IDENTIFICATION AND TREATMENT OF OUTLIERS IN FAILURE DATASETS – PRACTICAL ASPECTS OF STATISTICAL ANALYSIS

Lukasz Chmura[1*], P.H.F. Morshuis[1], E. Gulski[1], J.J. Smit[1] and Anton Janssen[2]
[1]Delft University of Technology, The Netherlands
[2]Liander – The Network Operator, Arnhem, The Netherlands
*Email: L.a.chmura@tudelft.nl

Abstract: Statistical analysis of failure data is a useful tool for assessing the condition of a population of high voltage components such as transformers, joints, bushings and cables. More specific, the results of the analysis give valuable insight into the future failure behaviour of the population of such components. Prior to the analysis, information about the age of the components that are still in operation, as well as time-to-failure data of the failed units need to be collected. Then, this information is used as an input for the analysis, and the proper statistical distribution model is fitted to the data. Using the model with properly fitted parameters a prediction can be made of the future failure rate. For a proper analysis, it is important to know how to deal with failures that do not seem to fit well the statistical distribution chosen - outliers. These outliers may affect the analysis considerably. In our paper we will describe how to deal with this problem in practice. We will use common techniques for identification of outliers and analyze the consequences of leaving out these outliers from the analysis. A case study is presented based on real failure data obtained on a population of high voltage transformers in the Netherlands.

1 INTRODUCTION

Statistical analysis of failure data is often considered with respect to the maintenance policy of electrical components, i.e. for transformers, joints, cables and bushings. In these cases, knowledge is necessary about replacement, refurbishment or back-up units in a case of failure. Taking into account past experiences with failures in certain populations of components, statistical analysis of lifetime data obtained for particular components proves to be a very powerful tool. This is true in particular, when information about future failures is desired. For a proper statistical analysis, the proper data need to be supplied. It means that for the particular population that is to be investigated, the data ought to fulfil the following requirements [1]:

- Homogeneity – the data have to be drawn from a single population, i.e. all potential parameters affecting the population must be kept constant for the whole time period of time.
- Independency – the data for different subjects are independent.
- Randomness – every outcome (e.g. time-to-failure) is equally likely to occur within the considered population
- Sufficient amount of data – the amount of failure information has to be sufficiently large in order to enable drawing conclusions

For a series of failure data presenting the mentioned features, the analysis can be performed. Ultimately, conclusions are drawn with respect to the whole population, based on the investigated sample of failure data.

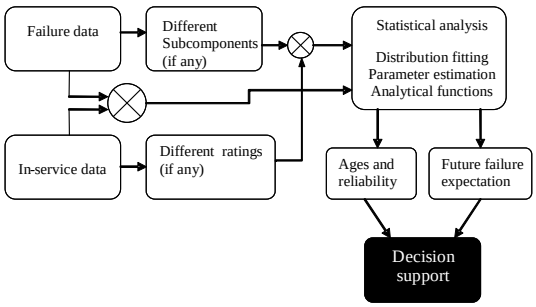

Figure 1: Schematic showing how data is processed for statistical analysis

However, during the analysis, some additional problems are met. When considering the times-to failures of the devices, it may happen that in some particular cases, points may not be members of a population, because they display either a very low, or a very high value with respect to the rest of population, and so they are influencing the results of the analysis by having a large impact on the parameters estimation. Evidently, deleting data points without particular reason and evidence should be avoided. This is particularly true, when dealing with small amounts of failure data. In such

a situation, during the analysis, the question arises on which factors should it be decided whether some points should be retained or rejected from the analysis as not belonging to the data-set. Based on the example of a population of high voltage transformers, the influence of such points on the overall analysis and on the results of future failure expectation are presented.

2 FAILURE AND IN-SERVICE DATA SUBJECTED TO THE ANALYSIS

Here, an example is presented of the analysis of a population of high voltage transformers belonging to Liander is presented. The failure data have been taken from the database of the utility, which contains entries since 1975. However, for some entries the moment of failure is either not exact or not known. Thus, for the population of transformers in total 52 failures are analyzed. The exact information, with respect to the number of failures and accompanying ages at the moment of failure is presented in Figure 2.

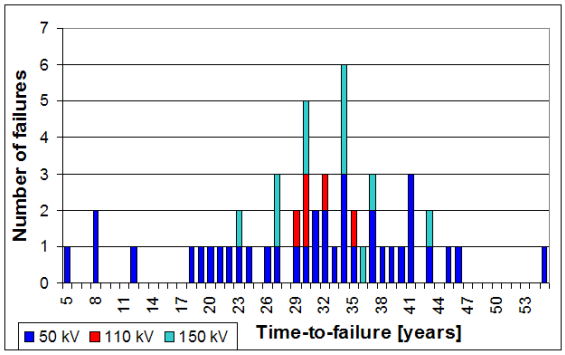

Figure 2: Number of failures registered for transformers of a particular voltage level.

The considered population of transformers can be divided into two main groups according to the voltage ratings, namely:

- Group I – Transformers of 150/50/10, 150/50, 150/20, 150/10 110/20 and 110/10kV
- Group II – Transformers of 50/10 and 50/6 kV

From Figure 2, it follows that the number of failures occurring for transformers of 150 kV and 110 kV is limited. For that reason, the failures for those transformers are taken together for the needs of statistical analysis. The failures can have different failure modes, such as failures of tap-changers, short-circuit in winding or in bushing or leakage of oil from main tank. However, here a further distinction will not be made. Considering the in-service population of transformers operated by the utility, the number of transformers installed in particular years with distinction to rated voltage can be seen in Figure 3.

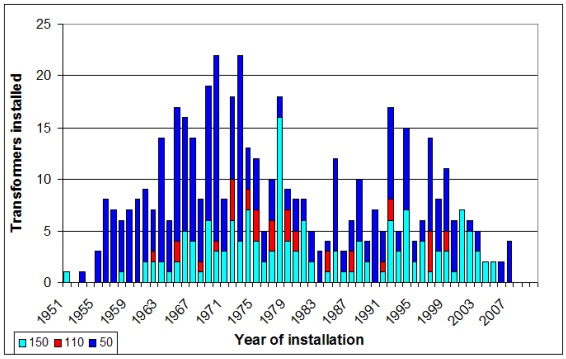

Figure 3: Number of installed transformers in particular years, discerned according to the voltage level, being operated by utility.

In total, the in-service population consists of around 500 transformers. The age span of the population is from 3 to 60 years. The average age of the transformers is 32 years. From Figure 3 it can be noticed that the majority of the transformers was installed before 1980's. The rated power of the transformers is between 10 and 175 MVA.

3 STATISTICAL ANALYSIS OF FAILURE DATA

For the statistical analysis, the failure and in-service data is taken into account. For the failed transformers, times-to-failure are considered, and for in-service transformers the actual ages. Regarding the status of the transformer after failure occurrence, a remark has to be made here. Namely, in some cases, the transformer can be either lost after failure occurrence. In other cases, it can be repaired and brought to service after long reparation, if damages are not severe. For the needs of analysis, although such information was available, no distinction has been made. This is due to small amount of failure data for population under consideration. Thus, all failure information was treated as regarding failures leading to failure. This assumption implies, that the failure expectation as an outcome of the analysis will be related to the failures leading to the loss of transformer. However, even repairable major failures cause unavailability of the transformer for a long time. The immediate replacement of failed unit is necessary. In this way, it can be seen that this assumption has no significant influence on the result of analysis.

3.1 The whole population of transformers

In order to get the overall picture of the failure occurrence, the whole population is investigated. This is done by using information from Figure 2 and Figure 3. The data of the population can be

fitted with a 2-parameter Weibull distribution. The parameters are: β=2.39 and η=94. The cumulative density function with 90% confidence bounds can be seen in Figure 4.

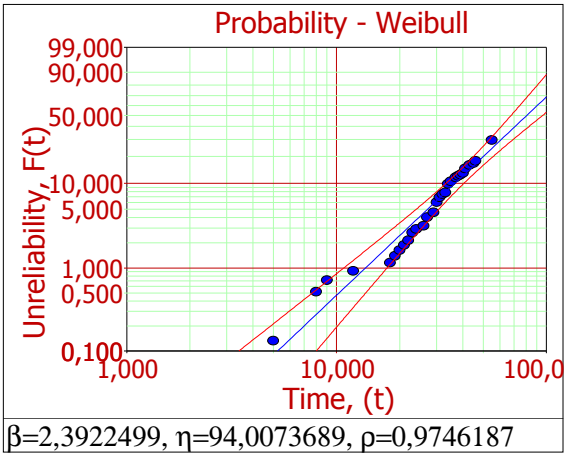

Figure 4: CDF of 2-parameter Weibull distribution with 90% confidence bounds obtained for the whole population of transformers.

From Figure 4 it follows that the first four points (failures occurring at the ages of 5, 8, 9, 12 years) ale not located well on the CDF line, and they are suspected as being outliers. Here a simple method will be presented in order to see if these points can be truly regarded as outliers [2]. The first step is to obtain the shape parameter (β), the scale parameter (η) and the mean life ($\bar{T}$) for the failed population. Only failure data has to be considered and all suspensions are to be neglected. The values of the parameters for the population are as follows: β=2.65, η=34 years and $\bar{T}$=30 years. Secondly, the standard deviation of the population has to be computed with (1)

$$\sigma_T = \eta\sqrt{\Gamma\left(\frac{2}{\beta}+1\right)-\left[\Gamma\left(\frac{1}{\beta}+1\right)\right]^2} \quad (1)$$

Where Γ is the gamma function, the value of which can be computed for particular arguments with available software packages. Substituting the obtained values into (1) a value of σ_T=12.72 is obtained. Finally, the approximate values of lower (2) and upper (3) limits for the failed population are obtained, for which the respective lower or higher times-to-failure may be considered outliers.

$$t_{lower} = \bar{T} - k\sigma_T \quad (2)$$

$$t_{upper} = \bar{T} + k\sigma_T \quad (3)$$

Where: t_{lower} and t_{upper} are lower and upper limits respectively, and k is a factor which value depends on the confidence level (CL) of the points to be considered as outliers. If *CL=99%* then *k=2.3264*

CL=95% *k=1.645*
CL=90% *k=0.1282*

Recalculating, it is found that the failures occurring at 5, 8 and 9 years can be rejected with a confidence level of 95%. Also the failure occurring at the age of 12 years can be rejected with a confidence level slightly lower than 95%.

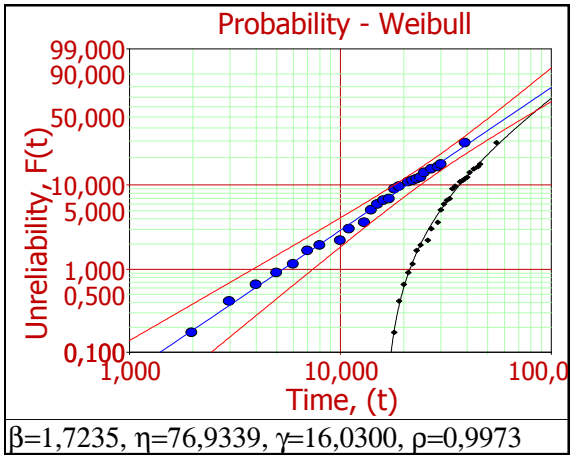

Figure 5: CDF of 3-parameter Weibull distribution with 90% confidence bounds.

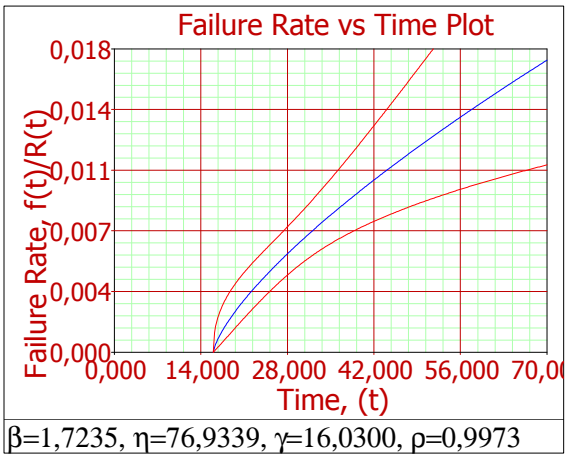

Figure 6: Failure rate versus time function of 3-parameter Weibull distribution with 90% confidence bounds.

Removing the data points mentioned from the considered population, as they are considered to be outliers will result in the creation of a new distribution with the parameters of β=1.73, η=77 years and γ=16 year The details can be seen in Figure 5 and in Figure 6. The introduction of the third parameter can be explained by the fact that in the new distribution there are no failures at an age lower than 18 years. However, a remark has to be

made here. As can be seen from (2) and (3), in the method used here the scatter of the data is taken into account. In other words, it is observed how far the mentioned point differs from the mean value, for the given confidence level. Beside mentioned calculations, it is also good to check the location of the point on the probability net and their correlation to the CDF line. This can be observed for the failure occurring at the age of 55 years, as it should be rejected as outlier with 95% confidence level. However, it can be seen that it fits very well the new distribution chosen where the first four failures are removed.

3.2 Transformers belonging to group I

Here, the failures occurring of group I are analyzed. In total, 16 failures are considered for this group of transformers. Using the in-service and failure data, as presented in Figure 2 and Figure 3, a 2-parameter Weibull distribution is fitted to the data. The parameters of the distribution are β=3.2 and η=71.15 years. The accompanying 90% confidence bounds are presented in Figure 7.

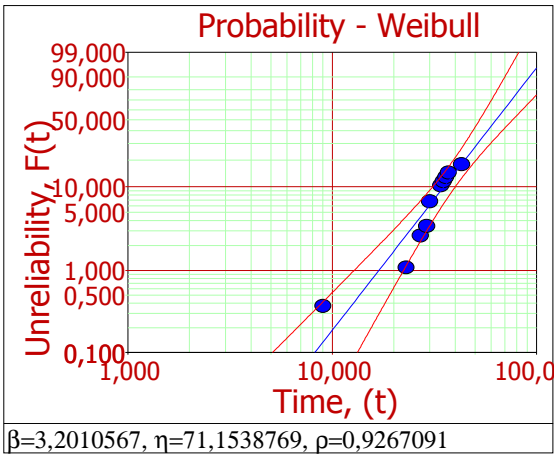

Figure 7: CDF of 2-parameter Weibull distribution with 90% confidence bounds obtained for the group I of transformers.

When taking a closer look at the CDF function, it can be seen that first point is not well located on the CDF function and is strongly influencing the distribution. Thus it might be considered as being outlier. Here, another method of identifying outliers, so called Natrella-Dixon test will be presented [3]. The advantage of this method, is that neither mean life nor standard deviation has to be known. After the data has been collected, the first step is to arrange all times-to-failure in ascending order, as given by (4)

$$T_1 < T_2 < T_3 < \ldots < T_{N-1} < T_N \quad (4)$$

Where: N is the sample size, number of units failed, including suspected outliers. For the given sample size N, a critical value r_{ij} has to be computed, if:

$3 \le N \le 7$ compute r_{10} (5)
$8 \le N \le 10$ compute r_{11} (6)
$11 \le N \le 13$ compute r_{21} (7)
$14 \le N \le 25$ compute r_{22} (8)

Where r_{ij} is calculated according to the sample size and given by the respective formula, as presented in Table I.

Table 1: Computation of r_{ij} with respect to the sample size and location of the suspected point.

r_{ij}	If T_N is suspect	If T_1 is suspect
r_{10}	$(T_N - T_{N-1})/(T_N-T_1)$	$(T_2 - T_1)/(T_N-T_1)$
r_{11}	$(T_N - T_{N-1})/(T_N-T_2)$	$(T_2 - T_1)/(T_{N-1}-T_1)$
r_{21}	$(T_N - T_{N-2})/(T_N-T_2)$	$(T_3 - T_1)/(T_{N-1}-T_1)$
r_{22}	$(T_N - T_{N-2})/(T_N-T_3)$	$(T_3 - T_1)/(T_{N-2}-T_1)$

For the group I of transformers (150 kV and 110 kV) sixteen failures are analysed. By arrangement in ascending order, the following population with times-to-failure in years is obtained: $T_1..T_N$={9, 23, 27, 27, 29, 30, 30, 30, 34, 34, 34, 35, 36, 37, 43}. The sample size N=16, then, as indicated by (8), r_{22} has to be calculated with the formula given in Table 1, where the first data point in the population is suspected to be an outlier. Substituting particular values into the formula, a value of r_{22}=0.667 is obtained.

Table 2: Criteria for rejecting suspected observations using the Natrella-Dixon test

	N	Critical values at the probability levels of α/2						
		0.30	0.20	0.10	0.05	0.02	0.01	0.005
r_{10}	3	0.684	0.781	0.886	0.941	0.976	0.988	0.994
	4	0.471	0.560	0.679	0.765	0.846	0.889	0.926
	5	0.373	0.451	0.557	0.642	0.729	0.780	0.821
	6	0.318	0.386	0.482	0.560	0.644	0.689	0.740
	7	0.281	0.344	0.434	0.507	0.586	0.637	0.680
r_{11}	8	0.318	0.385	0.479	0.554	0.631	0.683	0.725
	9	0.288	0.352	0.441	0.512	0.587	0.635	0.377
	10	0.265	0.325	0.409	0.477	0.551	0.597	0.639
r_{21}	11	0.391	0.442	0.517	0.576	0.638	0.679	0.713
	12	0.370	0.419	0.490	0.546	0.605	0.642	0.675
	13	0.351	0.399	0.467	0.521	0.578	0.615	0.649
r_{22}	14	0.370	0.421	0.492	0.547	0.602	0.641	0.674
	15	0.353	0.402	0.472	0.525	0.579	0.616	0.647
	16	0.338	0.386	0.454	0.507	0.559	0.595	0.624
	17	0.325	0.373	0.438	0.490	0.542	0.577	0.605
	18	0.314	0.361	0.424	0.475	0.527	0.561	0.589
	19	0.304	0.350	0.412	0.462	0.514	0.547	0.575
	20	0.295	0.340	0.401	0.450	0.502	0.535	0.562
	21	0.287	0.331	0.391	0.440	0.491	0.524	0.551
	22	0.280	0.323	0.382	0.430	0.481	0.514	0.541
	23	0.274	0.316	0.374	0.421	0.472	0.505	0.532
	24	0.268	0.310	0.367	0.413	0.464	0.497	0.524
	25	0.262	0.304	0.360	0.406	0.457	0.487	0.516

Here, similar to the previous case, it is investigated whether a given point is an outlier with a given

confidence level (CL). Firstly, CL=99% is investigated. If:

$$CL=99\%=0.99=1-\alpha/2 \quad (9)$$

Then $\alpha/2=0.01$.

For the given values of N and $\alpha/2$, the value of $r_{\alpha/2}$ has to be found from Table2, in this case

$$r_{\alpha/2}=0.595$$

The final step is to compare the obtained value of r_{ij} with $r_{\alpha/2}$. If $r_{ij} > r_{\alpha/2}$ then the suspected point can be rejected from the analysis with the given CL. Here:

$$r_{22}= 0.667 > r_{0.01}= 0.595.$$

Thus, it can be stated that the failure that occurred at the age of 9 years can be rejected from the analysis, as being outliers with the CL of 99% in the given data-set. Rejecting the suspected point from the analysis will result in obtaining a new 3-parameter Weibull distribution. The parameters of the distribution are β=1.74, η=57 years and γ=20 years. The introduction of the third parameter can be explained by the fact that the earliest failure considered in the dataset occurred at the age of 23 years.

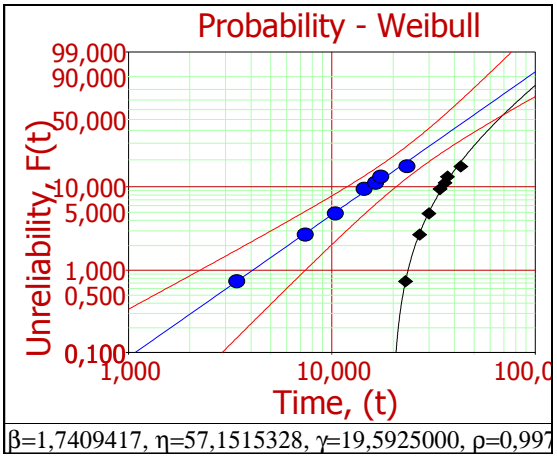

Figure 8: CDF of 3-parameter Weibull distribution with accompanying 90% confidence bounds.

In Figure 8, the cumulative density function of 3-parameter Weibull distribution is presented for the population from which the point suspected to be outlier has been removed. In Figure 9, time dependent failure rate function has been presented, for the same population.

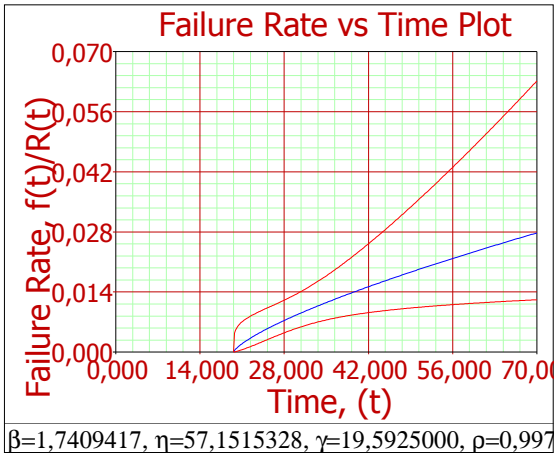

Figure 9: Failure rate versus time function, with accompanying 90% confidence bounds.

A remark has to be made regarding this method. In the presented case, there was only one point suspected, located at the beginning of the population. Thus, in (9) CL=1-α/2. If it turned out that there are two points suspected as being outliers, located in the beginning and in the end of population each, then the value of $r_{\alpha/2}$ corresponds to the CL=(1- α) instead of (1- α/2).

3.3 Summary

Based on real transformer failure data from the last 35 years, an example was presented of the analysis and investigation into the presence of outliers in the data-sets. Firstly the whole population was investigated and some points were found which can be called outliers with a given CL. For that purpose an approximate method has been presented for the outliers' identification. Secondly, a subsection of the transformers was investigated against the outliers and the Natrella-Dixon test was presented as an alternative method for the outliers' identification in the given data-set. In Table 3, an influence of retaining or rejection suspected points from the analysis is presented for the particular populations of transformers

Table 3: Values of B-life and mean-life for two populations, in each case influence of outliers on the analysis is presented.

	Total population [years]		Group I - transformers [Years]	
Outliers	**Included**	**Rejected**	**Included**	**Rejected**
B1 life	14 (11-17)	21 (20-23)	17 (13-23)	24 (22-27)
B10 life	36 (33-40)	36 (34-40)	35 (31-40)	35 (32-39)
Mean life	83 (69-99)	84 (71-100)	63 (53-77)	59 (51-68)

From Table 3, it follows that rejection of the outliers from the data-sets affected mainly the B1-life, as the removed outliers were in the beginning of the data-set and their removing resulted in changing the distribution. Values of B10 life and mean-life

were not significantly changed by the outliers' rejection.

4 INFLUENCE OF OUTLIERS ON THE FAILURE PREDICTION

Two real failure populations of transformers were investigated for the presence of possible outliers. For each population, a time dependent failure rate function was obtained. By using the function as well as information about the in-service population, it is possible to estimate the number of expected failures in the coming future. The results are presented in Figure 10 and Figure 11. For each population a prediction is made and the accompanying 90% confidence bounds are presented. Additionally, the influence of the rejection of outliers from the analysis on the failure prediction as well as on the confidence bounds is presented.

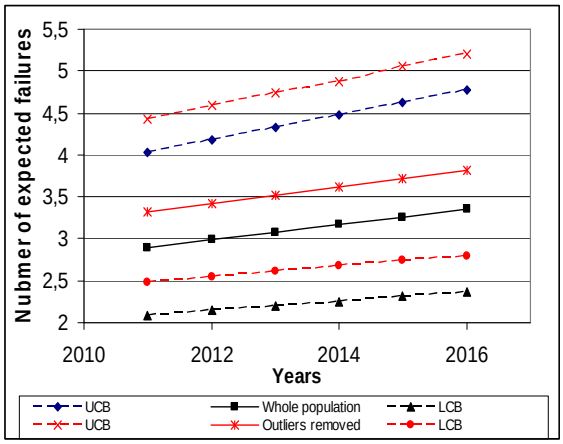

Figure 10: Failure expectation as calculated for the whole population, presented with 90% confidence bounds. Influence of outliers is pointed out.

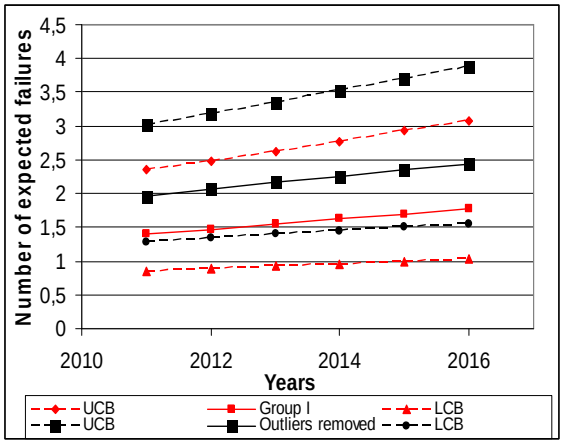

Figure 11: Failure expectation as calculated for group I, presented with 90% confidence bounds. Influence of outliers is pointed out.

From Figure 8 and Figure 9 it follows that outliers in the data-sets may have an influence on the number of expected failures in the future, as well as on the accompanying confidence bounds. In particular, for the group I of transformers, rejection of suspected outliers resulted in an increase of the number of expected failures in 2011 from around 1.4 to 2. In that case, also the 90% confidence bounds became wider.

5 CONCLUSIONS

From the statistical analysis of 110 kV and 150 kV transformer failure data obtained from a utility, the following is concluded:

- The failure expectation for the future can be estimated even with limited failure data. However, it might happen that due to the incomplete failure information, some points suspected to be outliers may be found.
- It is possible to prove with a certain confidence level that those suspected points are outliers.
- The presence of the suspected outliers may have a significant influence on the number of expected failures and accompanying confidence bounds, as well as on the particular values of B-lives.
- Considering differences in sizes of the whole population and group I of the transformers, it can be seen that the rejection of the outlier has stronger influence on the result of analysis for smaller population.

6 REFERENCES

[1] R.A. Jongen et al: "Application of Statistical Methods for Making Maintenance Decisions within Power Utilities" Electric Insulation Magazine, November/December 2006 – Vol. 22, No. 6

[2] N. P. Cheremenisinoff: "Practical statistics for engineers and scientists" Technomic 1987, ISBN 0-87762-505-0

[3] D. Kececioglu, "Reliability & Life Testing Handbook, Volume 1, PTR Prentice Hall Inc, New Jersey 1993. ISBN 0-13-772377-6