Sanitation in Ethiopia (2015), Turning Finance in to Services Beyond, an AMCOW Country Status Overview.
[13]. World Health Organization and UNICEF (2014), Water supply standards, Sanitation trends, Drinking water supply and distribution and Program evaluation.
[14]. Yaqin Wang, (2012). Estimation of accelerated failure time models with random effects, Retrospective Theses and Dissertations.