Machine Learning Methods for the Health-Indexing and Ranking of Underground Distribution Cables and Joints

***Jacco Heres*[*], *Roel Stijl*[†], *Frank Reinders*[*]**

[*]*Alliander N.V., the Netherlands, e-mail: {jacco.heres, frank.reinders}@alliander.com,*
[†]*BearingPoint B.V., the Netherlands, e-mail: roel.stijl@bearingpoint.com*

Keywords: Condition management, Underground Grid, Machine Learning, CBAM, Condition Based Assetmanagement.

Abstract

An aging asset population and a less predictable volatile electricity consumption and production pattern urge DSOs to get insight in the condition of their medium voltage (MV) and low voltage (LV) networks. Because visual inspections of underground networks are impossible and the number of measurements is still very limited, this paper proposes a method to rank underground assets by looking for trends and patterns in historical outages with help of Machine Learning methods.

Nine years of outages of MV and LV cables and joints in the network of a large Dutch DSO are analysed. A model is developed that couples each outage to the asset most probable responsible. Twentytwo different datasets are coupled with the asset database, ranging from load estimates of the asset to distance-to-a-railway. Each set could contain data that explains or correlates to some of the outages. Several Machine Learning techniques are benchmarked.

The final model, created by the Random Forest algorithm, is applied to rank current assets. It is operational to determine the positioning of an online monitoring system in the DSO's MV network.

1 Introduction

The main challenge for European DSOs will be the integration of decentralised and variable generation and new loads into the current ageing infrastructure [1]. As this energy transition is mainly due to the adoption of devices such as solar panels and electric vehicles, more insight in the condition of MV and LV networks is necessary. A large part of the electricity grid in Western Europe has been built during the post-World War II industrial investment recovery and economic expansion. These parts approach or already passed their technical lifespan of approximately 50 years. As 41% of the MV and 55% of the LV lines in Europe are underground [1], an important part of the challenge in this is how to monitor, refurbish and replace the underground cables and joints. Until now, the strategy for most underground assets is Run-to-Fail. There is normally no possibility to inspect the assets visually, and it is too costly to replace large parts of the network in advance or to place advanced monitoring systems throughout the grid.

In this paper an approach is proposed to rank underground cables and joints by their probability to fail. The key part is a Machine Learning method that learns from previous outages and data from the asset involved in the outage. The analysis is performed on Alliander's 42.000 km of MV and 74.000 km of LV network. Alliander is the largest DSO in the Netherlands, which has almost all of its distribution network underground. The analysis is performed on all of the individual cables and joints that were active between 2007 and the end of 2015. The philosophy and background of this approach is in line with the approach used in the Machine Learning project at ConEdison [2], [4], [5], [6]. There are, however, some important differences in the used data and the precise Machine Learning method.

In this paper we will first describe the creation of the collection of hypotheses of asset failures, which was the starting point in the search for different dataset that could help to predict outages. Following, the method to find the assets that have shown an outage in the past is described. Afterwards, the benchmark of different Machine Learning methods and its evaluation is discussed. Finally, we show some of the results and first applications of the predictions of the model.

2 Hypothesis creation and data collection

It is well known what the dominant degradation processes are for cables and joints [7], [8]. What is usually not known is how the possibility of certain degradation processes effects the probability of failure of an asset quantitatively. Equally important is that for almost all degradation processes not all the relevant data is available. The information from domain experts and literature is therefore used to generate hypotheses (see figure 1, step 1 and 2). These hypotheses are then used to start looking for relevant and available data (3). For instance, it is known that subsidence of the soil can cause bending of the conductor which in turn can cause a short-circuit. However, data about the exact subsidence of the soil is not available. Alliander currently only has a dataset that gives a classification for the severity of the subsidence for each area, this classification is a discrete number between 0 and 8. This dataset would be unusable for a physical model, but it is usable for a Machine Learning approach. Loads are estimated via the method described in [3].

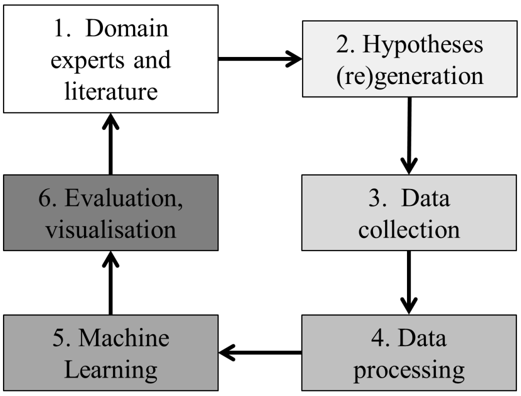
1. Domain experts and literature
2. Hypotheses (re)generation
3. Data collection
4. Data processing
5. Machine Learning
6. Evaluation, visualisation

Fig. 1: Process diagram of the loop used for building and improving Machine Learning models.

After the data sources have been selected and collected, it must be interpreted, filtered, cleaned, transformed, scaled an coupled in the right way (step 4). These step are obligatory to create a dataset suitable for high-performance Machine Learning methods. Usually this step requires the most effort, and it occurs often that findings later on in the process urge to return to the source data and domain experts to clarify certain aspects. On the other side, investing in this part of the process can hugely improve the resulting model, as is also stated in [2]. After the Machine Learning step (5), results are evaluated, creating an learning-loop for continuous improvement. The current models are evaluated in different ways, and the results and underlying data are visualised (6). Not only the performance of the model matters, but also the relative importance of the different parameters. These can be used to focus on the most important data to be cleaned or improved. Finally, the visualisation contains univariate plots of the relations between the variables, the percentage of failed assets and the predicted probability of failure, and maps of the assets with known and unknown data. The visualisation of the data and results can help to identify outliers and gaps in the data. Anti-causal relations, that are unwanted in the analysis, can be singled out.

3 Determining historical outages

A major challenge in the data processing step is the identification of assets that have shown a failure in the past. This data was not stored in the databases. The only geographical information that is available are the addresses from people who made a phone call to Alliander to report the outage. Sometimes some extra information about the location in the grid is available, e.g. the two MV substations between which an outage has occurred. Furthermore, the date of the outage is known, and some information about the type of asset involved in the outage. This last information is known to be not fully reliable. The search for the asset belonging to an outage is done by a Asset-Blackout Connection(ABC) method. This method tries to couple an outage to one or a few assets that were very probably modified or removed due to an outage. As underground electrical assets are mostly non-repairable after an outage, this has proven to be a relatively reliable method for finding failed assets, as these modifications of assets are stored in the database. Alliander keeps a monthly snapshot of all assets in its database, including length, type, location and status. Changes between these snapshots reveal the modifications of the network. These changes can be due to outages, but most changes are related to other construction works, such as creating more capacity or adding new customers to the network.

With the data of all these changes at hand, the search of assets involved in an outage can be performed.

1. Start at a customer that called Alliander and reported an outage, the extra information is also used here.
2. Search for changes in asset configuration data. This search starts at the postal codes in which customers with outages are located, then a geo-bounded box is drawn around all customers with outages, and finally customer connections are linked to sections and feeders (see fig. 2).

Usually there are many more changes within an area then only the relevant ones, so the next steps are to filter the changes to only keep the changes that are involved in an outage.

3. The first filtering step is done by keeping only changes that are registered within 70 days after the outage, and only assets that have been replaced by others and not have been removed entirely.
4. The remaining changes are ranked in a system that assigns points for the different search methods that have found this specific change and the validity of the asset type according to the asset type description of the outage. Finally the number of points is divided by the number of assets found for a single outage. Only assets that surpass a certain threshold are selected as failed. This threshold is determined by applying the method to small validation set of trusted ABC's.

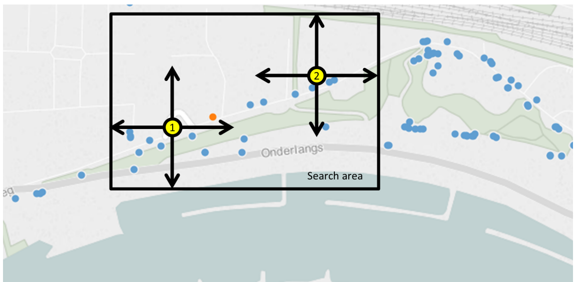

Fig. 2: Search for asset modifications (blue + orange) around customers adresses (yellow) that are known to have reported an outage. This is the geographical search method that is used besides the postal code and a method that uses the grid topology.

4 Benchmarking Machine Learning methods

Now the target variable of failed assets has been set, Machine Learning methods can be applied. Since a wide range of Machine Learning methods is available, and applying new ML algorithms is relatively easy when a suitable dataset is available, a benchmark is executed. The following Machine Learning methods have been tested:

- Decision trees learning, which splits the data according to the given variables in different parts and hereby creates a tree to classify the data as failed or non-failed [9].
- Logistic Regression, which fits a generalised linear model to the data [10].
- Random Forest, an ensemble model of decision trees on subsets of the data and variables [11].
- HyperCube, an algorithm that divides the high-dimensional space in hypercubes where the ratio of failed/non-failed assets is high as possible [12].
- Support Vector Machines, that transforms the high-dimensional space so that a linear model can be applied [13].
- Neural Networks, that transforms the variable-space with a single-hidden-layer [14].

These Machine Learning algorithms are all trained on the same data, with the same parameters that could possible help to predict failed assets. The model evaluation is then performed on a random selected test set that was not part of the train set. The model predicts for each asset a probability of failure. After that the assets are ordered according to their probability of failure. The sensitivity is calculated as a measure of the model performance. This measure, also called true positive rate, is defined as

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{total number of failed assets in testset}}. \quad (1)$$

This measurement is evaluated at different rates of the specificity or true negative rate, defined as

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{total number of nonfailed assets in testset}}. \quad (2)$$

The ultimate aim is to have a sensitivity and specificity of both 100 %, but for realistic models there is a trade-off. One therefore wants to have high sensitivity and specificity rates combined. For grid-operators, the most important thing is to be able to implement measures such as replacement and monitoring as purposeful as possible. Therefore the model evaluation is focussed on the Machine Learning methods to select as many failed assets as possible while only allowed to select a small part of the total asset population. This means that the specificity has to be high. In this benchmark it is chosen to be 99%, 95% and 90%. Finally, the area under the ROC-curve, which gives a measure of the total model performance, is calculated [15]. The results are shown in table 1. It is clear that the Random Forest algorithm outperforms all other Machine Learning methods on all measures evaluated. Apparently it is most suited to combine many local effects in the 20 to 25 different variables. The number of variables varies for MV/LV cables and joints. In total the algorithm creates 500 decision trees, each tree uses only part of the assets (ca 68 %) and only 5 variables. The prediction for a new asset is the average of the prediction of all these 500 trees. Because of the huge difference in model performance, the final model is chosen to be the model build by the Random Forest algorithm.

Measure	RF	HC	DT	LR	SVM	NN
Sensitivity at 99% spec.	24%	13%	10%	5%	8%	8%
Sensitivity at 95% spec.	47%	29%	31%	23%	31%	29%
Sensitivity at 90% spec.	61%	45%	47%	39%	48%	49%
Area under ROC curve	88%	81%	80%	79%	82%	82%

Table 1: Benchmark of Machine Learning methods at a standardised and more balanced dataset. The sensitivity at x% specificity is given, as is the total area under the ROC-curve. The benchmarked methods are: Random Forest (RF), HyperCube (HC), Decision Trees (DT), Logistic Regression (LR), Support Vector Machines (SVM) an Neural Networks (NN).

4.1 Dealing with class imbalance

A problem that quickly arises when analysing grid outages with Machine Learning techniques is the class imbalance between failed and non-failed assets. With the ABC-method it was possible to couple circa 80% of the outages within to an asset, but still this meant that only 0,4 % of the MV joints in the dataset had shown an outage within the timewindow of nine years. Most Machine Learning methods however, perform best on more balanced datasets. For instance, decision trees will reach their minimum node size that is still statistically relevant only after a few splits. This problem could be dealt with by creating a more balanced dataset out of the original, and choose only one in twenty non-failed assets. There are techniques available that can do this quite effective, for instance SMOTE [16]. The best and most robust result were received with Random Forest, that can deal with imbalanced dataset more naturally. For each of the 500 trees it selects 68% of the failed assets, and for each failed asset it picks 10 assets that have not shown a failure. Because this is done 500 times, the algorithm uses all assets multiple times, only the failed assets are used more often than the non-failed ones. One has to treat the predicted probabilities of failure with care, because the absolute value is effected by this stratification.

5 Results and applications

The current result is a ranking according to the probability to fail for each individual cable and joint in the MV an LV network. According to the model performance checks, 25% to 35% of the outages are taking place in the 1% of the asset population. In this procedure the results of the ABC-method have to be assumed fully correct. With the available data it is thus not possible to predict exactly where outages will occur next year. It is however feasible to select areas in which the probability is much higher than others, instead of a Run-to-Fail policy for every asset. This result can be used for a wide range of applications.

At Alliander the asset ranking as outcome of the Machine Learning model is currently in use to optimally place an online monitoring system in the MV network. This system, called Smart Cable Guard [17], monitors the number and size of partial discharges of MV cable circuits. By measuring the partial discharges, the occurrence of failures can be eventually predicted and prevented. When combining the frequency of failure with the number of customers effected by a possible failure, a risk based selection of cable circuits can be made. At Alliander a first version of an application (see fig. 3) is in use. In this application, operators can select cable circuits that are at high risk according to the predictions. They are provided with some extra information, such as the type of installation at the MV substations, and the cable configuration, which are needed as boundary conditions to place a Smart Cable Guard system.

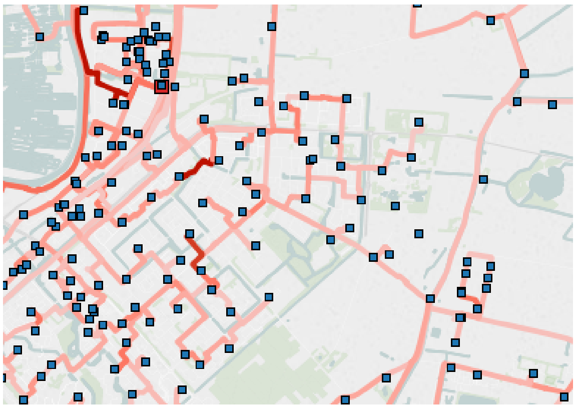

Fig. 3: Sample of the map of the grid health that is in use to place an online monitoring system. The MV substations are shown as blue squares, the colours of the connections are related to the probability of failure, darker red means worse.

6 Conclusions

A Machine Learning method for the ranking of underground distribution cables and joints is built and implemented. Critical steps were cleaning and coupling the data, a separate model to determine the assets that have failed in the past had to be constructed. Via benchmarks the best algorithm to construct models is selected. The models are capable of selecting parts of the asset population in which the probability of failure is more than an order of magnitude higher than the average. These models can be used to prioritize preventive measures such as the positioning of an online monitoring system or proactive replacement of vulnerable assets.

Acknowledgements

We would like to thank Michiel Musterd, Jerom Janssen, Werner van Westering, Maarten Wolf and Tongyou Gu for their valuable contributions and comments to our work.

References

[1] Eurelectric, 2013, “Power Distribution in Europe” www.eurelectric.org/media/113155/dso_report-web_final-2013-030-0764-01-e.pdf, accessed 22-01-2016

[2] C. Rudin et al., 2012 “Machine Learning for the New York City Power Grid”, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2

[3] W. van Westering et al., 2016 “Assessing and Mitigating the Impact of the Energy Demand in 2030 on the Dutch Regional Power Distribution Grid”

[4] P. Gross et al., 2006. “Predicting electricity distribution feeder failures using machine learning susceptibility analysis.” *The Eighteenth Conference on Innovative Applications of Artificial Intelligence*. Boston, Massachusetts

[5] P. Gross, A. Salleb-Aouissi, H.i Dutta, A. Boulanger, 2009. “Ranking Electrical Feeders of the New York Power Grid” *International Conference on Machine Learning and Applications*, Miami Beach, FL

[6] C. Rudin et al., 2014 “Analytics for Power Grid Distribution Reliability in NYC” *Interfaces*, vol. 44 no. 4

[7] E.F. Steennis, 2007 “Dominante degradatieprocessen van veel gebruikte middenspanningsmoffen in Nederland” *Kema*, 70745300-TDT 07-63295B

[8] Cigré, 2012 “Final Report of the 2004 - 2007 International Enquiry on Reliability of High Voltage Equipment”, 509

[9] L. Breiman et al., 1984 “CART: Classification and Regression Trees”. *Wadsworth Press*.

[10] D.R. Cox, 1958 “The regression analysis of binary sequences”. *J Roy Stat Soc B* vol 20: 215-242

[11] L. Breiman, 2001, “Random Forests, *Machine Learning*, vol. 45, no. 1, pp. 5-32.

[12] BearingPoint, http://www.hypercube-research.com/, accessed 14-03-2016

[13] H. Drucker et al., 1996 “Support Vector Regression Machines, *Proc. Advances in Neural Information Processing Systems*, vol. 9, pp. 155-161.

[14] W.N. Venables and B.D. Ripley, 2002 “Modern Applied Statistics with S”. Fourth edition. *Springer*.

[15] A.P. Bradley, 1197 “The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms” *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159.

[16] N.V. Chawla et al., 2011 “SMOTE: synthetic minority over-sampling technique.” arXiv:1106.1813.

[17] E.F. Steennis et al., 2014 “Smart Cable Guard for PD-online monitoring of MV underground power cables in Stedin’s network”, *International Conference on Condition Monitoring & Diagnosis*