WHERE TO REPLACE ASSETS? SPATIAL ANALYSIS ON DIFFERENTIAL AGEING OF LOW VOLTAGE PILC CABLES

Ruben VERWEIJ
Stedin Netbeheer B.V. – Netherlands
ruben.verweij@stedin.net

Dik van HOUWELINGEN
Stedin Netbeheer B.V. – Netherlands
dik.vanhouwelingen@stedin.net

Aad PREIN
Stedin Netbeheer B.V. – Netherlands
aad.prein@stedin.net

ABSTRACT

Experience has shown that a simple aged based replacement strategy for low voltage PILC cables will not allocate funds efficiently. A statistical survival analysis is executed to predict survival probabilities of cables with specific (spatial) characteristics. The predicted probabilities are multiplied by the potential local impact of interruptions to obtain the risk levels. Additionally is shown how a linear optimization would result in an optimized risk reduction strategy.

INTRODUCTION

As a DSO, Stedin owns a large LV cable network. Paper insulated lead cables (PILC) still account for approximately 7.000 kilometers within the Stedin network. Those cables are ageing and will reach the end of the presumed life soon. Experience has shown, that the useful life time of a cable depends on many factors, including soil conditions, excavation damage and number of clients connected. Thus, a simple aged based replacement strategy will not allocate funds efficiently. Ageing and condition assessment of low voltage cables is studied in previous work like [1-3]. From a strategic point-of-view Stedin aims to concentrate major replacement investments in areas with a high failure probability and high impact on key performance indicators like CAIDI, SAIDI and SAIFI. This study is conducted to relate the client oriented objectives with the technical condition of cables and optimize replacement from an Asset Management perspective. The chosen method to study the technical condition of cables is a statistical survival analysis [4-5] to explore the different probabilities of survival of cables in typical circumstances. A semi parametric and a non-parametric technique will be described further in the article. The probabilities of survival are combined with the impact of interruptions to estimate risks. The results give the opportunity to prioritize investments and optimize investment planning.

This analysis includes the following main steps:

1. Data collection;
2. Survival analysis to estimate survival functions S(t) of specific cable populations;
3. Using a Cox regression to investigate the covariate adjusted survival;
4. Model validation and comparison;
5. Geographic visualization of results and optimized investment planning by using results from (3) and (4).

DATA COLLECTION

Registered data in the interruptions database are the starting point for analysis. At Stedin, like many other DSOs, no direct link is available between the interruptions database and the asset database. The first step in the study is to relate historical outages to an asset. This was done by using an algorithm which combines asset information, outage information and the network topology to find the failed asset. When either of the systems does not provide accurate information at the connection point, the algorithm cannot find the asset and a "spatial join" procedure was used. The spatial join procedure finds pairs of multidimensional objects in Euclidean space satisfying a given relation between those objects. In this case the location of the outage is connected with the closest PILC cable which is repaired during the year of the outage. The combination of the techniques yielded information accurately enough for further processing. The pseudocode of the algorithm is presented below:

```
failedCables[]
for outage in outages:
    get postcode, house, year from outage
    get cables which feeds house
    if topology_trace= true
        for cable in cables:
                get year, length, id from cables
                if cable.year == outage.year and
                cable.length < 5 then failedCables =
                cable.id
    else topology_trace= false
        spatial join cables closest
        for cable in cables:
                get year, length, id from cables
                if cable.year == outage.year and
                cable.length < 5 then failedCables =
                cable.id
```

Figure 1 – Algorithm to relate interruptions to an asset in simple pseudo code

Different programming languages (SQL, R and Python) and tools (Anaconda, SQL developer, Rstudio and ARCGIS) where used to complete the data collection. When useful the names of modules (Python) and packages (R) will be named in the article.

SURVIVAL ANALYSIS

Different statistical techniques where used to estimate probabilities of survival for different cable populations. In this paper two techniques are described. The first technique is de well known Kaplan Meier Estimator [4]. This is a non-parametric technique to estimate the survival function S(t) mostly used in Medical Sciences but it is also applicable for Reliability Engineering.

Quality indicator

After the interruptions database and the asset database are connected, further analysis can be done. As a starting point a dummy variable named *quality indicator* is created. Networks with a high, average and low quality are defined based on historic interruptions. The following steps were taken:

1. Define all low voltage networks. The definition of a low voltage network is: a combination of cable connections which are coupled via one (radial) or more (meshed) LV busbar systems;
2. Define all PILC cables per network based on topology or spatial relation using the Python module *Arcpy;*
3. Calculate the sum of PILC cable length per low voltage network;
4. Calculate the number of ageing related interruptions per low voltage network;
5. Calculate overall failure rates $\lambda(t)$ [interruptions/km/installation year] per low voltage network;
6. Calculate the weighted arithmetic mean age $\bar{t}$ of the installed PILC cables per network, here l_i is the length of cable i and t_i the year of installation;

$$\bar{t} = \frac{\sum_{i=1}^{n} l_i * t_i}{\sum_{i=1}^{n} l_i}$$

7. Divide the failure rates with the normalized $\bar{t}$ to obtain a quality indicator called q for all LV networks.

$$\mathrm{t} = \begin{vmatrix} \bar{t}_1 \\ \bar{t}_{..} \\ \bar{t}_n \end{vmatrix}$$

t is the array including the weighted arithmetic mean age for all LV networks.

$$q_i = \frac{\lambda_i}{(t_i - \frac{\min(t))}{\max(t) - \min(t)})^2}$$

The denominator is a quadratic equation because of the assumed increasing failure rate over time. Using the R package *Hmisc,* three equally sized (in terms of total cable length) populations can be defined. The result is visualized in the figure below.

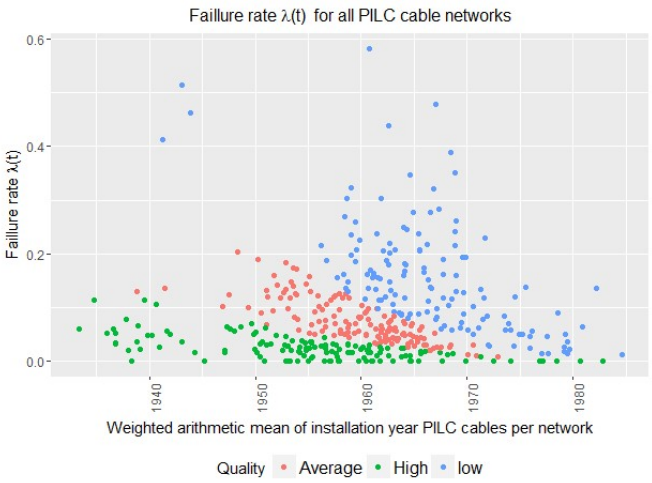

Figure 2- Failure rates for three defined cable populations

In the figure the blue dots are networks with a relative young age and high number of interruptions. The red dots are networks with an average increasing failure rate over time. The green dots are cable networks with a slow increasing failure rate. These networks can be very old and still have a relative low number of interruptions. This *quality indicator* will be included as covariate later in the survival analysis. The figure below shows the geographic visualization of the quality indicator (low, average, high) in Stedin area.

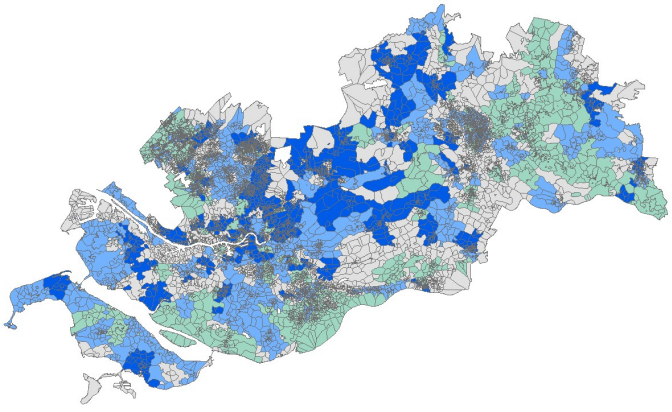

Figure 3 - Geographic visualization of the quality indicator (a low quality equals dark blue, average equals light blue, high equals green, no PILC equals grey)

Kaplan Meier estimator

The Kaplan Meier Estimator [4] is defined by:

$$S(t) = \prod_{t_j<t} \frac{n_j - d_j}{n_j}$$

In this formula $S(t)$ is the estimated survival function and n_j the number of cables which did not fail yet. d_j is the number of failed cables in year j. The technique was used in R by the package *survival* on the Stedin data. $S(t)$ is the probability that the lifetime of a cable in a certain population is greater than the time t. Note that $S(t)$ equals $1 - F(t)$ where $F(t)$ is the cumulative distribution function. The derivative of $S(t)$ is $-f(t)$ and the hazard function $\lambda(t)$ is defined by:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

The hazard function $\lambda(t)$ can be seen as a conditional probability of failure. The condition is that the cable did not fail until *t*.

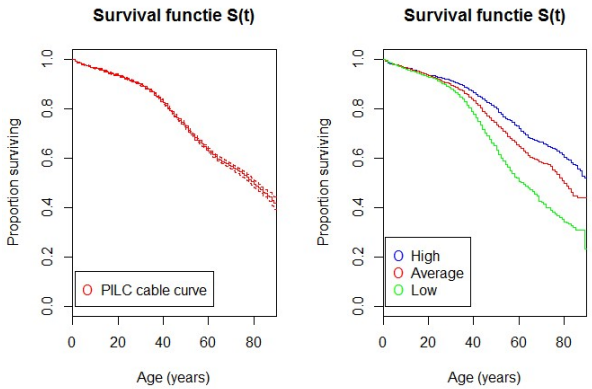

Figure 4 – Survival curves for PILC cables as one population on the left side and the diversification for three cable quality indicators on the right side (blue = high, red = average and green = low).

The summary of statistics of the Kaplan Meier curves in the right plot in figure 4 are presented in the table below:

Table 1- Summary statistics of the Kaplan Meier curves

Quality	Median (year)	0.95 LCL (year)	0.95 UCL (year)
High	100	96	104
Aver.	88	86	83
Low	62	59	65

This outcome means that 50% of the cables in high quality networks have failed in approximately 100 years. For cables in low quality networks this is 62 years. The Kaplan Meier model will be evaluated later in the model validation part. The disadvantage of Kaplan Meier is that it is hard to examine the effect of multiple covariates. This will be done by using Cox regression later in the article.

Log-rank test

In order to test whether the survival probabilities are significantly different, a log-rank test is executed. The log-rank test compares estimates of $\lambda(t)$ for all network qualities (high, average and low) at each observed event time. The log-rank statistic (Z) is defined as:

$$Z = \frac{\sum_{j=1}^{J}(O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}}$$

In this formula j = 1,..,J are the unique times of observed interruptions for each quality. O_{1j} are the observed number of events for group 1. E_{1j} is the expected value under H_0 where O_{1j} has the hypergeometric distribution. . V_j is the variance. More information is to be found in [7]. The P value of the log-rank statistic is nihil in this case which means that H_0 (= the groups have equal survival and hazard functions) can be rejected.

Cox proportional hazard model

Second step in the analysis is examine the impact of different covariates like soil type and ground deformation on the relative hazard. For this purpose the Cox proportional hazard model [5] is useful. The Cox proportional hazard model allows to examine the influence of multiple factors on the survival probability of cables. The following factors have been examined:

Table 2 - Included covariates or predictor variables

Factor	Scale of measure
Year of installation	Ratio
Cable length	Ratio
Number of clients connected	Ratio
Amount of transported energy	Ratio
Number of previous repairs	Ratio
Cables pieces renewed	Ratio
Number of joints	Ratio
Quality indicator	Ordinal
Deformation level	Ordinal
Groundwater level	Ordinal
pH level soil	Ordinal
Soil type	Categorical
Number of trees on trace	Ratio
Average height of trees on trace	Ratio
Average surface of trees on the trace	Ratio
Age variation on cable	Ratio
Urban area	Nominal

Cox proportional hazard model is defined by:

$$\lambda(t|X_i) = \lambda_0(t) * e^{(X_i * \beta_i)}$$

In this formula $\lambda_0(t)$ is the baseline hazard function. This

baseline function describes how the risk of event (failure) per unit time changes over time at baseline levels of covariates. $X_i = \{X_{i1}, X_{i2}, X_{in}\}$ are the realized values of covariates for cable i. β are the coefficients. For this study a backward stepwise variable selection was used. The function for selection is called *fastbw* and is available in the R package *rms*. The Akaike information criteria (AIC) was used to select the covariates for the best model. The AIC is a well-known quality measure for statistical models. The covariates are checked for multicollinearity (include when < 0.8). Below shows the code to fit the regression model in R.

Table 3 – Summary statistics Cox regression model

Covariate	Coefficient	P
Install_year	0.0405978	2.00E-16
Length	0.2818967	1.01E-13
Num_previous_outage	0.3158489	2.00E-16
Num_short_cables	0.1093555	2.00E-16
Num_renewed_cables	0.1104384	2.00E-16
WaterLev_high	0.0702397	4.37E-07
WaterLev_low	-0.0414598	0.00691
Tree_height	0.0158129	2.00E-16
Sigma_age_cable	0.0518380	2.00E-16
Quality_indicator	0.0624635	4.34E-10

Variables with positive coefficients are associated with increased hazard and decreased survival times. An example is cable length, if the length of the cable increases with one kilometer, the hazard rate increases with $e^{(0.2818)}$ 32%. The table above shows significant impact P < 0.05 of the covariates. The concordance is 0.77 which means that the predictive power of the model is average, as 1.0 means a perfect prediction and 0.5 is the same as flipping a coin. The R^2 equals 0.29. This metric reveals how well the Cox model explains the variance in the outcome. A higher concordance and R^2 could be obtained by 1) increasing the asset/interruptions data quality and 2) include higher resolution data on the included covariates.

Model validation

Prediction error curves are computed by the *pec* function to compare predictions of different tested survival models. The integrated Brier score summarizes the prediction error curve and measures the accuracy of probabilistic predictions. More on the mathematical details in [6]. Zero is the best score, one is de worst score. A score of one is obtained when the model predicts a 100% failure probability and in reality (historic values) there is no failure. The reference model is the Kaplan Meier model scoring a 0.161. The predictive performance of Cox.optimal scores best (0.104).

Table 4 – The integrated Brier score (IBS) for four tested models.

Model	IBS
Reference (Kaplan)	0.161
Cox.optimal	0.104
Cox.length	0.160
Cox.previous-outage	0.146

The prediction error curves are shown below. At t=0 no cables have failed yet and the prediction error equals zero.

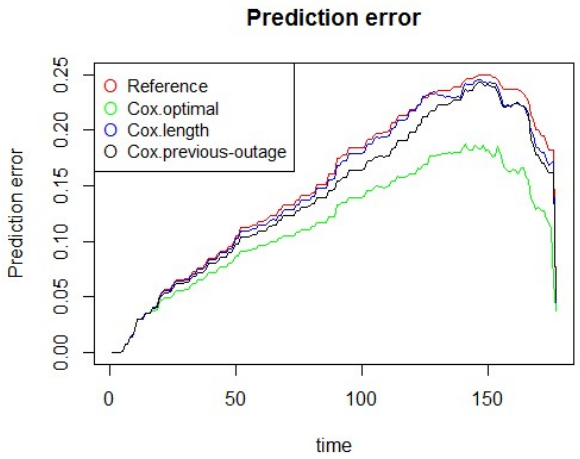

Figure 5- Prediction error curves of four tested models including one Kaplan Meier model and three Cox regression models

Using the *cox.zph* function, of the package *survival* the proportional hazard assumption was examined. It was shown that this assumption holds for all included covariates. Predictions based on the cox regression model where made with the function *predictSurvProb* of the R package *pec*. The survival probabilities are calculated for all individual PILC cables based on the specific variable combination (soil type, year of installation, length etcetera) for any time horizon.

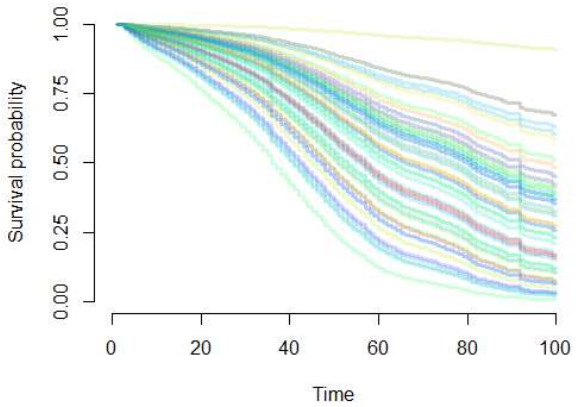

Figure 6- Cable specific survival functions of 50 random cables in the dataset with time in years

OPTIMIZING INVESTMENT PLANNING

The predictions of the Cox model are now combined with the impact estimate per cable. Predicted survival probabilities for a time horizon of 5 years are multiplied with the potential impact. The impact consists of 1) customer minutes lost [minute] 2) societal impact in [euro/client/hour] 3) costs for repair [euro/interruption]. An aggregated visualization of networks with a high impact is presented below:

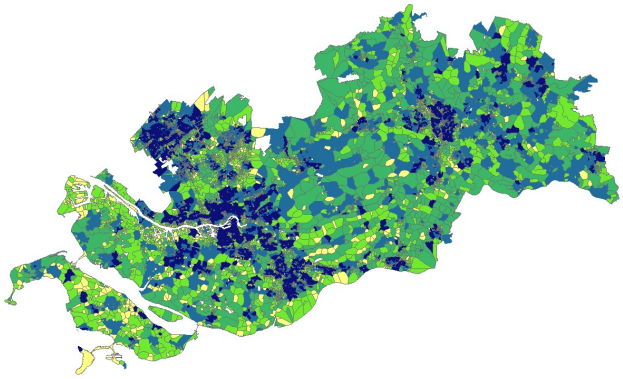

Figure 7- Visualization of LV networks with the highest expected impact of interruptions in dark blue

Risk is calculated as probability multiplied with impact. For optimal financial benefit the risk is divided by the cost of replacing the cable. This results in a risk reduction per million euro. Now it is possible to apply a linear optimization like 1) available budget or 2) on a defined risk border or 3) optimize risk in areas where gas networks are going to be replaced. Below a geographical visualization is presented on a optimization where risk reduction was maximized for a limited budget of 30 million euro spread over five years. The red cables in the figure are then selected to be replaced.

Figure 8 – Where to replace assets? Replacing the red colored cables will result in a maximized risk reduction against minimal cost.

CONCLUSION AND FUTURE WORK

Survival analysis is a suitable method to study survival probabilities and optimize investment planning. There are however limitations to achieve a desired level of detail. The largest limitation is the availability of detailed data on assets (e.g. types, year of installation), interruptions (e.g. year, cause) and network topology. High resolution data on all included covariates is essential to achieve an acceptable outcome. The second limitation is data quality. Improvements on these points are necessary to enhance the predictive power of the models. Stedin accomplished a software system change to integrate the GIS and interruptions database. In 2017, interruptions will be coupled to the *asset id* of the component that failed. This will be an important step to reduce information loss. The less accurate coupling algorithm will no longer be necessary. From a modelling perspective further work is required on the statistical modelling of repair since cables will not be replaced in total when interrupted. An overview is presented in [8].

LITERATURE

[1]. B. Kruizinga, P.A.A.F. Wouters and E.F. Steenis, 2016 “Comparison of polymeric insulation materials on faillure development in low-voltage underground power cables“, 2016 IEEE Electrical Insulation Conference (EIC) (pp. 444-447) Piscataway : Institute of Electrical and Electronic Engineers Inc.

[2]. B. Kruizinga, P.A.A.F. Wouters and E.F. Steenis, 2015 “Accelerated Aluminum Corrosion upon Water Ingress in Damaged Low Voltage Underground Power Cables” Jicable 2015, Versailles.

[3]. B. Kruizinga, P. Wouters et E. Steennis, 2015, "Fault Development on Water Ingress in Damaged Underground Low Voltage Cables with Plastic Insulation, Proceedings IEEE Electrical Insulation Conference (EIC) 2015, Seattle, submitted.

[4]. E.L. Kaplan, and P. Meier, 1958, “Nonparametric estimation form incomplete observations“, J. Almer Statist. Assn. 52 (282) : 457-481.

[5]. D.R. Cox, *1972,* "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B. 34 (2): 187–220.

[6]. U.B. Mogensen, H. Ishwaran, T.A. Gerds, 2012, “Evaluating Random Forests for Survival Analysis Using Prediction Error Curves”, Journal of statistical software, Vol. 50, issue 11

[7]. N, Mantel 1966, “Evaluation of survival data and two new rank order statistics arising in its consideration”, Cancer Chemotherapy Reports, 50, 163-70

[8]. B.H. Lindqvist, 2007, “On the Statistical Modeling and Analysis of Repairable Systems“, Statistical Science, Vol. 21, No. 4, 532-551.