Analyzing Data of the Medium Voltage Grid and the Weather to Predict Power Outages

Internship Graduation Project at Alliander

Thesis MSc Mathematics

by

Mees van Osch
s4330714

Radboud University Nijmegen

Supervisors
Dr. Wieb Bosma (Radboud University) Dr. Sander Rieken (Alliander N.V.)

July 30, 2021

Abstract

Alliander maintains the medium voltage grid for a large part of the Netherlands. This network consists almost entirely of underground cables. Power outages are often caused by the breakdown of the cable joint connecting two cables in a circuits. The power outages can be prevented by replacing joints that are likely to fail. Smart Cable Guard (SCG) system monitors many circuits in the medium voltage grid. SCG measures, among other variables, partial discharges (PD) and faults in the circuits. Faults are short-circuit currents, which usually lead to circuit breakdown. The measurements consist of the timestamp, location and charge of every partial discharge that was registered by the SCG system.

The data of the faults, PD and weather are analyzed in this thesis. Most potential faults preceded by PD are prevented because of the warnings that SCG operators assign based on PD. Most of the faults that still occur take place in the summer. So there is a relation between the faults and the weather conditions.

The data from the PD is used to predict faults in the network to avoid power outages. Many faults are preceded by PD, so many faults can be predicted by analyzing the PD. Many of the PD measurements are noise, for example from nearby industry. Alliander uses a cluster algorithm to cluster the PD measurements from the same source to roughly filter the noise. Second, a classification model is used to determine which clusters of PD are likely to be followed by a fault, and which clusters are noise. This model uses various features of the clusters, such as duration, location, discharge magnitude and type of closest joint, to classify the clusters. Alliander currently uses 44 features. The model benefits from many features with predictive power.

In this thesis, we construct 33 new features to add to Alliander's total set of features. Many of these features are based on the relation between discharge magnitude and soil temperature during the discharge. The correlation between them appears to have a high predictive power. Also, the distribution of the PD across the seasons contributes to the prediction of the faults. In contrast to the faults, there is significantly more PD in winter than in summer. The distribution of the discharge magnitude is also very informative. The shape and scale of these distributions are two features extensively used by the classification model to predict faults.

The parameters of the classification model can be adjusted, but the first results when adding these 33 functions suggest better performance in predicting faults. This ultimately leads to the prevention of more power outages.

Preface

This thesis is the graduation project of the Master Mathematics at Radboud University Nijmegen in the specialization called Mathematical Foundations of Computer Science (MFoCS). I followed several courses on cryptology, machine learning and programming in this Master. I found out that I really like data analyzing and wanted to do a graduation assignment in this field. I was determined to combine my thesis with an external research internship. I wanted to experience how mathematics is used in business, work in a team and learn how to efficiently achieve concrete results in practice with mathematics.

Alliander offered me a great opportunity to analyze data with a clear goal in mind. Unfortunately due to corona I was unable to work in the office and I experienced the entire internship period from home. The many video meetings, on the other hand, have shown me broadly how things work at Alliander. They also definitely made me feel like I was involved with a team of nice and knowledgeable people. During this internship, I learned how to use Python in a professional way and work result-oriented.

A big thank you to Dr. Sander Rieken, my supervisor at Alliander. He was very involved with my thesis, my goals and my feelings. I felt like I could talk about everything with him. He helped to come up with clear research questions, ways to approach them and gave many tips for writing the thesis. He is very critical and communicates his criticism in a friendly manner. He also helped to understand Alliander's structure and culture. I was able to learn how he keeps many balls in the air at the same time with his drive and skills. His approach to team members is friendly and non-dominant, which makes him a great leader.

I would also like to thank Ruud Wassink. He helped me feel comfortable in the team through the light-hearted conversations we had. He also helped me understand the classification model I used to put my results in perspective.

Final thanks go to Dr. Wieb Bosma, my supervisor at Radboud Univeristy. He answered many questions about writing the thesis, came up with ideas about the content and gave feedback. Special attention is paid to his pleasant way of communicating about his feedback.

The target group of the thesis is a master's student in mathematics. Some prior knowledge of mathematics makes it much easier to read.

Mees van Osch

Nijmegen, July 27, 2021

Contents

1 Introduction

We start this thesis by introducing Alliander, Smart Cable Guard and the purpose of the research. Alliander is the company where I did my internship, and the Smart Cable Guard systems provide the data that we will use to answer the research questions. An overview is given at the end in which we briefly explain what is discussed in each chapter.

Alliander

Alliander N.V. is a utility company that develops and manages energy networks. It takes care of the distribution of energy in a third of the Netherlands. More than three million Dutch households and companies receive electricity, gas and heat via its cables and pipelines. Alliander consists of the parts Liander, Qirion and Kenter. During my internship I was part of Qirion. Qirion focuses on the construction and maintenance of complex energy infrastructures for customers.

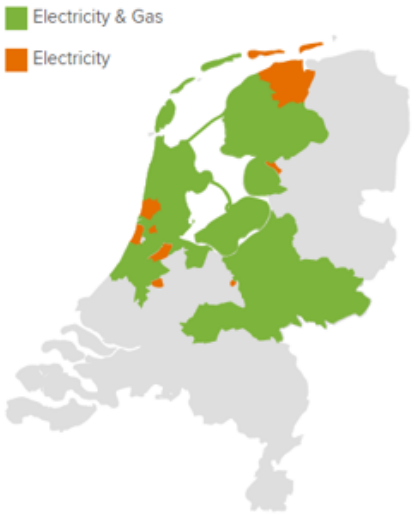


Figure 1: Service area of Alliander in the Netherlands [1]

Smart Cable Guard

DNV (Det Norske Veritas which translates to "The Norwegian Truth") provides digital solutions for managing risk and improving safety and asset performance for ships, pipelines, processing plants, offshore structures, electric grids, smart cities and more [2]. It provides Smart Cable Guard (SCG) to Alliander to monitor its electricity grid. SCG is a sensor-based digital monitoring platform that puts owners in control of their medium voltage cable network. Combining patented technology with 24/7 monitoring and support, it detects and locates faults and weaknesses in underground cables [3]. Alliander installs the SCG-systems to monitor their medium voltage network. The collected data about the network is used to predict power outages.

Research Questions

Cables in the grid are connected by joints. These weak spots of the medium voltage network are usually the cause of power outages. These defects can be prevented by replacing the right joints in time. Alliander used the data of SCG to predict the power outages.

SCG registers faults and partial discharges (PD) on the joint and the cables. Power outages are preceded by faults and faults are often preceded by partial discharges. These data are used to predict the outages, by first clustering the partial discharges from the same source. Subsequently, features of these clusters are calculated and fed into a prediction model. Alliander has implemented models for the clustering, calculation of features and classification of clusters. The prediction model uses the cluster features to classify the clusters, and that leads to the prediction of faults.

The goal is to improve these models, so that the prediction model is better able to predict which PD clusters will lead to a fault.

The model classifies clusters that are likely to be followed by a fault, as dangerous. Alliander acts on this by replacing the relevant joints in the network. Strictly speaking, this model does not predict faults. Using the data of the PD the model predicts when and where a fault would occur if no maintenance work was carried out on the network. That said, for the sake of convenience, we still talk about the prediction of faults.

Faults without a warning are often caused by external influences, like damage from excavation activities. These faults are unpredictable from PD data. However, there are still predictable faults, showing PD activity beforehand, that occur without warning. Experience has shown that these predictable faults depend on the weather conditions, so combining the fault data with the weather data and PD can lead to improved prediction of these faults.

Research question 1.
Can we predict faults from the relation between faults, PD and weather conditions?

We are going to use Alliander's classification model to predict the faults. So we need to define features based on weather data that have predictive power to feed to the model.

Research question 2.
How can we compose features of these relations so that they improve the current model performance to predict faults?

First we will investigate the relation between weather conditions, faults and PD. We will answer the next questions:

- How many faults occur each year, and per circuit?
- In how many cases is a cluster of PD found beforehand?
- Is this related to the seasons?

Next, we will compose features that allow the model to better predict the faults. Finally, we use the model to test the contribution of these new features to the prediction of faults.

Many of the used terms are explained in Chapter 2 and the List of terms. You can click on a term to go directly to its definition. All plots in this thesis are created using the libraries matplotlib.pyplot [4] and plotly.offline [5].

Chapter overview

We begin in **Chapter 2** by explaining in detail what a circuit is and how the SCG-systems are used to collect the data about the faults and PD from the circuits.

In **Chapter 3** we elaborate on three mathematical functions that are used throughout the thesis: Pearson's correlation coefficient, percentiles and the Weibull distribution.

In **Chapter 4** we explore the fault data and the PD in detail. The effect of the weather conditions on faults and PD is also investigated.

In **Chapter 5** we examine the data about the temperature further to find out to what extend it can be used to predict faults.

In **Chapter 6** we define three methods to composes 25 cluster features that describe the relation between PD and temperature. Two methods rely on the correlation between PD and temperature, for long and short periods of a cluster. The third method relies on the distribution of the PD across time to determine certain features.

In **Chapter 7** we describe two methods to composes eight cluster features related to the charge of PD. We look at PD with a high charge and we try to fit a Weibull distribution to the distribution of the charges of the partial discharges.

In **Chapter 8** we briefly explain the prediction model used by Alliander. The 35 features of Chapters 6 and 7 are evaluated by this model. The prediction of the faults using the current features is compared with the prediction of the faults using the current and the new features combined.

We discuss and conclude this thesis in **Chapters 9** and **10**.

These chapters are followed by the **List of features** and **List of terms** which give an overview of all the features and the important terms, with their definitions.

2 Technical background information and explanation of terms

The power grid in the Netherlands can be divided into three parts:

- The High Voltage (HV) network transports the power of > 36 kV over large distances;
- Medium Voltage (MV) network transports the power of 10-20 kV to regions;
- Low Valtage (LV) network transports the power of 400 V to customers in the neighborhood [6].

Figure 2: Secondary substation at De Randweg in Arnhem [Van Osch, 2021]

A station that transforms the power from high voltage to medium voltage is called a substation. A **secondary substation** transforms the power from medium voltage to low voltage. Alliander's medium-voltage network is designed as rings but operates radially: for each secondary substation there is precisely one way to get power from the substation [7]. Each secondary substation is equipped with a medium-voltage ring main unit (RMU). A ring main unit (RMU) is a set of switchgear used at the (secondary) substations of a ring distribution network. The RMU are nodes in this ring, which explains the term ring main unit. An RMU in a ring structure can be powered from either side, so if a cable fails on one side, the switchgear of the nearby RMU can be used to restore power through the other side. This allows the power supply to customers to be restored quickly, as the cable does not need to be repaired first. The ring structure also allows maintenance work to be carried out without customers running out of power. [6]. The SCG-systems are placed in the RMUs.

2.1 SCG on circuits

Alliander started to monitor weak cables by placing SCG-systems in nearby RMUs. A SCG-system consists of two SCG-devices: a master unit that sends pulses and a slave unit that receives the pulses. Figure 3 shows three RMUs and a SCG-system. The master unit is installed in RMU A and the slave unit is installed in RMU B. A SCG unit consists of a Controller Unit (CU) and a Sensor/Injector Unit (SIU). The CU controls the data collection and provides the data communication. The SIU of the master unit injects a pulse to the SIU of the slave unit every minute, to detect faults and partial discharges on the cables and joints. When there is a defect between the two SCG-devices, like at X in Figure 3, both RMUs receive pulses from which the location of the defect can be determined.

Definition 1. A **fault** is a short-circuit current. It usually leads to circuit breakdown.

Definition 2. Partial discharges (PD) are small charge displacements in the cavity or layer of the insulation of a component. PD is a good predictor of faults.

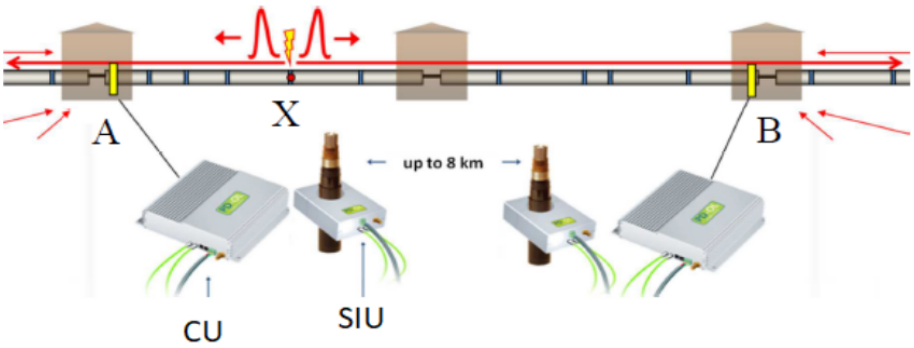


Figure 3: SCG setup to monitor a circuit [7]

A circuit is defined by the placements of its SCG-system:

Definition 3. A **circuit** consists of all cables, joints and RMU's between the master and slave unit that of a SCG-system.

Definition 4. The **circuitlength** is the cumulative length of the cables of the circuit: the distance between the master and slave unit.

Circuits have different lengths up to 15 km. Figure 9(a) shows the distribution length of the circuit lengths of the SCG monitored circuits in the Alliander network. Figure 4 shows the parts of a particular circuit. This circuit contains five RMUs from which only the start and end has got a SCG-device. The cables are made of different materials and the length of the cables is shown in the second column of Figure 4. The cables are connected by either an RMU or a joint. The function of a joint is just to connect the cables. These joints are usually the weak spots of the circuits, depending on the material of the joints. The cumulative length tells the location of the RMUs and joints. The location of the master unit on the circuit is 0 m and the location of the slave unit is the length of the circuit.

Component type	Length (m)	Cumulative length (m)
RMU		
Termination (unknown)		
Cable (XLPE, 3 cores, common earth screen)	70	70
Joint (heat shrink)		
Cable (PILC, 3 cores, belted)	35	105
Termination (unknown)		
RMU		
Termination (unknown)		
Cable (PILC, 3 cores, belted)	49	154
Joint (cold shrink)		
Cable (XLPE, 3 cores, common earth screen)	146	300
Termination (unknown)		
RMU		
Termination (unknown)		
Cable (XLPE, 3 cores, common earth screen)	21	321
Joint (cold shrink)		
Cable (PILC, 3 cores, belted)	72	393
Joint (unknown)		
Cable (PILC, 3 cores, belted)	101	494
Joint (unknown)		
Cable (PILC, 3 cores, belted)	15	509
Joint (unknown)		
Cable (PILC, 3 cores, belted)	98	607
Termination (unknown)		
RMU		
Termination (unknown)		
Cable (PILC, 3 cores, belted)	4	611
Joint (cold shrink)		
Cable (PILC, 3 cores, belted)	106	717
Joint (heat shrink)		
Cable (XLPE, 3 cores, common earth screen)	19	736
Termination (unknown)		
RMU		

Figure 4: The cable configuration of circuit 3107

2.2 Alliander's use of SCG

Alliander's entire network in the Netherlands consists of 91,000 km of cables for electricity and 42,000 km of pipelines for gas. 3000 km of cables is monitored by SCG. In June 2021, Alliander is using 1837 SCG-systems to monitor these cables on 1837 circuits, and the number of SCG-systems that Alliander uses continues to increase. There are 19.000 joints on these circuits, so they are measured simultaneously. In addition to the data from the faults and PD, Alliander receives warnings from DNV. These warnings tell which locations of the circuits need to be monitored closely. Alliander uses these warnings combined with the data of faults and PD to decide which joints need to be replaced, to prevent circuit outages. Alliander works on a model to predict circuit outages without making use of the warnings. This ultimately allows Alliander to better predict faults in the joints that Alliander uses.

3 Mathematical background information

3.1 Pearson's correlation coefficient

Correlation is a statistical relationship between two variables. Correlation coefficient is the measure of the correlation that exists between two variables. The coefficient is a real number between -1 and . A correlation coefficient of 1 indicates a perfect positive correlation. As variable X increases, variable Y increases and as variable X decreases, variable Y decreases. If the correlation coefficient is greater than 0, it is a positive relationship. Conversely, a correlation coefficient of -1 indicates a perfect negative correlation. If the coefficient is less than 0, it is a negative relationship. A value of 0 indicates that there is no relationship between the two variables.

One has to note that correlation is not causation. Correlation does not necessarily mean that one variable causes the other. For example, palm size is negatively correlated with longevity. This does not mean that the size of your palm causes you to live or die. In fact, women tend to have smaller palms and live longer. To find causation, you generally need experimental data, not observational data. In this thesis we only use observational data.

The most common measures of correlation is the Pearson correlation coefficient, developed by Karl Pearson in 1895 [8]. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between -1 and 1.

Definition 5. Pearson's correlation coefficient of X and Y: $\text{Pearson}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$.

Here $\text{cov}(X,Y)$ is the covariance between X and $Y : \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$;
σ_X is the standard deviation of X: the square root of the variance of $X : \sqrt{\mathbb{E}[(X-\mu_X)^2]}$.

This thesis only uses the Pearson correlation coefficient to express a correlation between two variables. When we mention a correlation coefficient, we always mean the Pearson correlation coefficient.

3.1.1 Significance

The correlation coefficient measures the strength of a relationship in samples only. It doesn't say whether what we see in the sample is expected to be true for more data. To test whether we have enough data points to conclude whether there is a correlation between two datasets, the significance test is used.

The correlation coefficient of two variables is indicated with ρ. We test whether ρ is close to 0 or significantly different from 0. We decide this based on the sample correlation coefficient r and the sample size n. We specify the null hypothesis H_0 and the alternative hypothesis H_a:

$$H_0 : \rho = 0;$$

$$H_a : \rho \neq 0.$$

If we fail to reject the null hypothesis that $\rho = 0$, we say that the two variables are not significantly correlated. Then the correlation occurred on account of chance coincidence in the sample. We use the Student's t-test to find out if we can reject the null hypothesis. The t-test is one of the most commonly used techniques for testing a hypothesis on the basis of sample data. The value of the t-test is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

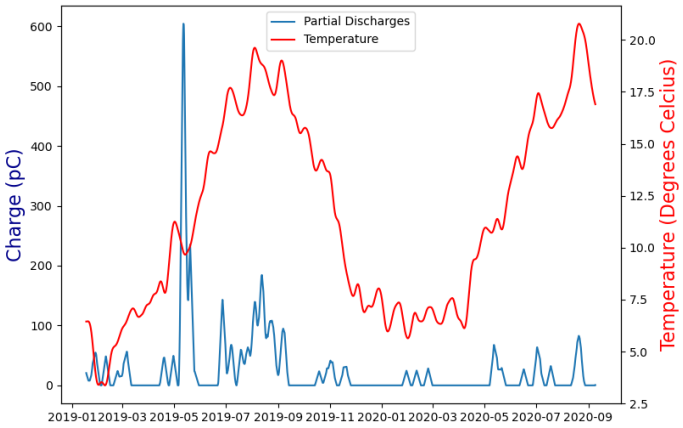
Here r is the sample correlation coefficient and n is the sample size.

The bigger the t-value, the more likely it is that the correlation is repeatable. To interpret the t-value we need to find the p-value. A p-value is the probability that the null hypothesis is true. Like in most research, we consider a p-value ≤ 0.05 significant. A p-value of 0.05 means that there is only 5% chance that results from the sample occurred due to chance. The p-value can be looked up in a t-table using the t-value and the the number of degrees of freedom ($df = n-2$). A t-table can be found in Table 13 in the Appendix.

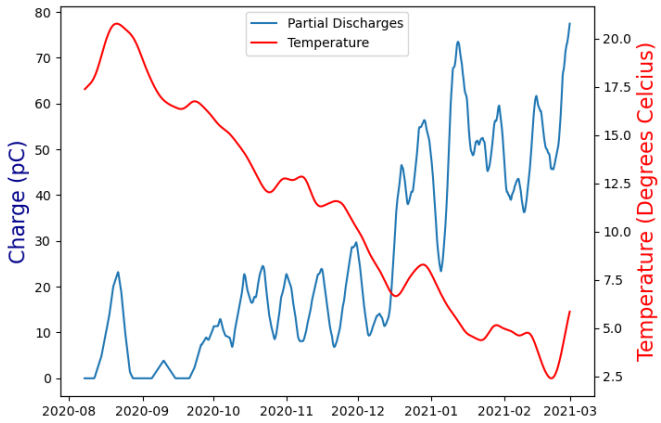
3.1.2 Interpreting the value of the correlation coefficient

A correlation coefficient of 1 is a perfect correlation, 0 indicates no correlation between two variables. Determining whether a coefficient between 0 and 1 represents a strong correlation is subjectively determined.

The correlation coefficient is used in this thesis to quantify a pattern in the data. These quantifications are used to compare different datasets. In Figure 5 we see two plots of partial discharges and temperature. We do not see a clear pattern between the PD and temperature in Figure 5(a). This is why the correlation coefficient is relative low: 0.18. There is a clear pattern in Figure 5(b). The temperature decreases while the charge of the PD rises. We see a relative high (absolute) correlation coefficient of -0.84. The opposite trend explains the negative coefficient. These correlation coefficients are used to predict power outages. Stronger correlations have greater predictive value.



(a) Partial discharges of cluster 8 of circuit 2719 and the soil temperature. The correlation coefficient is 0.18.



(b) Partial discharges of cluster 1 of circuit 2719 and the soil temperature. The correlation coefficient is -0.84.

Figure 5: Data with weak correlation and strong negative correlation

3.2 Percentile

Percentile is a specific form of the more general concept of quantile. Although the term quantile appeared first in 1940 [9], Fracnis Galton used the term equi-postile to describe the idea of quantile in 1902 [10]. **Quantile** comes from the Latin word quantus (how much or how great), and is defined in the Oxford English Dictionary (OED) as "each of any set of values of a variate which divide a frequency distribution into equal groups, each containing the same fraction of the total population; also, any one of the groups so produced, e.g. a quartile, decile, or percentile."

Already back in 1885 [11] Galton used the term **percentile** and is defined by the OED as "each of a series of values obtained by dividing a large number of quantities into a hundred equal groups in order of magnitude; that value which is not exceeded by the lowest group is the first percentile; that not exceeded by the lowest two, the second percentile; and so on."

This definition is not sufficient for this thesis because not all datasets can be divided into a hundred equal groups. In this thesis we use the *pandas* function *quantile()* [12], that calculates the n-th percentile for each $n \in [0, 100]$. This function is well-defined and can be described by the next method:

If *data* is the increasing sequence of the data points $[data[0], ..., data[N-1]]$ with length N, then the n-th percentile P_n ($n \in [0, 100]$) can be found as follows.

Define a list A of length N of easy computable percentiles:

$$\forall i \in \{0, ..., N-1\}: \quad A[i] = \frac{100i}{N-1}.$$

Now we have a list $A = [0, \frac{100}{N-1}, \frac{200}{N-1}, ..., , \frac{100(N-2)}{N-1}, 100]$ of the same length as *data*. Next we determine the percentiles of A that surround n and call their indices m and M:

$$m \text{ is the index for which } A[m] = max\{a \in A | a \leq n\};$$
$$M \text{ is the index for which } A[M] = min\{a \in A | a \geq n\}.$$

$m, M \in \{0, 1, ..., N-1\}$. Note that $M = m+1$ if $n \notin A$, and $M = m$ otherwise. $A[m] \leq n \leq A[M]$ by construction. Finally we define the n-th percentile P_n:

$$P_n = data[m] + (data[M] - data[m]) * \frac{n - A[m]}{A[M] - A[m]}.$$

Note that $P_n = data[m]$ if $n \in A$.

The next example finds the 83th percentile P_{83} for $data = [3, 5, 8, 9, 13, 20]$:
We have $n = 83$ and $N = 6$.

$$A[0] = \frac{100i}{N-1} = \frac{0}{5} = 0; \qquad A[1] = \frac{100i}{N-1} = \frac{100}{5} = 20; \quad A[2] = \frac{100i}{N-1} = \frac{200}{5} = 40;$$
$$A[3] = \frac{100i}{N-1} = \frac{300}{5} = 60; \quad A[4] = \frac{100i}{N-1} = \frac{400}{5} = 80; \quad A[5] = \frac{100i}{N-1} = \frac{500}{5} = 100.$$

$80 \leq 83 \leq 100$, so $m = 4$ and $M = 5$, and we get:

$$\begin{aligned} P_{83} &= data[4] + (data[5] - data[4]) * \frac{83 - A[4]}{A[5] - A[4]} \\ &= 13 + (20 - 13) * \frac{83 - 80}{100 - 80} \\ &= 13 + 7 * \frac{3}{20} = 14.05. \end{aligned}$$

3.3 Weibull distribution

The Weibull dsitribution is a a continuous probability distribution, named after Swedish mathematician Waloddi Weibull. He described it in detail in 1951 [13]. The Weibull distribution is widely used in reliability and life data analysis due to its versatility and relative simplicity. Depending on the values of the parameters, the Weibull distribution can be used to model a variety of life behaviors. [14]

The Weibull distribution is defined by the two parameters α and β:

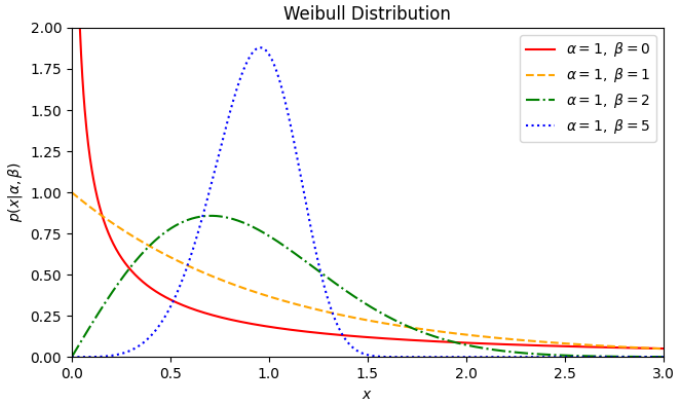
$$W(x;\alpha,\beta) = \begin{cases} \frac{\beta}{\alpha}(\frac{x}{\alpha})^{\beta-1}e^{-(x/\alpha)^{\beta}} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

$\alpha > 0$ represents the scale of the distribution and $\beta > 0$ represents the shape.

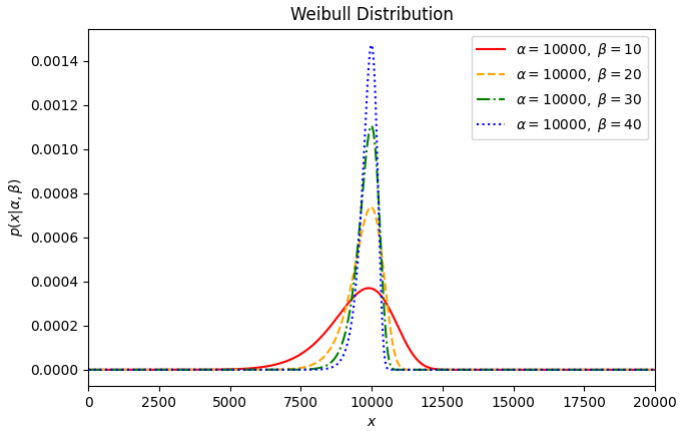
The shape parameter, β, is also known as the Weibull slope. This is because the value of β is equal to the slope of the line in a probability plot. Different values of the shape parameter can have marked effects on the behavior of the distribution. Figure 6(a) shows the effect of different values of β on the shape of the probability density function (pdf). One can see that the shape of the pdf can take on a variety of forms based on the value of β. The skewness depends only on the shape parameter.

For $0 < \beta < 1$, the pdf tends to infinity as x approaches 0 from above and is strictly decreasing. For $\beta = 1$, the pdf tends to $\frac{1}{\alpha}$ as x approaches 0 from above and is also strictly decreasing. For $\beta > 1$, the density function tends to 0 as x approaches 0 from above, increases until its mode and decreases after it. Also the slope of the pdf at $x = 0$ is determined by the shape parameter. The slope is negative if $0 < \beta \leq 1$, positive if $1 < \beta \leq 2$ and it is a null slope at $x = 0$ if $\beta > 2$. In Chapter 7, the Weibull distribution is used to simulate distributions of partial discharge. The shape parameter, β, of these distributions is always bigger than 1.

Figure 6(b) shows that the Weibull distribution converges to a Dirac delta distribution centered at $x = \alpha$, as β goes to infinity. The Direac delta is a hypothetical signal that lasts infinitely short and at the same time is infinitely high, such that the integral is exactly equal to 1.



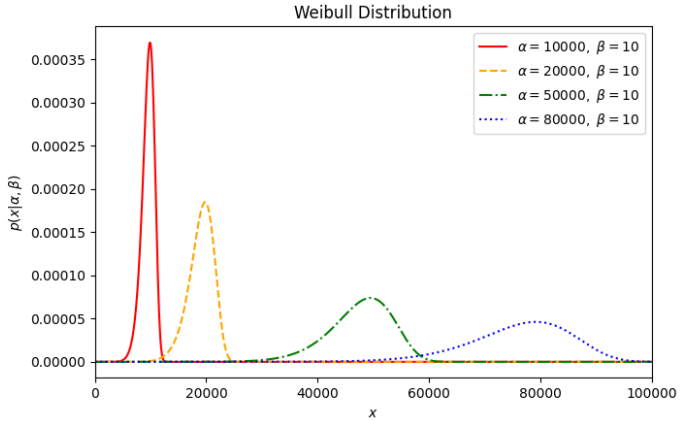
(a) For constant $\alpha = 1$, the shape of the distribution changes for different β



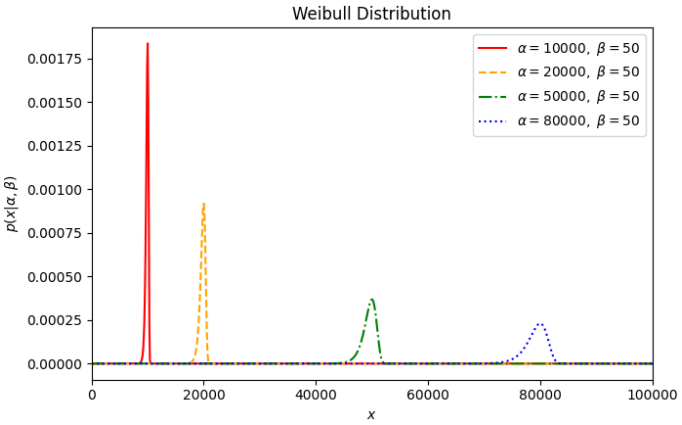
(b) Centered at $\alpha = 10000$ and converges to a Dirac delta as β goes to infinity

Figure 6: Probability density functions for several scale and shape parameters α and β

Figure 7 shows that the pdf is stretched out if the scale parameter, α, is increased. Since the area under a pdf curve is a constant value of one, the peak of the pdf curve will also decrease with the increase of α. For large β the mode is approximately equal to α.



(a) For constant $\beta = 10$, the distribution stretches out if α increases



(b) For constant $\beta = 50$, the distribution stretches out if α increases

Figure 7: Probability density functions for several scale and shape parameters α and β

The Weibull distribution will be used in Chapter 7 to quantify features of clusters. Figure 8 shows the distribution of the partial discharges of a cluster and the Weibull distribution that fits best.

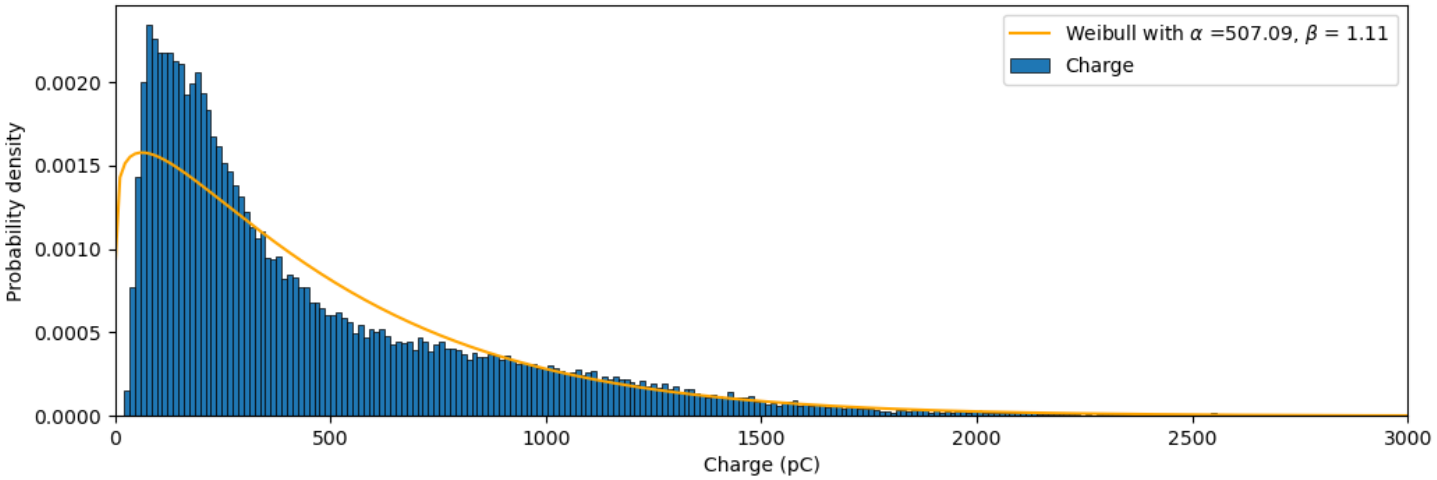


Figure 8: Distribution of the particles of cluster 1 of circuit 4082. The orange line shows the Weibull distribution with parameters $\alpha = 707.09$ and $\beta = 1.11$ that fits best.

4 Data exploration

In this chapter we look at the data we are going to use to gain more insight into it. Alliander uses SCG to retrieve the data of the faults and partial discharges (PD), and we have access to a weather application programming interface (API) [15]. Using this weather API we can collect all kinds of weather variables. In 2014, changes were made to the hardware of the SCG-systems. SCG produced other numbers from then on. Therefore, in this thesis we focus on the data from December 2014. In Section 3.1 and Section 3.3 we take a close look at the data of the faults and partial discharges respectively. In Section 3.2 we will explore which weather variables are the most valuable for our research to predict both faults and PD. Section 3.4 shows the available data of the warnings and we conclude with the findings of the data exploration in Section 3.5.

4.1 Faults

When there are short-circuit currents in a circuit, SCG registers and classifies them as faults. The faults usually lead to circuit breakdown. See Chapter 2 for more details on the measurements. The dataset of the faults available to us looks like Table 1. This table only shows a small fraction of the entire dataset. We have access to five variables of all 822 faults registered between December 2014 and February 2021.

Date/time (UTC)	Location in meters (m)	FaultCount	FaultGroup	circuitnr
2015-08-30 20:24:00	9695.72822	1	150	1108
2015-11-14 15:17:00	9695.72822	1	150	1108
2017-05-05 14:39:00	3539.33803	1	743	1108
2014-12-04 14:37:00	1234.19103	1	44	1219
2014-12-04 15:01:00	2486.98796	2	46	1219
2014-12-06 08:13:00	6964.80669	2	47	1219
2015-02-26 05:34:00	2619.66522	125	71	1225
2015-12-22 11:57:00	10384.12186	1	201	1235
2015-12-22 11:57:00	11332.26557	7	202	1235
2015-12-22 11:57:00	11760.87848	1	203	1235
2017-05-19 16:18:00	11760.87848	1	203	1235
2017-10-29 12:56:00	344.18915	4	1031	1235

Table 1: Dataset of the faults. On circuit 1225 occurred a fault with a huge faultcount in 2015.

- Every circuit with an SCG-system, has a unique circuit number which is written as **circuitnr**. Circuits with the same circuit number are in fact the same circuits.
- **Date/time (UTC)** is the time the fault is registered. SCG measures only every minute so we do not have the exact time of the fault. As we can already see there are circuits with multiple faults in different locations. See for example circuit 1219 in Table1.
- **Location in meters (m)** tells the distance to the starting point of the circuit. SCG makes a calculation to determine the location of the faults, and presents it with five decimals. However the precision of the localization is approximately 1% of the circuitlength, according to SCG.
- **Faultcount** is the number of large sparks that occur during the minute of the fault. One fault can consist of multiple detections. If a short circuit occurs in a cable, multiple blows (sparks) can occur. SCG combines these per minute. So if there are multiple detections per minute, a fault gets a faultcount of more than 1.

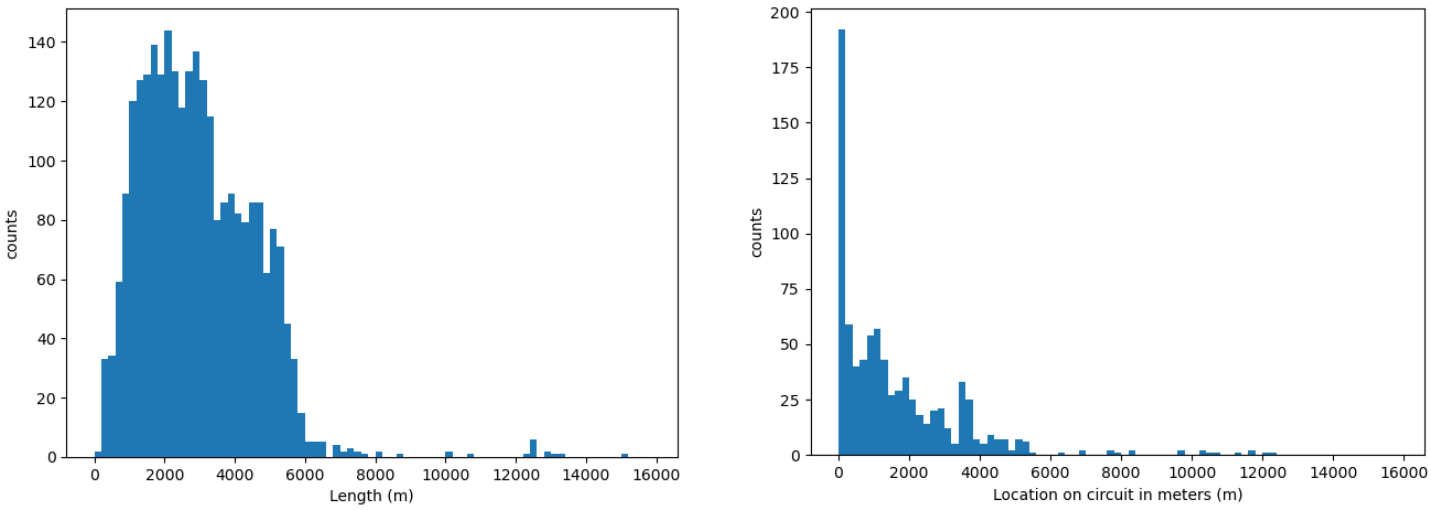
- Some faults belong to the same **faultgroup.** Fault grouping is a way of combining multiple faults, based on location and time. DNV gives faults the same faultgroup if they have the same circuit number and location is within and including ±2% of the circuitlength and their time is within 180 days.

  For example, when a fault occurred on one specific circuit at 2020-10-01 00:00:00, at location 30% and then another fault, on that circuit, occurred on the 2020-10-05 00:00:00, at location 32%, then those faults will have a same fault group id.

  The precision of the localization of the faults is 1% of the circuitlength, so two faults that are 2% of the circuitlength apart may have occurred at the same location. This explains DNV's choice of 2%.

4.1.1 Location

We take a closer look at the data of the population of all Smart Cable Guard circuits. Before we look at the locations of the faults it is good to know that not all circuits have the same length. In Figure 9(a) we see the distribution of the lengths of the circuits.



(a) Circuitlengths

(b) Locations of the faults

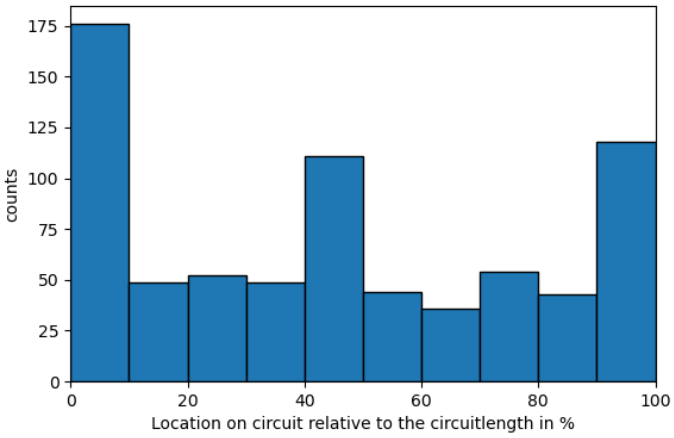
Figure 9: Distribution of the lengths of the circuits and the location of the faults on the circuits

The distribution of the absolute locations of the faults in Figure 9(b) shows a big outlier: there are relatively many faults at 0 meter. This raises the question if there would also be such an outlier at the end of the circuits. Because all circuits have different lengths, we look at the relative location of the faults. In Figure 10(a) we see the faults are divided into 10 bins. Each bin represents 10% of the length of a circuit. A fault is allocated to a bin if the location of the fault divided by the length of the particular circuit lies in that bin:

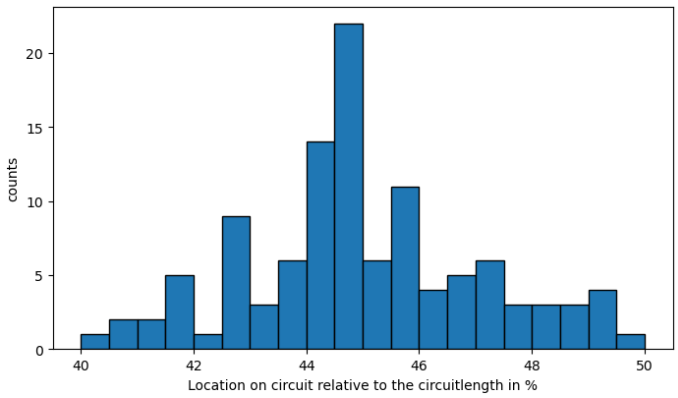
$$\frac{location}{circuitlength} * 100\%.$$

There happen to be 8 faults for which $\frac{location}{circuitlength}$ is larger than 1. This is caused by errors in the data. The circuitlengths of those circuits have changed in the past and are not up to date. This causes errors in the calculation of the relative locations because the locations of the faults, on the other hand, are up to date. Previous investigation showed that this is very incidental, so only for a few circuits the circuitlength is not up to date and we can assume that the locations are correct.

If the location does not have any influence on the faults, we expect a uniform distribution in Figure 10(a). However we see three clear outliers: 0-10%, 40-50% and 90-100%. A closer look tells us that a lot of faults are exactly at 0% and 100%. This has to do with the ring main units (RMU) at the start and end of the circuits. Some faults that do not occur on a monitored circuit are still registered by a nearby circuit. These faults do not occur between the two SCG-systems of a circuit but they are registered by them. SCG determines incorrectly that the location of these faults is the location of the nearest RMU. This is why we see many faults at 0% and



(a) All faults



(b) Faults for which the relative location is between 40-50%

Figure 10: Distribution of the relative location of the faults on the circuits. The relative location is the absolute location on the circuit divided by the circuitlength.

100%. It also happens sometimes that the maintenance work at the RMU's cause the SCG-system to register a fault. Both reasons explain the high bins of Figure 10(a).

For the outlier between 40% and 50% we take a closer look at the data. Circuit 3532 has 15 faults in the exact same location, so they also have the same relative location: 44.84%. It concerns an exceptional case. Alliander has reviewed this situation carefully and decided to divert this circuit and connect another circuit to the relevant customers. This way circuit 3532 is still monitored, but a possible circuit breakdown will not effect any customers. Alliander wants to see how long it takes for the cable to actually break. This is why there were faults again and again at this location without Alliander intervening.

These 15 faults still do not explain the huge 40-50%-bin. It turns out that on this specific location there are two parallel circuits. Because of this exceptional case the faults on one circuit can also be detected by the other circuit. The circuits at this location appear to have the same length so the relative location is the same and consequently we see a huge outlier in Figure 10(b).

So not all faults result in a breakdown and circuit outage. A fault can be registered in one circuit although it occurred in a nearby circuit. In these cases the faults are registered on both circuits: One inside a circuit with a clear location, and the other outside the circuit that leads to a location of 0% or 100% of the circuitlength. There are also events which trigger SCG to register a fault while there is no danger to the circuit at all. Lightning and human activities during the maintenance of the RMU's are examples of such events.

We also have to note that there occur **intermittent faults**: "faults with very short durations (a few milliseconds), after which they disappear. Such self-healing faults typically happen in fluid-filled oil and mass-filled joints. Sometimes this can happen many months before a full breakdown occurs, on which the protection equipment can operate" [7].

The provider of SCG, DNV, has developed a decision tree to classify faults into true positives (actual faults) and false positives (wrong observations). Most of the the false positives are filtered by means of this decision tree. We further assume that all remaining faults are relevant for our analysis.

4.1.2 Faultgroup

We investigate the column *FaultGroup* a bit further. The 822 faults are divided into 753 faultgroups. Figure 11 shows two distributions of the faultgroups. In Plot 1 of Figure 11 we see that most faultgroups consist of only 1 fault. Note the logarithmic y-axis. We also see that there are no very big faultgroups. The biggest faultgroup consists of only 8 faults which occurred roughly on the same time and location.

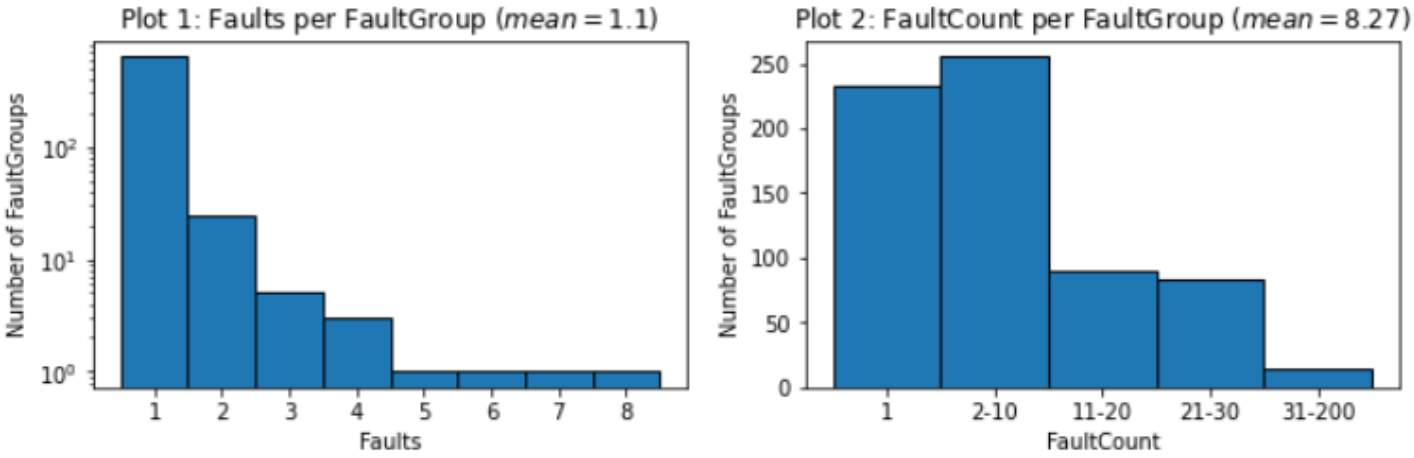


Figure 11: Distribution of number of faults and faultcount per faultgroup

For Plot 2 of Figure 11 we added up the faultcount of the faults which belong to the same faultgroup. For each bin in the graph we see how many faultgroups there are for which the faultcount of the faults adds up to the relevant bin. For example there are 231 faultgroups which consist of faults with a total faultcount of 1. The faultcount of a fault is an integer ≥ 1, so these faultgroups must consist of precisely 1 fault. This is why we see those large bins at the left of both plots of Figure 11.

We see that there are not many faultgroups with a large faultcount but there are a few faultgroups with a faultcount larger than 30. Those faultgroups do not consist of many faults as we can see in Plot 1, so the faults of these faultgroups have a large faultcount.

4.1.3 Circuit numbers

We investigate the column *circuitnr* a bit more. The 822 faults occurred in 527 circuits. So 2313 of the 2840 circuits in February 2020 have not experienced a fault. This is why we make a distinction between circuits without faults and circuits with at least 1 fault. The three left plots (plot 1, 3, 5) of Figure 12 show data of the circuits with fault(s). In plots 1 and 3 we see that most of the circuits have just 1 fault and a faultcount of 1: their shapes are very similar to plots 1 and 2 of Figure 11 about the faultgroups. In plots 2 and 4 of Figure 12 the circuits without any faults have been added. We see that most of the circuits do not have any faults and therefor their faultcount is 0. There is one circuit that has 15 faults. This corresponds to circuit 3532 that was previously mentioned in Section 3.1.1. Unlike Plot 1, this can be seen in Plot 2 due to the logarithmic scale.

In Plot 5 we see the unique faultgroups per circuit. Some circuits have ≥ 2 faults. If these faults are not in the same faultgroup, the circuit has ≥ 2 different faultgroups. We call them unique faultgroups. As before, we see that in most cases there is one unique faultgroup.. This is due to the fact that most circuits have just 1 fault. But we also see that there is 1 circuit which has 8 unique faultgroups.

Plot 6 shows the faultcount per fault. Most of the faults have a faultcount of 1. Plot 3 has a very similar shape as Plot 6, because most circuits with a positive number of faults only have 1 fault (which can be seen in Plot 1).

All plots of Figure 11 and 12 show a logarithmic decline. This is especially clear in Plot 2 and 5 of Figure 11 and Plot 1 of Figure 12 because of the logarithmic scale. Most circuits have no faults and the circuits with a fault, usually only have one faultgroup with only one fault with a faultcount of one.

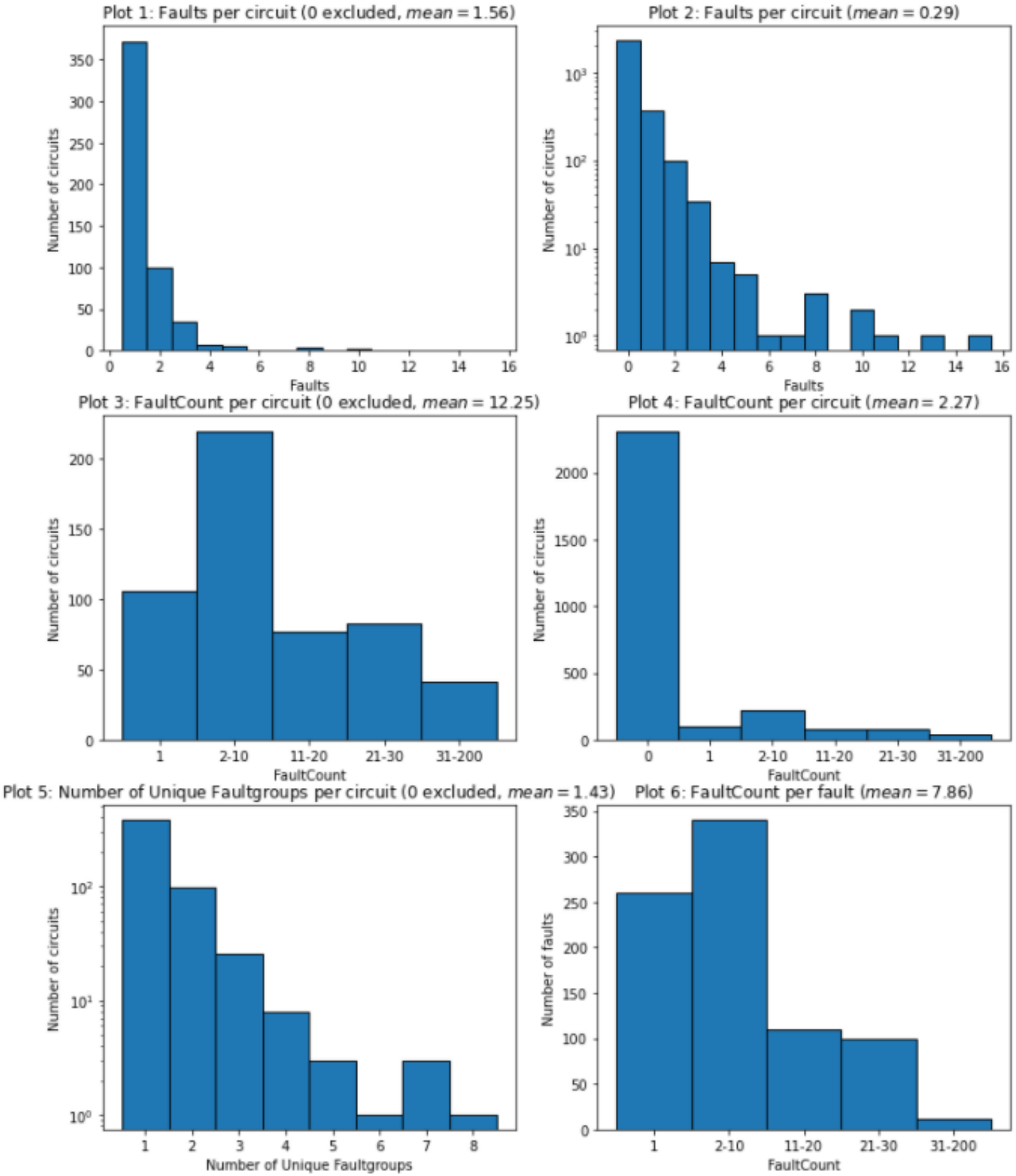


Figure 12: Plots about circuit number

4.2 Weather

In this section we collect the data using the weather application programming interface (weather API) and relate it to the data of the faults. We show that the number of faults correlate positively with the temperature.

4.2.1 Application programming interface

Using the Alliander weather API, we can access 57 different weather variables, including temperature at different elevations, wind speed, and air pressure. The API uses two sources for this data. The Dutch meteorological institute called *Koninklijk Nederlands meteorologisch instituut* (KNMI) provide 39 features for each day [16]. 22 of them are also available for each hour [17]. On top of that *The Climate Data Store* (CDS) provides [18] the other 19 features including for example the soil temperatures. Also the data from CDS is registered for each hour. All these 57 variables are shown in Table 14.

The weather variables are measured at 670 weather stations across the Netherlands [19]. Since the weather at the circuits does not vary significantly, we use (unless stated otherwise) the data measured in De Bilt, the central gauge of the weather in the Netherlands (52° 06' N.B. 05° 11'O.L) [20]. See Chapter 5 for more details about the variation of the temperature across the circuits. At the time of application, only the data of the past 3 years was available in CDS. However in Section 5.4 we need soil temperatures from many years back. This is why we use an alternative source of KNMI [21] for this. Unfortunately, these soil temperatures from KNMI are registered only for each 6 hours. Because the soil temperatures from CDS are registered for each hour, we will always use these except in Section 5.4.

4.2.2 Day

In the next figures (13, 14, 16) we just use the general air temperature: the temperature measured at a height of 1.5 meters [22]. In Figure 13 we see the temperature and all the faults distributed across the hours of a day. Each

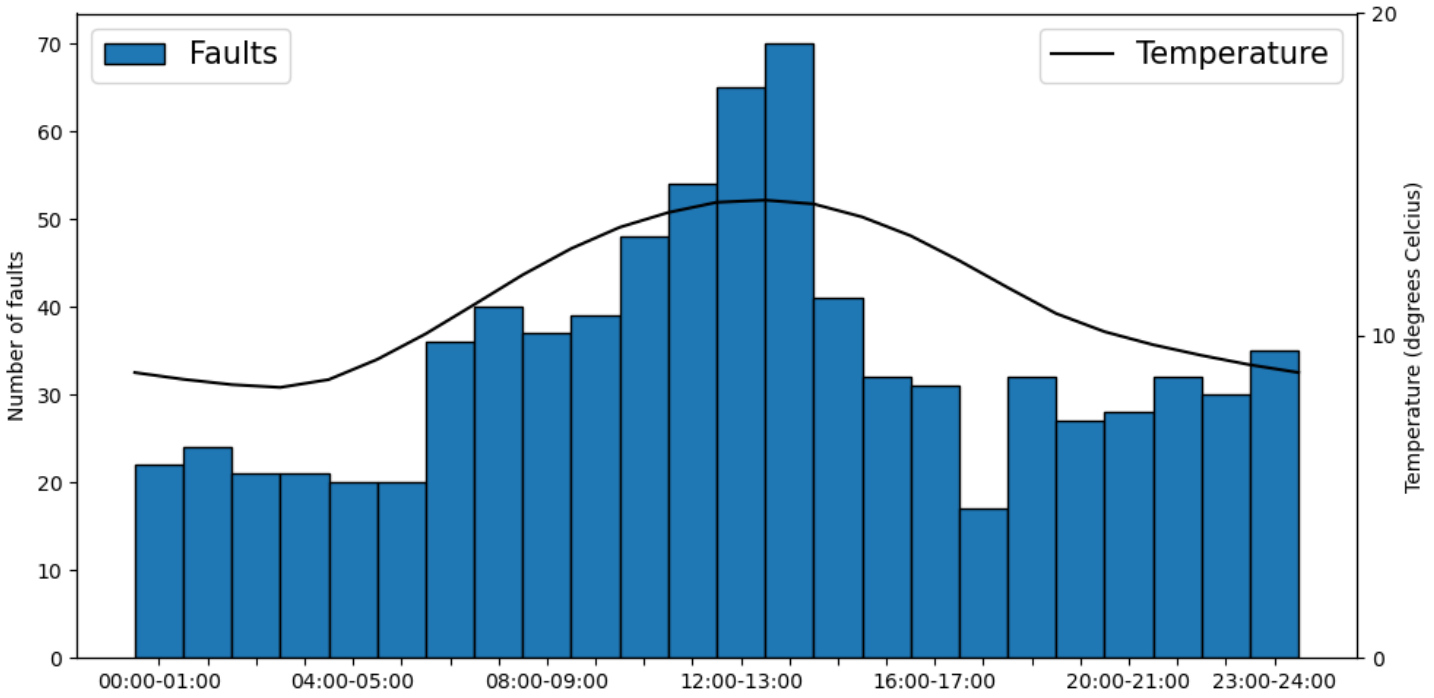


Figure 13: Mean of the air temperature and sum of number of faults per hour of the day for the period December 2014 until February 2021

bin represents the total number of faults registered during a period of the day of 1 hour. So we see 24 bins and they add up to the total number of faults, which is 740. For example there have been 19 faults between 00:00 and 01:00 over the period December 2014 till November 2020, and 68 faults between 13:00 and 14:00. After 14:00 there is a sudden decline of the number of faults. Some faults are caused by human activities, for example maintenance work and excavation damage. It is likely that these activities hardly take place after 2 pm.

For the same time period (December 2014 till November 2020), we have been looking at the temperature measured in De Bilt. For each day of this period we have the data of all the hours of the day. In Figure 13 we see the average of all these days. We see the maximum at the middle of the day and the minimum at night. We can already see a positive correlation between temperature and faults. The calculations confirm this: there is a correlation of 0.69 between the temperature and the number of faults distributed over the hours of a day. We use the Pearson correlation coefficient for this. See Section 3.1 for more details on the correlation and the definition of the Pearson.

4.2.3 Month

The temperature in summer is higher than in winter. Would the number of faults also be higher in summer? For this we have looked at the temperature and the number of faults per month for the entire period from December 2014 to February 2021.

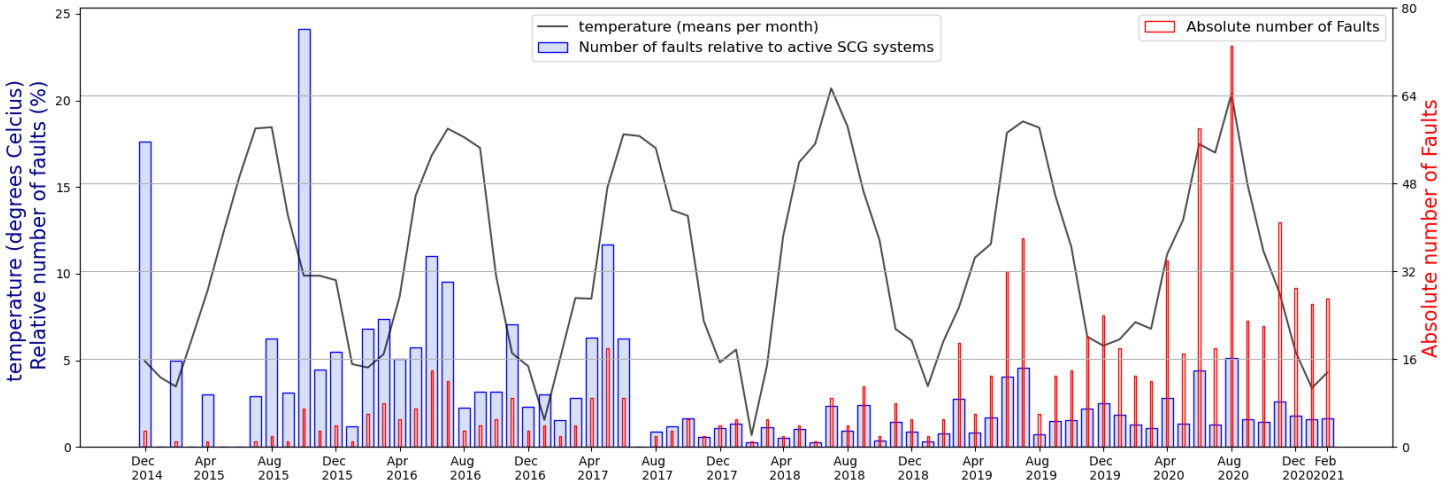


Figure 14: Faults and relative number of faults per month compared with the average temperature per month from December 2014 until February 2021

For the temperature we take the average temperature of all days of each month. This is shown in Figure 14 on the left vertical axis. We immediately see the difference in temperature between the summer months and the winter months. For each year we see that the maximum temperature is reached approximately in July. The blue bars indicating the relative number of faults are explained soon. The red bars show the number of faults for each month. This is shown on the right vertical axis. We see that in 2015 there are several months without faults and many more faults in 2019 and 2020 than before. The most likely explanation for this is that there are many more SCG-systems that register the faults. The distribution of the number of active SCG-systems is shown in Figure 15.

The number of active SCG-systems started to increase very rapidly from 2017. Returning to Figure 14, we see this pattern again in the evolution of the red bars. This makes it difficult to compare months from different years. Finding a pattern between the weather and the number of faults is more difficult if the number of faults is also effected by the number of active SCG-systems. That is why we will look at the relative numbers of faults per month. That is the number of faults per month divided by the number of active SCG-systems that month. In Figure 14 you see these values represented by the blue bars with the numbers on the left vertical axis. The relative number of faults is multiplied by 100 such that it can be shown using the left vertical axis of the graph.

The blue bars are much bigger at the start of the period. This suggests that the SCG-systems, which were active back then, are much more effective. This has probably something to do with the strategy of the placement of the systems. At first Alliander only placed the SCG-systems on vulnerable circuits. The strategy of the placements changed in 2017. So it makes sense to look at the relative number of faults from 2017 on.

In Figure 16 we see the same data as in Figure 14 but here from 2017 on. The blue bars show maximums in the summer months. So it seems there is a relation between relative number of faults and temperature. Calculation shows us that the correlation in this period is 0.40 which is a significant high correlation.

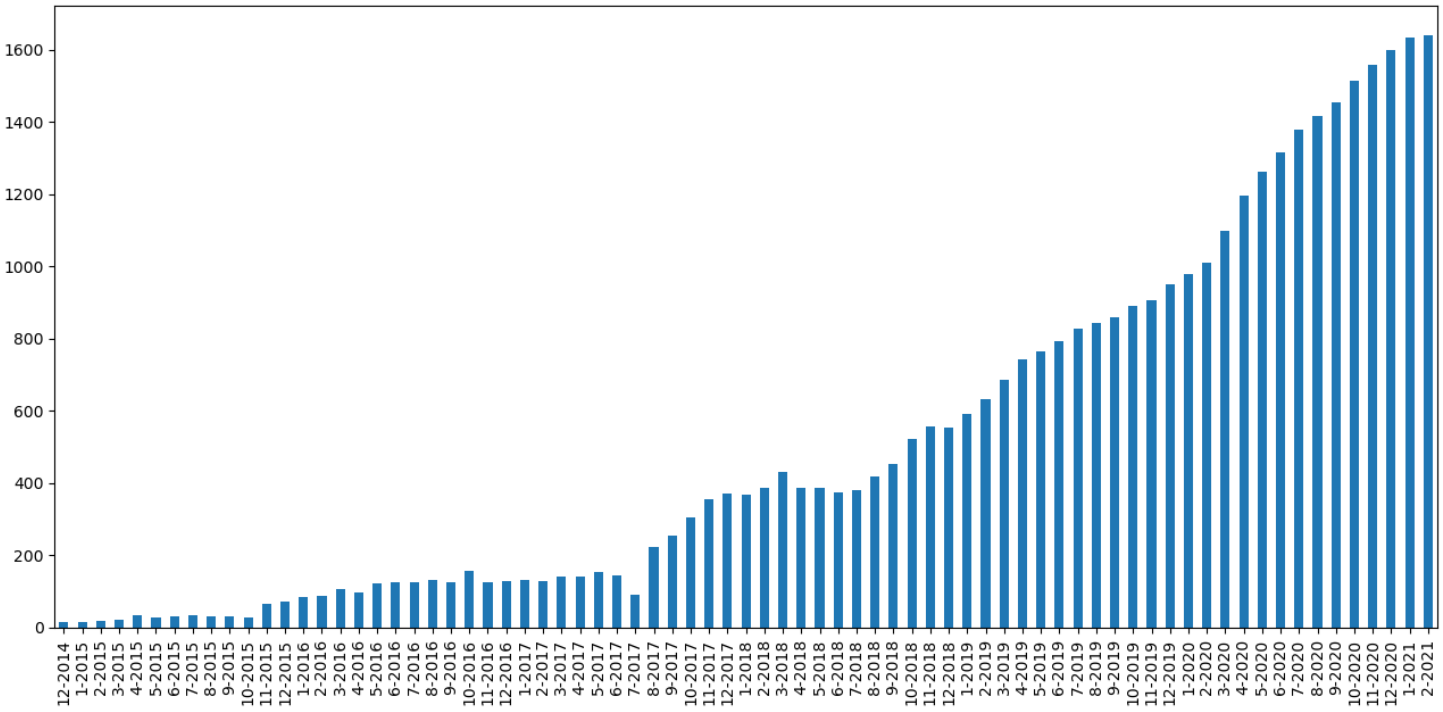


Figure 15: Number of SCG-systems in use over time. Note that there are more circuits (2840) than active SCG-systems (1640) in February 2021 because some circuits stopped being monitored by SCG.

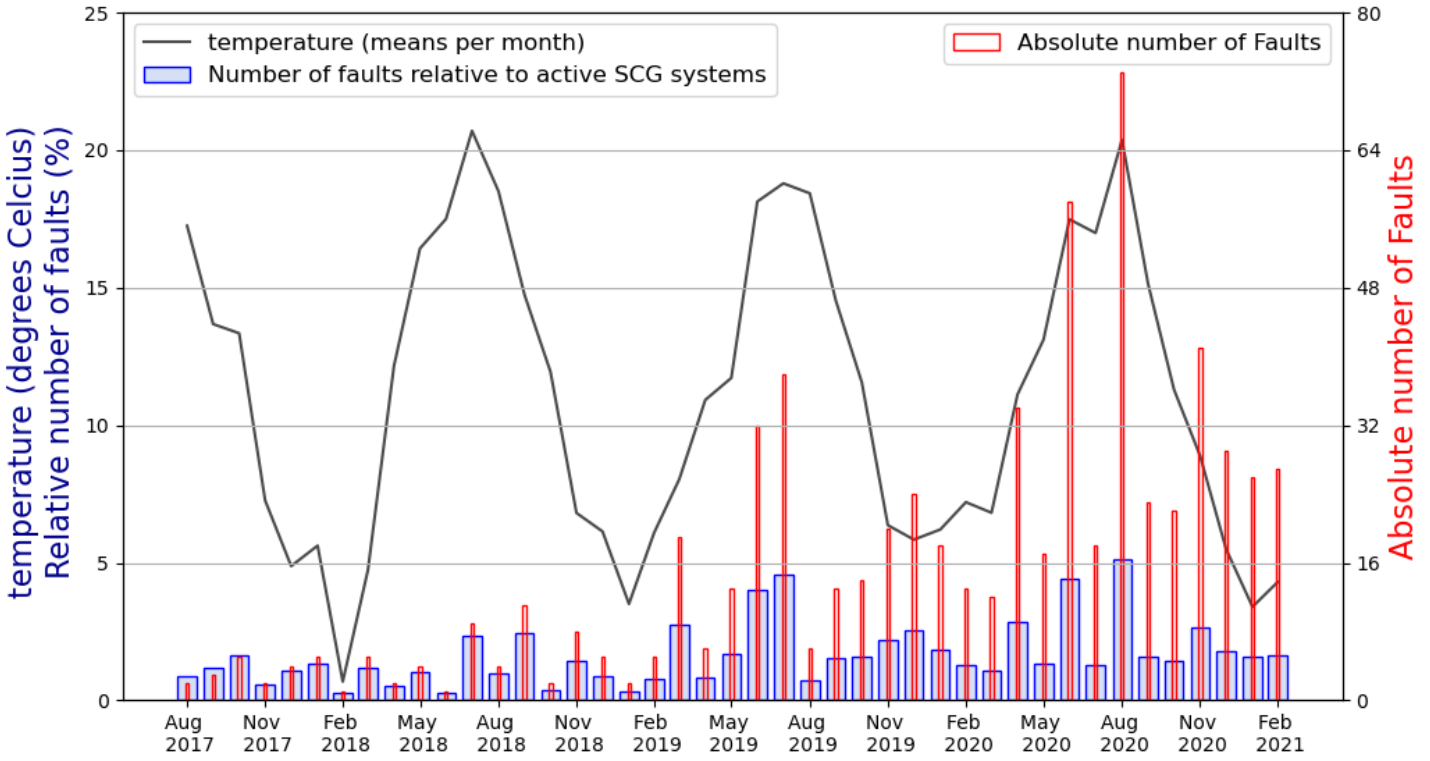


Figure 16: Faults and relative number of faults per month compared with the average temperature per month from August 2017 until February 2021

4.2.4 Correlate faults and weather variables

In the previous section we only used air temperature versus faults, which resulted in a correlation of 0.40. However we can compare many more weather variables with the number of faults. Let's investigate which variable correlates best with the faults. We will look both at the absolute and the relative number of faults. We compare these with three periods: the whole period (December 2014 - February 2021), the period in which the strategy of the placement of the SCG-systems changed (Augustus 2017 - February 2021) and the period in which we also have access to the CDS features including the soil temperatures (January 2018 - February 2021). The results are in Table 14 in the Appendix.

The features are sorted on the last column: Relative Faults in the period January 2018 - February 2021. Table 2 shows the weather features for which the correlation coefficient between the weather feature and the relative number of faults from January 2018 is at least 0.35. We see that the weather feature *T10N* has the highest correlation. *T10N* is the minimal temperature measured at a height of 10 cm. The soil temperature is measured at several depths. *soil_temperature_level_1* is at a depth of 0-7 cm, *soil_temperature_level_2* at 7-28 cm and *soil_temperature_level_3* at 28-100 cm. We pay more attention to the soil temperatures in Chapter 5. The meaning of many other features is evident. Detailed explanation can be found on the website of KNMI [23, 21].

The column *Relative 12-2014* shows very low correlations. This is due to the fact that the strategy of placement of the SCG-systems was very thought through. This caused relative many registered faults in December 2014 for example. The numbers we find in the columns of 2017 are a little bit lower than the correlations in the columns of 2018. This is caused by the low number of faults in the summer 2018 in comparison to the summers of 2019 and 2020. However both in 2017 and 2018 we see that the columns of the relative faults show higher correlation, meaning that this is a better indicator for the relevance of the weather features.

Judging on the last column, the relative faults from 2018, we see that all kinds of temperatures appear at the top of the table, meaning they correlate well with the faults. And thus the weather features about the temperature are the most relevant to predict faults.

Alliander knows that the cables are 80-100 cm deep in the ground. This is why we are most often going to use the feature *soil_temperature_level_3*, which is the temperature of the soil measured at a depth of 28-100 cm. It is surprising that this feature does not correlate best with the relative number of faults. We have no clear explanation for this. Although the temperature above the ground fluctuates more which could lead to a higher correlation coefficient. Also one can argue that the difference between the correlation coefficient between *soil_temperature_level_3* and the relative number of faults from January 2018 (0.35) and the correlation coefficient between *T10N* and the relative number of faults from January 2018 (0.44), is not significant. Moreover, the temperature is not at all the only factor for the number of faults.

	Feature	Absolute 12-2014	Relative 12-2014	Absolute 8-2017	Relative 8-2017	Absolute 1-2018	Relative 1-2018
0	T10N	0.29	0.05	0.34	0.41	0.4	0.44
1	temperature_min	0.29	0.05	0.35	0.41	0.4	0.44
2	precipitation_max	0.27	0.06	0.38	0.4	0.43	0.43
3	temperature	0.29	0.04	0.34	0.4	0.36	0.42
4	temperature_max	0.28	0.03	0.32	0.38	0.34	0.39
5	soil_temperature_level_1	nan	nan	nan	nan	0.27	0.39
6	soil_temperature_level_2	nan	nan	nan	nan	0.27	0.39
7	2m_dewpoint_temperature	nan	nan	nan	nan	0.28	0.39
8	2m_temperature	nan	nan	nan	nan	0.26	0.39
9	humidity_min_hour	0.13	0.01	0.17	0.3	0.19	0.35
10	soil_temperature_level_3	nan	nan	nan	nan	0.25	0.35

Table 2: Correlation coefficient between absolute/relative number of faults and several weather variables for three periods: from December 2014 until February 2021, from August 2017 until February 2021 and from January 2018 until February 2021. Some weather features are only known from January 2018. This table only contains the available weather variables whose correlation coefficient with the relative number of faults from 2018 is at least 0.35. See Table 14 in the Appendix for all weather variables.

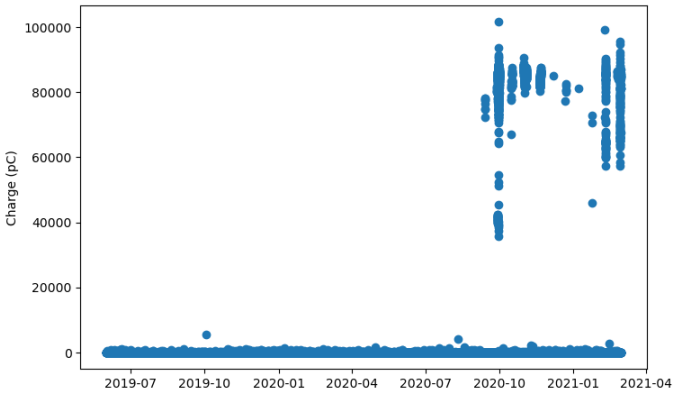
4.3 Partial Discharges

The main functionality of SCG is measuring partial discharges (PD) to predict the faults. A partial discharge is a small charge displacement in the cavity or layer of the insulation of a component. See Chapter 2 for more details on the measurements of partial discharges.

SCG measures these discharges every minute using a measurement session of 20 ms. on each circuit and we get the data as shown in Figure 17(a). Many times there is no discharge measured. That is why we see a lot of empty cells in the dataframe. In Figure 17(b) are only the observations that happened on a location of 2213.84 meter with an bandwidth of 1% of the circuitlength, which is 2215 meter for this circuit.

	Date/time (UTC)	Location in meters (m)	Charge (picocoulomb)
818560	2021-02-28 23:09:00	nan	nan
818561	2021-02-28 23:10:00	2213.84097	85270.50000
818562	2021-02-28 23:11:00	nan	nan
818563	2021-02-28 23:12:00	nan	nan
818564	2021-02-28 23:13:00	nan	nan
818565	2021-02-28 23:14:00	nan	nan
818566	2021-02-28 23:15:00	nan	nan
818567	2021-02-28 23:16:00	nan	nan
818568	2021-02-28 23:17:00	nan	nan
818569	2021-02-28 23:18:00	nan	nan
818570	2021-02-28 23:19:00	nan	nan
818571	2021-02-28 23:20:00	98.79283	471.00000
818572	2021-02-28 23:20:00	2213.84097	87107.00000

(a) Dataframe



(b) Scatterplot of measured charge around 2213.84 meters

Figure 17: Charge of circuit 4099

4.3.1 Clusters of Partial Discharges

Many of the observations of SCG are noise instead of PD. In addition to PD, SCG also measures noise signals from, for example, power electronics or industry. Noise is no predictor of faults at all which is why we would like to filter this. Alliander has developed models to filter this noise. The cluster algorithm clusters the discharges that come from the same source. In Figure 18 is an example of circuit 2719. Most of the clusters are noise and we use a classification model to classify the clusters as noise or dangerous. We pay more attention on the classification model in Chapter 8. If we do not know yet if an observation is noise or a partial discharge, we call it a **particle**. Although it is often not known whether it is PD or noise, particle are hereinafter often referred to as PD for convenience.

We see all particles measured on circuit 2719 in Figure 18. On the vertical axis is the time from September 2019 until May 2021. On the horizontal axis is the location of the circuit in meters. The dots in the plot are the particles. The magnitude of the charge is not shown in this plot. However much can be deduced from the density of the observations. Between September 2019 and November 2019 there were almost no particles registered, probably because the circuit was in maintenance during this period. The cluster algorithm clusters the PD and each cluster is given a unique color. The noise is left grey.

4.3.2 Table of clusters and their features

The particles that are not clustered are noise and they are further ignored, because Alliander assumes noise cannot be used to predict faults. The clusters are used to predict faults but most clusters still consist entirely of noise. To determine which clusters are noise and which clusters are likely to be followed by a fault, a **masterframe** is built.

The masterframe is a table with all clusters of every circuit. Each row represents a single cluster, and each column represents a feature of the cluster. From the experience of the experts, Alliander has developed a list of features that is known to be predictive. Examples of the most basic cluster features are the time, location, shape, and magnitude of charge. Faults most often occur in the joints instead of other parts of the circuit. This is why the nearest joint type is another important feature of the cluster: the type of joint closest to the median location of the cluster. In Table 3 we see a small part of the masterframe. The actual masterframe consists of thousands

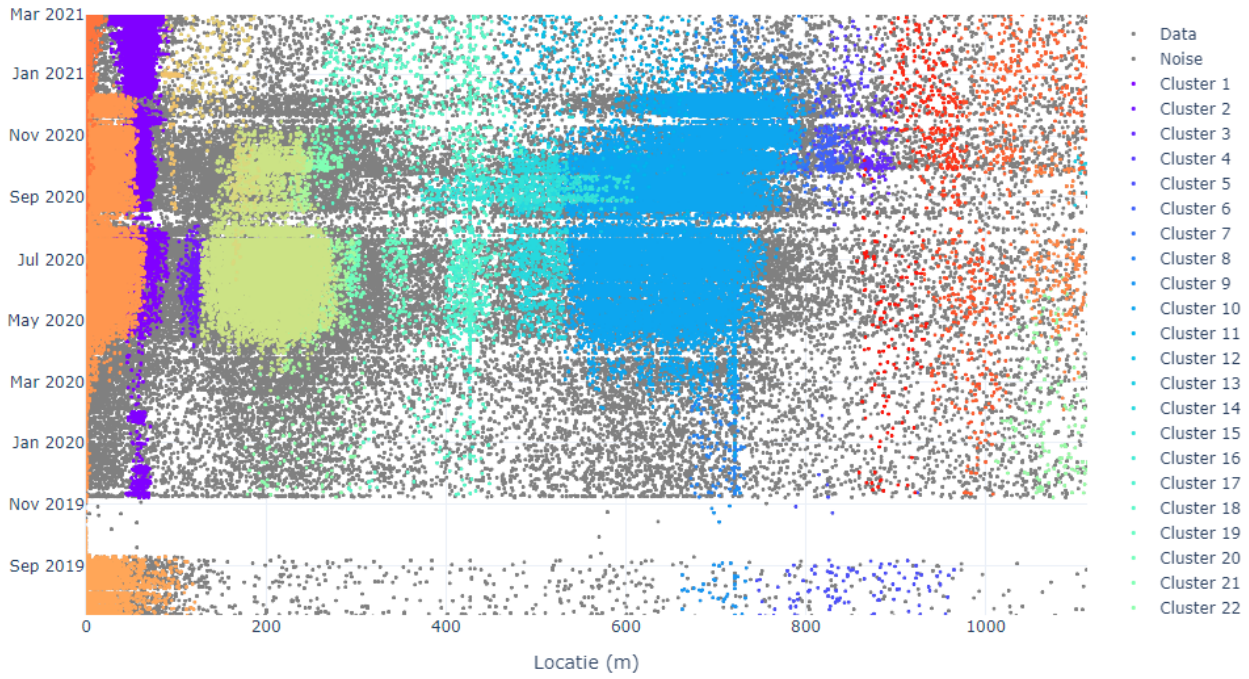


Figure 18: Particles of circuit 2719 clustered by the cluster algorithm. Each color represents a separate cluster. The particles that are not clustered are left grey and considered noise.

of clusters and more than hundred features. The first cluster in the masterframe started in December 2014. For both the clusters and the faults we have data from the same period.

The masterframe is used to predict faults for the given clusters. An important task is to construct cluster features that have a high predictive value. In the following chapters, we will look for those features and convert them into numbers so that they can be used in the masterframe. Chapter 8 explains how the faults are predicted by the masterframe and shows how well the features contribute.

circuitnr	startdate	enddate	circuitstartloc	circuitlength	Charge (picocoulomb)_sum	Location in meters (m)_median	nearest-jointtype	nearest-jointlocation	nearest-jointdistance	secondnearest-jointtype
4385	1-4-2019	1-11-2020	NEWTONSTRAAT 46	1101.00000	4970746.00000	657.40642	Joint (oil)	662.00000	4.59358	Joint (unknown)
4385	1-4-2019	1-11-2020	NEWTONSTRAAT 46	1101.00000	284011.50000	911.93495	Joint (oil)	921.00000	9.06505	Joint (heat shrink)
4385	1-4-2019	1-11-2020	NEWTONSTRAAT 46	1101.00000	838500.00000	883.34322	Joint (oil)	921.00000	37.65678	Joint (heat shrink)
4385	1-4-2019	1-11-2020	NEWTONSTRAAT 46	1101.00000	78226156.00000	716.96758	Joint (oil)	662.00000	54.96758	Joint (oil)
4387	1-4-2019	1-11-2020	26A	1822.00000	574343.50000	1591.99217	Joint (oil)	1580.00000	11.99217	Joint (oil)
4387	1-4-2019	1-11-2020	26A	1822.00000	450337.00000	1063.44612	Joint (heat shrink)	1113.00000	49.55388	Joint (unknown)
4387	1-4-2019	1-11-2020	26A	1822.00000	64677859.00000	731.93359	Joint (heat shrink)	721.00000	10.93359	Joint (heat shrink)
4387	1-4-2019	1-11-2020	26A	1822.00000	292666.00000	97.97958	Joint (oil)	55.00000	42.97958	Joint (oil)
4387	1-4-2019	1-11-2020	26A	1822.00000	14299590.50000	904.74620	Joint (oil)	914.00000	9.25380	Joint (heat shrink)
4388	1-4-2019	1-11-2020	LANGE BALKWEG 1	3233.00000	561897.50000	1142.85768	Joint (heat shrink)	1148.00000	5.14232	Joint (heat shrink)
4388	1-4-2019	1-11-2020	LANGE BALKWEG 1	3233.00000	461518.50000	973.66103	Joint (heat shrink)	974.00000	0.33897	Joint (resin)
4388	1-4-2019	1-11-2020	LANGE BALKWEG 1	3233.00000	162142.50000	2907.47999	Joint (unknown)	3003.00000	95.52001	Joint (unknown)
4388	1-4-2019	1-11-2020	LANGE BALKWEG 1	3233.00000	462380.00000	2701.21746	Joint (unknown)	2656.00000	45.21746	Joint (unknown)
4388	1-4-2019	1-11-2020	LANGE BALKWEG 1	3233.00000	9577407.00000	889.67665	Joint (resin)	907.00000	17.32335	Joint (heat shrink)

Table 3: Fraction of the masterframe of clusters and their features. Only a few clusters and features are shown.

4.3.3 Adding faults to the masterframe

Alliander has not analysed the data of the faults yet so as part of the research we add faults as a feature to the masterframe to gain insight into the masterframe in relation to the faults. As explained in Section 4.1, a location in meters and a time are also stored when the faults are registered. A cluster of PD has a different location for each particle, as can be seen in Figure 18. So we have to come up with a definition which will be used to determine which faults belong to which clusters such that we can add the faults to the masterframe.

Definition 6. A fault is linked to a cluster if

1. the fault occurred at a location between the 5th percentile and the 95th percentile of the location of the particles of the cluster;

2. the fault occurred after the first particle of the cluster has been detected.

Note that Definition 6 allows a fault to be linked to multiple clusters. This makes some sense because many faults occur months after a cluster and some clusters are right after each other in time, so sometimes it is not clear whether one cluster caused the fault or that multiple clusters contributed to the fault.

By only looking between the 5th percentile and the 95th percentile of the location, we ignore the outliers. That makes sense because the partial discharges occur close to where the fault will occur in general, so a fault is not likely to occur close to the sides of a cluster. The 5th and 95th percentiles were chosen to ensure that the correct faults are linked to the cluster, given that the precision of the localization of both the particles and the faults is 1% of the circuit length.

A fault and a cluster of PD are often caused by degradation of the insulation in a joint or cable. However, the PD precedes the faults so we interchange the following expressions: faults linked to clusters and faults caused by clusters.

We add columns to the masterframe that provide information for each cluster about any fault(s) caused.

- *fault-count_inside/after_cluster*: Number of faults linked to the cluster;
- *Date/time (UTC)_of_first_fault*: Time of the first fault linked to the cluster;
- *location_of_first_fault*: Location of the first fault linked to the cluster;
- *locationdelta_of_first_fault*: The distance between the location of the first fault and the median location of the cluster;
- *locationdeltarelative_of_first_fault*: The distance between the location of the first fault and the median location of the cluster, divided by the circuitlength.

It is cumbersome and unnecessary to store the data of all faults in this format of the masterframe. This is why we have chosen to only store the data of the most important fault of the cluster, the one that occurred first. If *fault-count_inside/after_cluster* is 0, then the cells in the other columns are left empty. In Table 4 we see a part of the masterframe in which the columns related to the faults are added.

circuitnr	label	Location in meters (m)_q5	Location in meters (m)_q95	fault-count_inside/after_cluster	Date/time (UTC)_of_first_fault	location_of_first_fault	locationdelta_of_first_fault	locationdeltarelative_of_first_fault
1478	9.00000	127.91991	2857.42880	3.00000	21-12-2019 00:11	547.15532	901.53807	0.30253
1478	1.00000	2850.34847	2978.94045	2.00000	16-1-2020 03:06	2978.94037	38.27910	0.01285
1478	0.00000	577.96321	611.87038	1.00000	20-12-2019 20:32	611.88919	12.34869	0.00414
1478	2.00000	683.76414	727.25127	0.00000	nan	nan	nan	nan
1478	3.00000	396.08167	683.75160	0.00000	nan	nan	nan	nan
1478	4.00000	0.00000	118.67256	0.00000	nan	nan	nan	nan
1478	5.00000	2578.76411	2894.41950	0.00000	nan	nan	nan	nan
1478	6.00000	2266.08959	2777.29727	0.00000	nan	nan	nan	nan
1478	7.00000	1790.94538	2213.23057	0.00000	nan	nan	nan	nan
1478	8.00000	1572.10401	1673.82338	0.00000	nan	nan	nan	nan
1478	10.00000	1245.36986	1458.05757	0.00000	nan	nan	nan	nan
1478	11.00000	792.91606	884.74674	0.00000	nan	nan	nan	nan

Table 4: Fraction of the masterframe with all clusters of circuit 1478 and the features related to the faults.

4.3.4 Exploration of the clusters in the masterframe

Using the columns about the faults, we can explore the masterframe. There are 2840 unique circuits, of which 1549 occur in the masterframe. So for 1291 circuits there is no cluster of PD. There are 12345 clusters from which 3137 are still active. We call a cluster active if their last particle is measured less than 24 hours ago. Depending on the density of the cluster, new particles could be added to the active cluster, even if there is a small gap in time.

Also not all faults can be found in the masterframe: there are 822 faults, of which 190 have a circuit number that does not occur in the masterframe. So at least this number of faults could not have been predicted using the masterframe. Multiple faults occur on some circuits, so this means that there may be even more faults that cannot be found in the masterframe. There were more faults on 214 circuits than the number of clusters they are linked to. This is possible if the fault is not caused by a cluster, or if the cluster caused multiple faults.

So at least 214 faults are either caused by a cluster causing multiple faults or cannot be predicted by a cluster, and for at least 190 faults we know for sure that they cannot be predicted by a cluster.

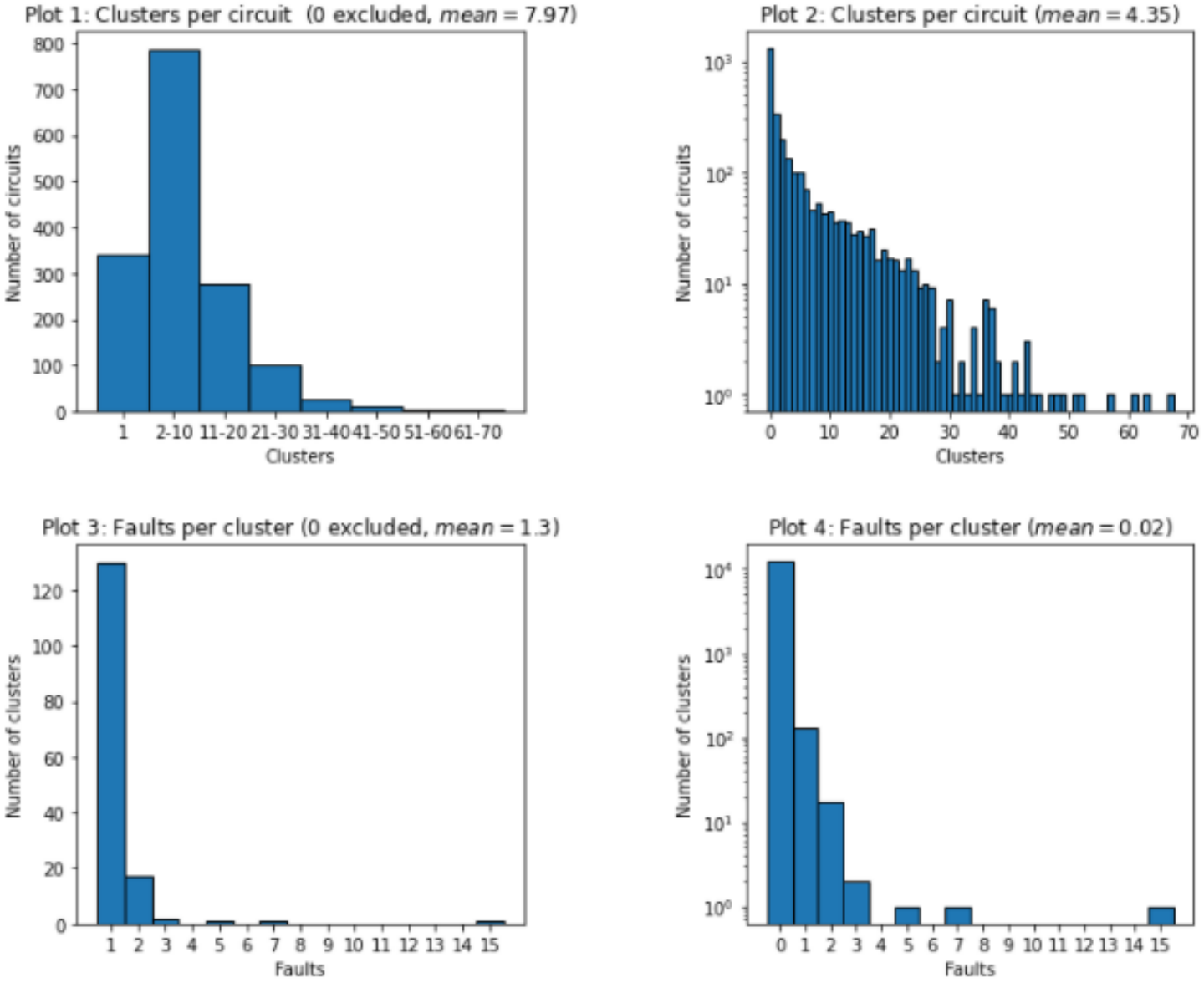


Figure 19: Distributions of the circuits across the number of clusters per circuit and the clusters across the number of faults per cluster.

The distribution of the 1549 circuits is shown in Plot 1 of Figure 19. Most circuits have multiple clusters of PD. A circuit in the masterframe has 7.97 clusters on average and most of the circuits have less than 10 clusters. In Plot 2 of Figure 19 we see that most circuits only have 1 cluster of PD. Note that also the circuits which do not occur in the masterframe are taken into account in Plot 2. Also note the logarithmic scale.

Plot 3 and 4 show the number of faults each cluster is linked to, using Definition 6. All faults are caused by 152 clusters and we see in Plot 3 that by far the most clusters cause only 1 fault. There are many more clusters that do not cause any faults as can been seen in Plot 4, note again the logarithmic scale. However we see that there is a cluster which has caused 15 faults. This is a cluster with median location 3574.18 on circuit 3532. These faults were already mentioned in Section 3.1.1.

From Plot 4 one cannot say how many clusters have caused faults because multiple faults could be caused by a single cluster. For this we look at Figure 20. We see that most clusters do not cause any faults, 51 faults overlap only one cluster and 5 faults overlap 2 clusters. We can conclude that for most of the faults a single cluster can be identified as the causer. However we cannot deduce from Figure 20 how many clusters cause faults because a cluster can cause multiple faults. Plot 4 of Figure 19 and Figure 20 combined tell us that there are 56 faults caused by 152 clusters. So only about $\frac{56}{822} = 7\%$ of the faults are caused by the clusters, so by looking at the masterframe only 7% of the faults could have been predicted. Note that many faults have already been prevented by replacing joints preventively. Also note that although filtering has taken place, the fault dataset still contains false positives: events that were incorrectly issued as faults.

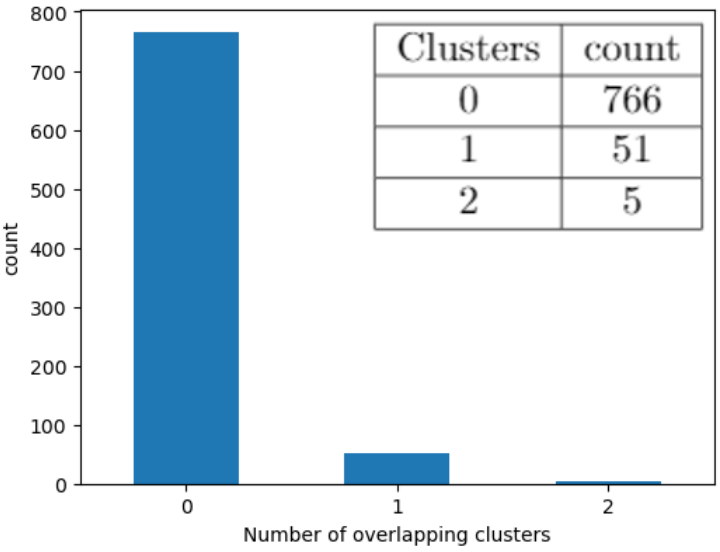
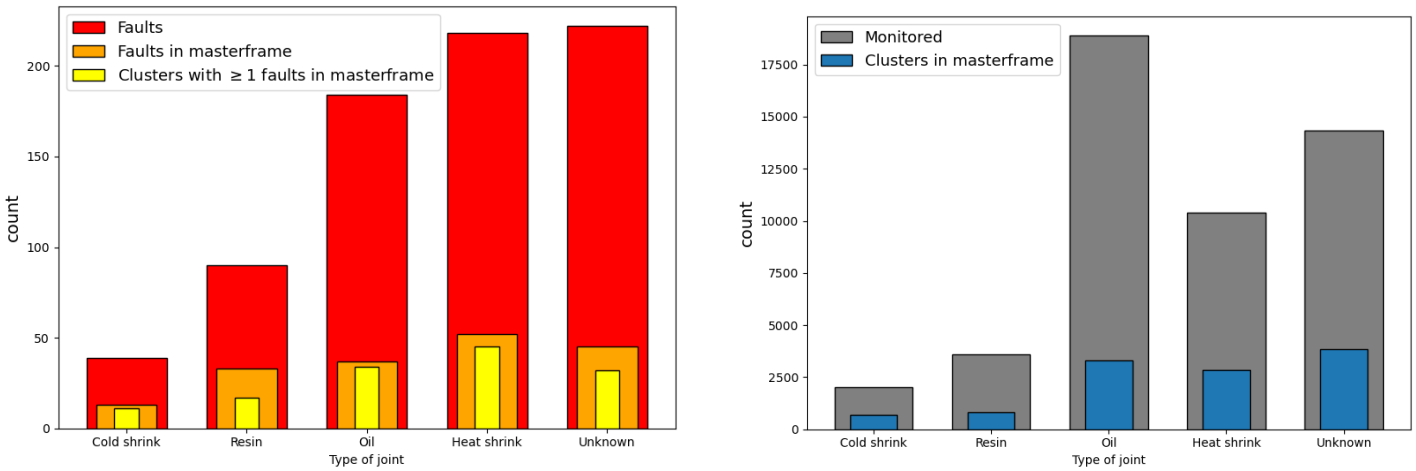


Figure 20: Distribution of the 822 faults across the number of clusters they are linked to. Most faults are not linked to a cluster of the masterframe.

Would a certain type of fault be more predictable by the masterframe? To investigate this we use the feature nearest joint type. Almost all faults that are predictable using the masterframe, occur in a joint. So we will categorize the faults by the type of joint closest to the faults. Even if the location of the nearest joint and fault are not identical we can assume that the fault occurred in this joint because the precision of the measurements is not perfect. Figure 21 and Table 5 show the results.

For a lot of joints there is no data about the type. Beyond that most faults occur in heat shrink joints (red bars in Figure 21) despite the fact that more oil joints are monitored (grey). There are also more clusters with at least one fault in heat shrink joints (yellow) even though there are more clusters around oil joints (blue). Figure 21(a) also shows that most of the faults do not appear in the masterframe which explains why the red bars are much higher than the orange bars. The orange bars are higher than the yellow bars because there are more clusters with multiple faults than faults that appear in multiple clusters.

Only the most common types of cable joints are shown in the bar graphs. For many joints the type is unknown because many joints were made a long time ago and the historical data is not always complete. Nowadays if a fault occurs in a joint, the joint is most often replaced by a heat shrink joint. This is why the number of heat shrink joints is very high. Although not all replacements are registered correctly. The data contains outdated joint types. After a replacement the joint type usually changes but this is not included in our data for all replacements. So for some faults we see that it occurred in some type and for other faults we can only see what the current type of the joint is, and we do not know which joint types are recent and which are outdated. That is why it is hard to draw conclusions from these data.



(a) Three measures of faults per joint type

(b) Monitored joints and their clusters

Figure 21: Distribution of faults across the joint types in comparison with the number of clusters and monitored joints

Type of joint	Clusters with ≥ 1 faults	Faults in masterframe	Faults	Clusters	Monitored
Paper lapped	0	0	0	0	1
Grease	0	0	0	0	11
Premolded	0	0	1	29	83
Taped	0	0	5	19	139
Bitumen	2	2	14	175	1218
Silicone fluid	5	8	16	201	917
Polymer	6	7	33	305	958
Cold shrink	11	13	39	708	2019
Resin	17	33	90	835	3607
Oil	34	37	184	3314	18883
Heat shrink	45	52	218	2848	10404
Unknown	32	45	222	3911	14344
Total	**152**	**197**	**822**	**12345**	**52584**

Table 5: Total number of monitored joints per joint type and the number of faults and clusters per joint type

4.4 Manually assigned warnings

Alliander receives warnings from DNV. These warnings indicate whether a cable or cable joint is likely to cause a fault in the near future. They concern a circuit, the location on the circuit, the level of the warning and the start and end time of the warning. The levels of these warnings depend on the seriousness of the danger to cause a fault, see Figure 22 for a detailed description by DNV.

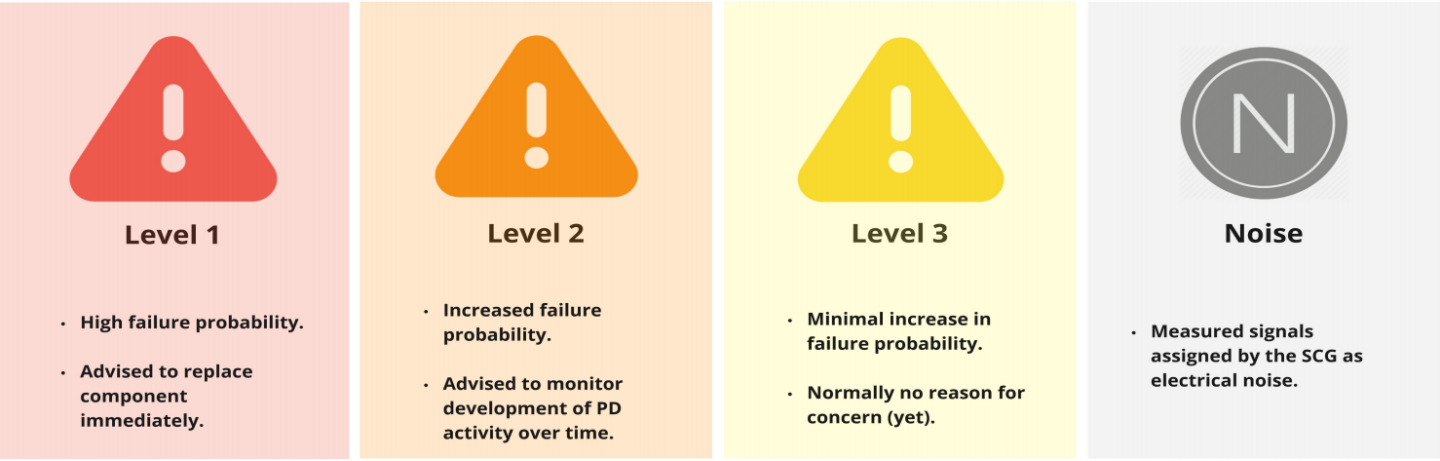


Figure 22: Warning levels assigned by DNV and their meaning according to DNV

We have access to the warnings from June 2012. However there were only 5 warnings before December 2014, so this data can be compared well to the faults and clusters. We check for each fault if there has been a warning for it. For each fault the overlapping warnings are considered. Then only the warning with the most serious level is linked to the fault.

Definition 7. A warning is linked to a fault if

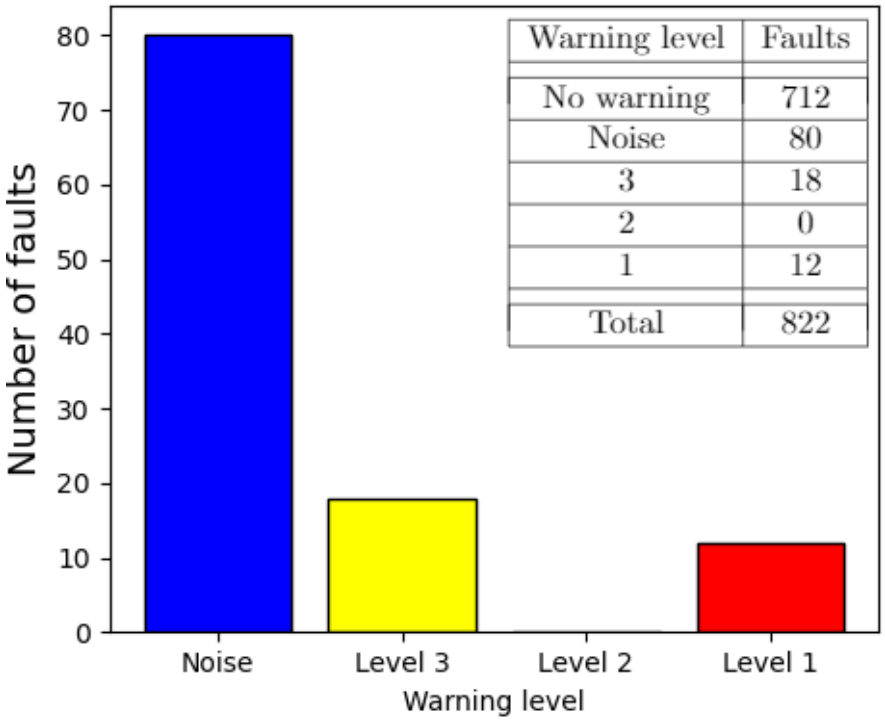
1. the difference between the locations of the warning and the fault is less than 3% of the circuitlength;
2. the warning start time is before the time of the fault.

If more than one warning is linked to a fault, we always chose the most severe warning to link the fault to.

The precision of the localization of the faults and the warnings should be approximately 1% of the circuitlength. We have noticed that this could be more in practice. For example the number of faults that overlap with a warning depends slightly on this bandwidth. When we choose 2%, 3% and 4%, respectively 16, 18 and 20 faults would be linked to a level 3 warning. The number of level 1 and 2 warnings would be the same in these three cases. Further investigation is necessary to choose the best bandwidth. For now we choose 3% so that we do not link warnings to faults incorrectly.

On the one hand we expect that the clusters which have caused a fault have received a warning. On the other hand the clusters that get a warning are very well inspected by hand and the relevant joint is replaced in time to prevent a fault.

In Figure 23 we see that 712 of the 822 faults have not received any warning at all. 80 of the remaining 110 faults have had a warning with the level 'noise'. Faults get a noise warning if the particles measured at the fault location are considered to be noise. So Alliander does not react to these 'warnings'. Only 30 faults are linked to an actual warning. It happens to be circuit 3532 again that plays a big role here. All 12 warnings with a level 1 warning occurred on this circuit. Also 3 other faults on this circuit got a level 3 warning. This remains 15 other faults with a level 3 warning that were not prevented.



Warning level	Faults
No warning	712
Noise	80
3	18
2	0
1	12
Total	822

Figure 23: Distribution of the 822 faults across the warnings they are linked to

To investigate how the warnings relate to the clusters of the masterframe, they are linked to the overlapping clusters. If multiple warnings overlap a cluster, only the most serious warning is linked to the cluster.

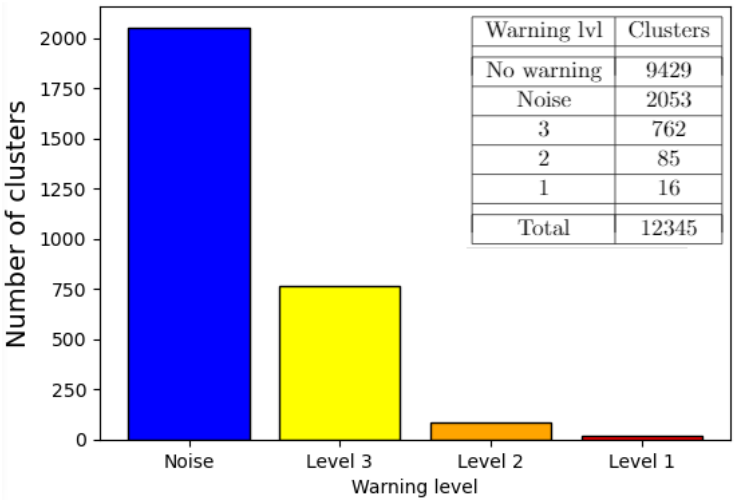
Definition 8. A warning is linked to a cluster if

1. the location of the warning is between the minimum and maximum location of the particles of the cluster;
2. the time period of the warning overlaps the time period of the cluster.

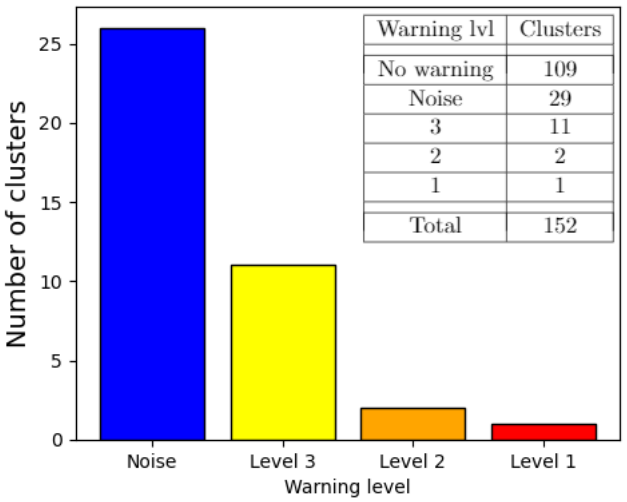
If more than one warning is linked to a cluster, we always chose the most severe warning to link the cluster to.

Figure 24 shows that there were 863 clusters with a warning (noise warnings excluded) and in only 14 of them occurred a fault. The main reason for this is that Alliander prevents faults that would have occurred without maintenance work. So warnings actually lead to prevention of faults. Note that level 3 warnings are no reason for concern (Figure 22), so even without any replacements of joints, there would be still many more clusters with a warning than the number of faults. It seems like assigning more warnings leads to more faults being prevented, because most of the clusters linked to at least one fault are not linked to a warning. Assigning the warnings to the right clusters is the challenge. Unfortunately we do not have access to data of the replacements of the joints. We do not know how many false positives there are: replacements of joints which would not cause a breakdown in short term.

Figure 24(b) shows that there are 14 clusters with at least one fault that are linked to a warning level 1, 2 or 3. However according to Figure 23 there are 30 faults linked to a warning level 1, 2 or 3. To explain this difference we need to look at the number of faults caused by clusters with a warning in Figure 25. There is only 1 cluster that is linked to a fault and got a warning level 1. This happens to be the cluster that caused 15 faults (12 of which are linked to warning level 1 and 3 are linked to warning level 3). We already discussed this cluster (of circuit 3532) as a consequence of Plot 3 of Figure 19. We also see there are 2 clusters with a level 2 warning. Further inspection showed that these clusters were on circuit 2719 and 20133 from respectively 2019-11-07 and 2019-10-22 and were still alive in 2021-02-28. The faults occurred respectively on 2020-08-08 and 2020-08-14 while the level 2 warnings were given on 2021-01-12 and 2020-11-27, so these faults could not have been prevented by these warnings. However the clusters kept on existing after this so the level 2 warnings were linked to them. The clusters also received level 3 warnings before the faults (on 2019-11-14 and 2019-10-28 respectively), but Alliander did not respond to it because there should be no reason for concern (Figure 22). We can conclude that the assigned warnings should have been level 1 or 2 instead of level 3.



(a) All clusters



(b) Clusters that are linked to at least one fault

Figure 24: Distribution of the clusters across the warnings they are linked to

There remain 13 faults with a level 3 warning and 11 clusters with a fault and a level 3 warning, to be inspected further. These 11 clusters have all been identified as causing only one fault. Those numbers are not equal for several reasons. Only 7 of these 11 clusters were linked to a fault while their faults had been linked to a warning. 2 of those 7 clusters (of circuit 2737) did overlap in both time and location and consequently both were both linked to the same fault. The other 4 of these 11 clusters kept on existing after the fault, which is why there could have been given the warning for this cluster while it is not linked to a fault. So this explains 6 faults with a level 3 warning. For the remaining 7 faults with a level 3 warning there is no cluster that overlaps the warning. So there were warnings for the faults, but there were no clusters for these faults.

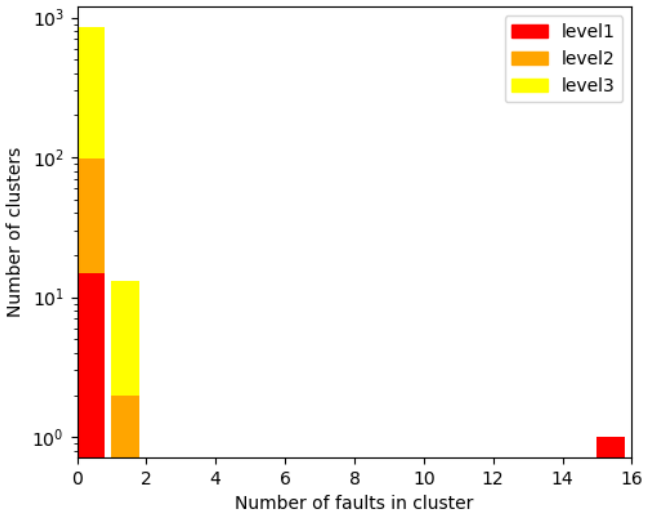


Figure 25: Distribution of the clusters that are linked to a warning across the number of faults they are linked to. The vast majority of clusters linked to a warning did not cause a fault.

In general Alliander reacts well to the warnings because almost all faults occur without a warning. These warnings are manually assigned by DNV. The ultimate goal of Alliander is to be independent of manual work by automatically assigning warnings using the masterframe.

4.5 Conclusions of the data exploration

4.5.1 Faults

The available data contains 2840 circuits with 822 faults. Not all these faults actually cause immediate outage of the circuits in which they are perceived. Some faults are intermittent and will cause a breakdown later on, and some faults are registered in a circuit while they actually occur in another nearby circuit. Since we cannot easily distinguish the different cases above in our data, **we will assume that all faults in the data are relevant for our analysis to predict faults.**

4.5.2 Weather

There are weather features that correlate with the number of faults. In general the temperature correlates strongly with partial discharges, and consequently we can use this to predict faults. The most important weather feature is the temperature of the soil measured at a depth of 28-100 cm. **So we will use the data of the soil temperature in the next chapters.**

4.5.3 Partial Discharges

The masterframe is a table of clusters of particles and features of the clusters. We assume that the particles are clustered well. The goal is to use the masterframe to classify the clusters as either PD or noise, without using the manually assigned warnings. We can predict the faults better if we introduce new features to the masterframe. **The next chapters search for cluster features that help the classification model to predict the faults better.**

4.5.4 Manually assigned warnings

Only 9% of the clusters are linked to a warning. This suggests that many clusters are noise. 7% of the faults were linked to a cluster and 4% to a warning. **So many faults cannot be predicted by analyzing the PD.**

We have access to the data of the faults, clusters and warnings between December 2014 and February 2021, and the data of the soil temperature from January 2018. **This is why we will focus on the period from January 2018 until February 2021 in our research.**

5 Temperature

Section 3.2.4 showed that temperature is related to the faults, because there were many more faults during the summers. In Section 3.3 we have seen that the masterframe with its clusters of PD is assumed to have a predictive power. So it seems that PD are related to the temperature. In this chapter we further examine the available temperature data to see if and how we can use this to relate to PD.

To what extend can we use the temperature data to relate them to the partial discharges?

5.1 Soil temperature measured at several depths

Most electricity cables and joints in the medium voltage grid lie underground. Therefore it seems reasonable that the temperature of the soil correlates more strongly than the temperature above the ground. KNMI measures the soil temperature and categorises the measurements to four levels depending on the depth. See Table 6. We only have access to the data from 2018.

The four different levels and the air temperature, used in figures 14 and 16, are plotted in figures 26 and 27. The measurements come from the gauge of De Bilt. As can be seen by visual inspection, the value of soil temperature level 4 has a delay of half a month on the regular temperature at +150cm. In fact the deeper the measurement the larger the delay. This also causes the difference in inconstancy: the deeper the measurements the less erratic the temperature.

The cables in the ground are located at a depth of 80-100cm, so we use the soil temperature level 3 for all analysis.

Level	Depth (cm)
1	0-7
2	7-28
3	28-100
4	100-289

Table 6: KNMI measures the soil temperature and categories the measurements to four depth levels

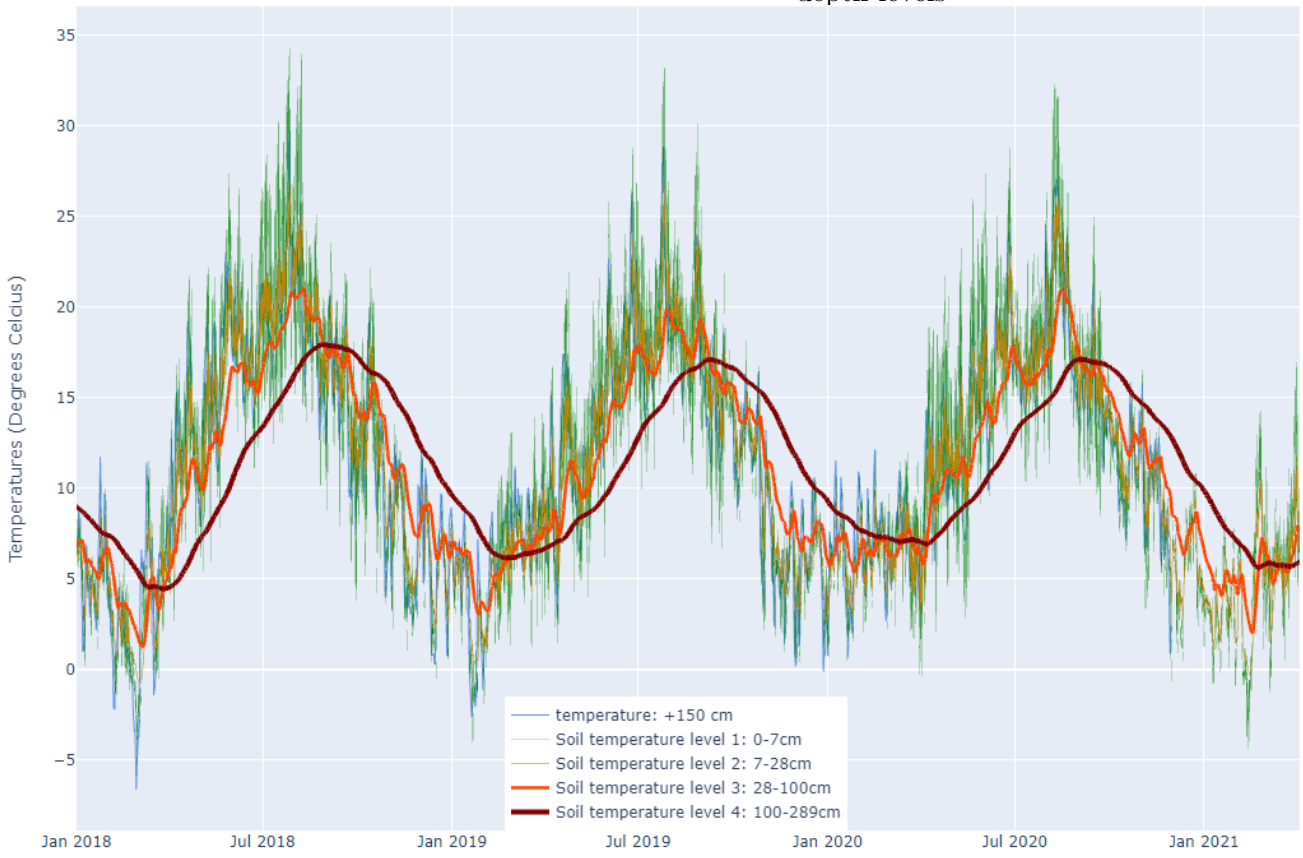


Figure 26: Temperature at +150cm, and depths of 0-7cm, 7-28cm, 28-100cm and 100-289cm in De Bilt

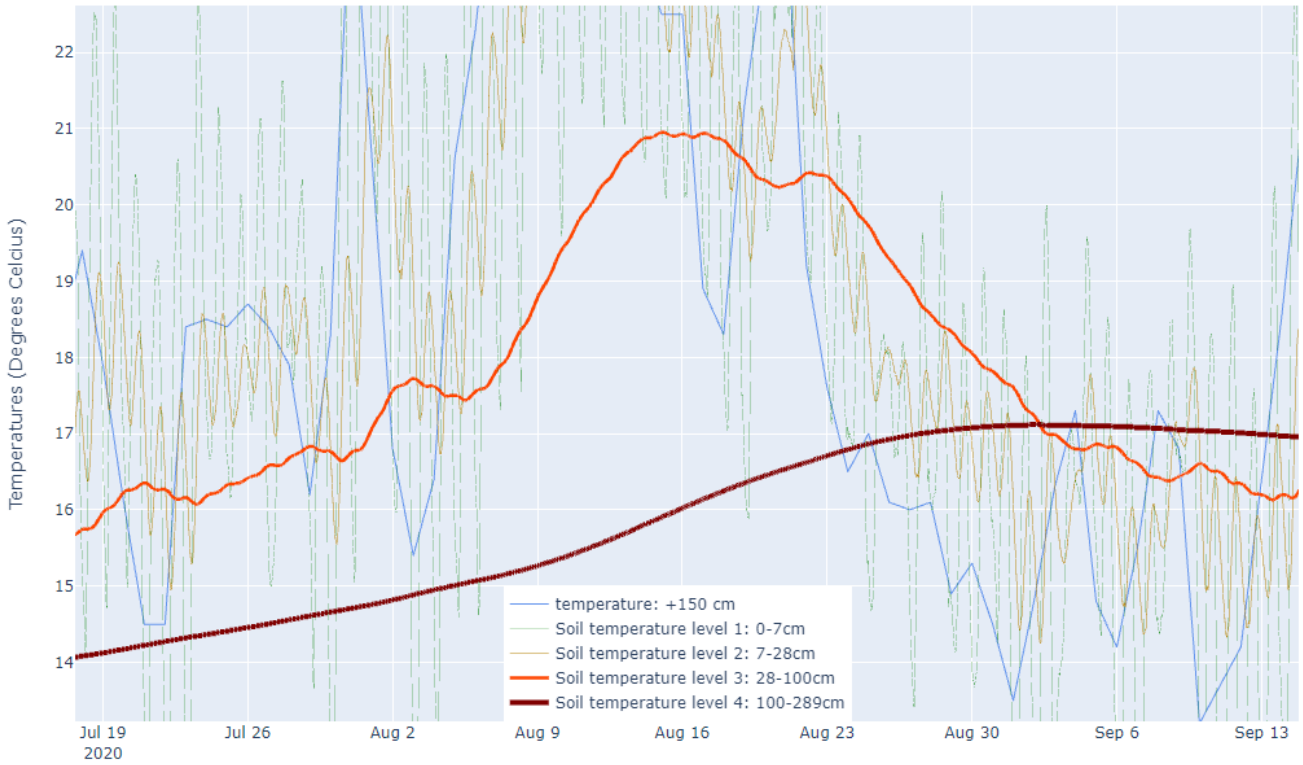


Figure 27: Temperatures at several depths in August 2020 in De Bilt

5.2 Variation between temperature measurements across the Netherlands

To relate the temperature to PD it is useful to know where the circuits are, such that the temperature of the specific regions can be used. Due to constraints in the availability of the data, location of SCG circuits has to be queried manually. So we cannot do extensive research on the locations of the circuits. However De Bilt is centrally located in the Netherlands, so it makes sense to use these measurements for all circuits across the Netherlands. To what extend can we use the temperature of De Bilt to say something about PD of other cities? To investigate this we look at the temperatures of De Bilt and two distant cities: Groningen and Maastricht. See Figure 28.

The variations between the three cities are made visible in Figure 29. The three curves are pretty similar, especially the slopes. But after zooming in, we see in Figure 30 that there are days in which the temperature in one city rises while it drops in another city. To see if we can still use these measurements to say something about the relation between PD and the soil temperature somewhere else in the country, we are first going to investigate how the measurements across the country correlate to each other.

Figure 28: The Netherlands and the locations of De Bilt, Groningen and Maastricht

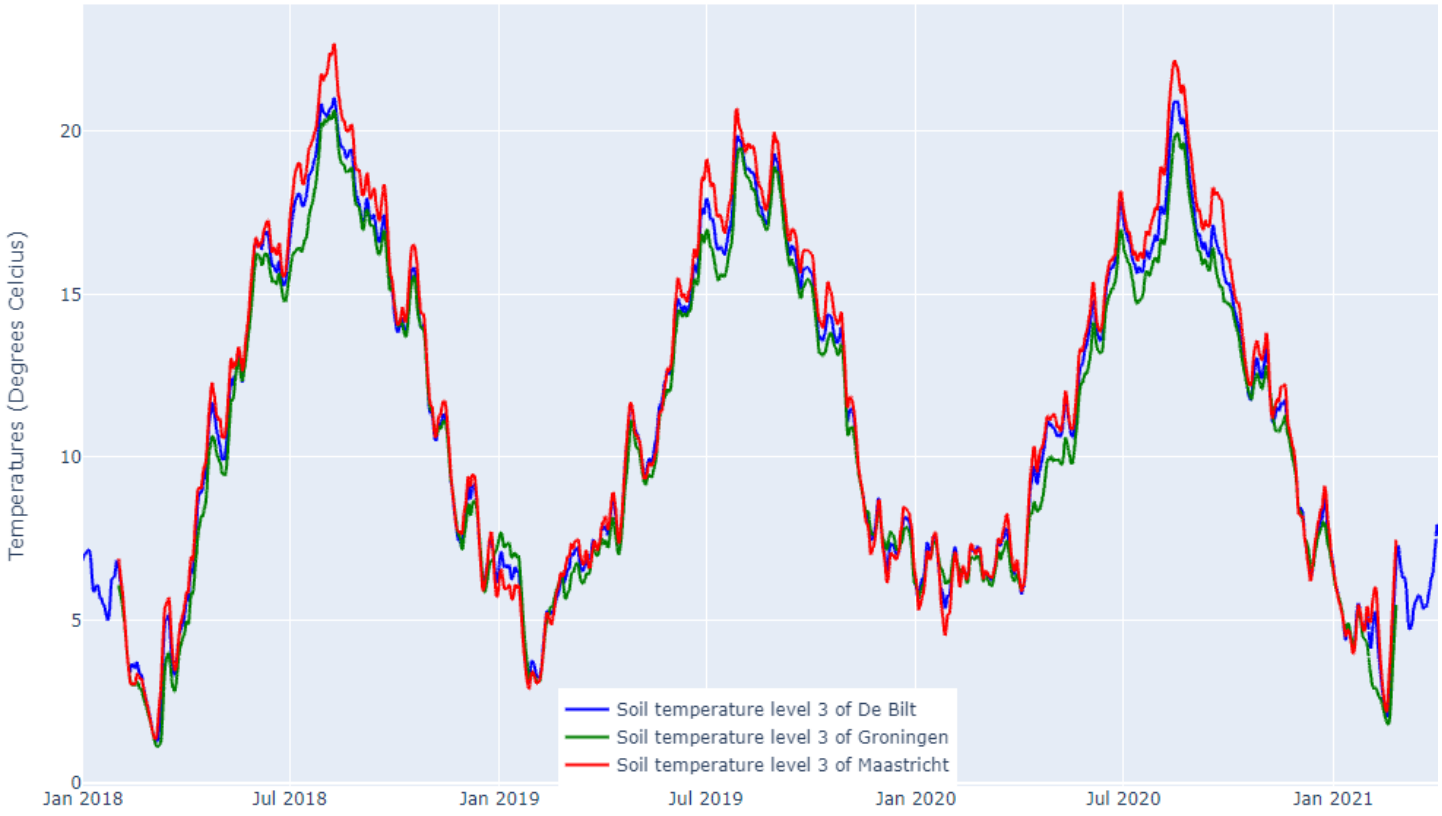


Figure 29: Soil temperature level 3 in De Bilt, Groningen and Maastricht

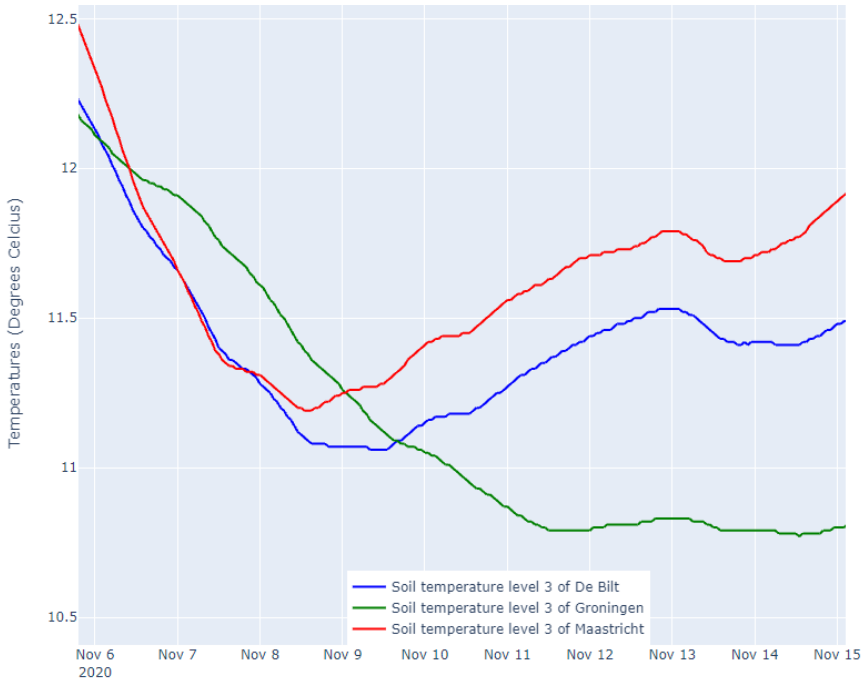


Figure 30: Soil temperature level 3 in November 2020 in De Bilt, Groningen and Maastricht

In all three cities the soil temperature is measured each hour and we use this data over a period of 3 years, so for each city we have 26304 data points. The mutual correlations of the cities are 0.9969, 0.99856 and 0.99371, for respectively [De Bilt - Groningen], [De Bilt - Maastricht] and [Groningen - Maastricht]. As expected, the correlation between Groningen and Maastricht is lowest, because these cities are the furthest apart. Note that the correlation coefficients are very high. We would like to predict PD on a much smaller window. Do the temperatures also correlate when we take a time window of a single day?

We have 24 data points for each day and each city. For each day we calculate the three mutual correlations. Then we have three lists of 1096 numbers, because the three years consist of 1096 days. To get some insight in these numbers we calculate the minimum, mean, median and some percentiles of the lists. See Table 7. See Section 3.2 for more details on the percentiles.

Cities	Mean	Median	25%	10%	5%	1%	Minimum
De Bilt - Groningen	0.845583	0.989404	0.922305	0.518964	-0.068412	-0.819790	-0.997063
De Bilt - Maastricht	0.888970	0.992834	0.956815	0.708682	0.244546	-0.643664	-0.969608
Groningen - Maastricht	0.780022	0.984655	0.878705	0.115848	-0.525885	-0.929132	-0.997418

Table 7: Mutual correlations between the three cities for which the rolling window consist of 1096 correlation coefficients. For each day a coefficient is calculated using two lists of 24 data points: temperature of each hour of the day of two cities.

We see that the fifth percentile is around 0 for all city combinations. This means that for 5% of the time there is no significant correlation. Also the mean of around 0.85 is not promising. We determine that the temperature for a circuit far away cannot be used to say something about PD. So when the temperature is rising during part of the day, we cannot link that to PD of a circuit far away. However if we take a much bigger time window, we may find a correlation between pd and temperature.

To get rid of the biggest fluctuations we use the mean of the temperature per day. We receive one data point per 24 hours. The mutual correlations are even higher now: 0.99691, 0.99857 and 0.99374, for respectively [De Bilt - Groningen], [De Bilt - Maastricht] and [Groningen - Maastricht]. Using the means of the days we calculate for each day the mutual correlations of the past 20 days. The results are in Table 8.

The correlation coefficients are much higher despite the fact that we use less data points (20 versus 24 before). There is even a significant correlation for the fifth percentile which means that at least 95% of the time there is a significant correlation and therefore it is very reasonable to look at the temperature in one city while predicting PD of another city. We can use the soil temperature level 3 in De Bilt to relate it to PD in Groningen and Maastricht.

Cities	Mean	Median	25%	10%	5%	1%	Minimum
De Bilt - Groningen	0.948948	0.982833	0.953099	0.89732	0.810756	0.302128	-0.436738
De Bilt - Maastricht	0.977815	0.990724	0.977196	0.951936	0.914735	0.801646	0.483294
Groningen - Maastricht	0.909520	0.968878	0.914933	0.787942	0.630983	-0.107660	-0.669359

Table 8: Mutual correlations between the three cities for which the rolling window consist of 1096 correlation coefficients. For each day a coefficient is calculated using two lists of 20 data points: daily temperature of the past 20 days of two cities.

5.3 Conclusion

The soil temperature differs between cities in the Netherlands. There are cities that for 5% of the days do not show significant correlation between the hourly temperature of the cities, so one cannot make statements about the relation between temperature and PD of a single day. However the centrally located De Bilt makes it possible to relate its temperatures of 20 consecutive days to circuits in other parts of the Netherlands. We will use this conclusion in Section 6.2 to determine the correlation between the soil temperature and the PD of the circuits.

6 Relation between partial discharges and temperature

In this chapter we investigate how to create features related to temperature that can be assigned to a PD cluster in the masterframe. This feature is intended to have a significant contribution to predict faults. This chapter uses the partial discharges and the temperature to create such a feature. Alliander knows that the cables and joints are influenced by the temperature. We are going to determine for each cluster the degree of influence of the temperature on PD. We do this by investigating to what extend they are correlated for each cluster.

There may be some clusters for which the temperature correlates well with PD, some for which there is a negative correlation and some for which no correlation can be found because, for example, there is insufficient data. These features are not predictive for the latter type of clusters. For the other clusters, the features could greatly improve the classification model to predict faults, especially when combined with other types of features, such as the circuit construction material.

In this Chapter we only use the soil temperature level 3, measured at a depth of 28-100cm, because the cables are located at a depth of 80-100cm. In Chapter 5 we have seen that we have to be careful how to use the data of the temperatures, because the temperature in the regions of the circuits differs slightly. Because we do not know where the clusters are, we use the measurements from De Bilt.

How can we create masterframe features that quantify the relation between the partial discharges and the temperature?

2019-05-23T14:13:00.000000000	295.00000
2019-05-23T14:17:00.000000000	138.00000
2019-05-23T14:31:00.000000000	261.50000
2019-05-23T14:35:00.000000000	227.50000
2019-05-23T14:40:00.000000000	418.50000

(a) Charge

2019-05-23T14:00:00.000000000	11.71000
2019-05-23T15:00:00.000000000	11.72000
2019-05-23T16:00:00.000000000	11.74000
2019-05-23T17:00:00.000000000	11.75000
2019-05-23T18:00:00.000000000	11.77000

(b) Temperature

Figure 31: Samples of the charge and temperature series of cluster 1 of circuit 4082

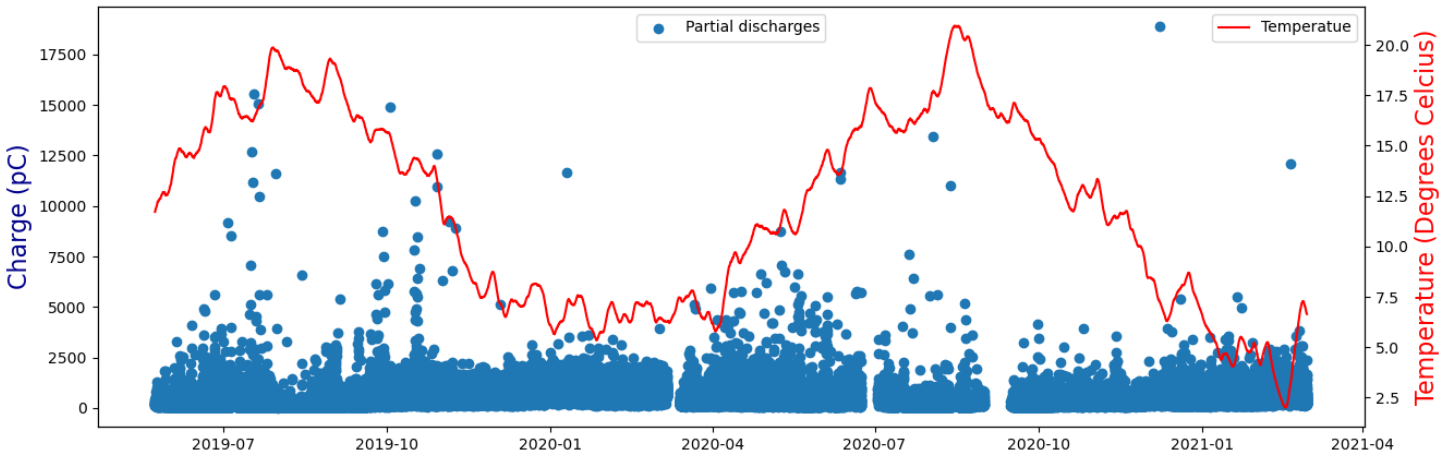


Figure 32: Measured partial discharges and corresponding temperature of cluster 1 of circuit 4082.

The partial discharges are all particles which form the cluster according to the cluster algorithm. For the temperature we use the soil temperature level 3 during the time period of the cluster. Only the soil temperature from 2018 is available. The particles are measured each minute, and the temperature is measured each hour. Figure 31 and 32 show the visualisation of the data of one cluster. We will correlate the temperature with the partial discharges for each cluster in Section 6.1. Section 6.2 creates feature which identify short periods of the cluster that correlate well with the temperature. Sections 6.3 and 5.4 improve these features by looking at the fluctuations of the temperature. Section 6.5 and 5.6 relate the seasons to the partial discharges.

6.1 Correlation coefficients

To calculate a correlation coefficient for the charge and the temperature, the two series need to be of equal length. We define variables for the charge and temperature per hour:

q_h^c : The sum of the charge of the measured PD of cluster c during hour h, in picocoulomb (pC);
t_h : The temperature during hour h, in degrees Celcius;
$q^c = [q_{h_1}^c, ..., q_{h_{max}}^c]$: The series of summed charges per hour in cluster c;
$t^c = [t_{h_1}, ..., t_{h_{max}}]$: The series of temperatures per hour in cluster c.

h_1 is the first hour of the cluster and h_{max} is the last hour of the cluster, such that q^c and t^c both have the length of the number of hours in cluster c. Note that the series t^c depends on cluster c, because it depends on the time period of the cluster. However the hourly temperatures t_h are independent of c, because for all clusters the same measurements of De Bilt are used. A cluster does not have holes: hours without PD are still included. So t^c is completely fixed by the first and last hour of cluster c. The superscripts of the charge and temperature series are omitted if it is clear which cluster is being referred to.

Figure 33 shows a sample of q: there are many hours without PD but these data points are included. t is the same as Figure 31(b), because the raw data of the temperature is already hourly. q and t are plotted in Figure 34. The corresponding correlation coefficient is Pearson$(q, t) = -0.52$. See Section 3.1 for more information about the Pearson correlation coefficient. We see a lot of short peaks in Figure 34. This is because the measured PD fluctuates greatly every hour, and there are many hours without PD. The overall trend of the PD is difficult to see. We would like to adjust the data so that the trend is clearer. Subsequently, the PD can be better compared with the temperature.

2019-05-23T14:00:00.000000000	1473.50000
2019-05-23T15:00:00.000000000	801.00000
2019-05-23T16:00:00.000000000	405.50000
2019-05-23T17:00:00.000000000	907.50000
2019-05-23T18:00:00.000000000	583.00000

Figure 33: Sample of the series of the hourly summed charge of cluster 1 of circuit 4082

How can we resample the data so that the relationship between PD and temperature is visible?

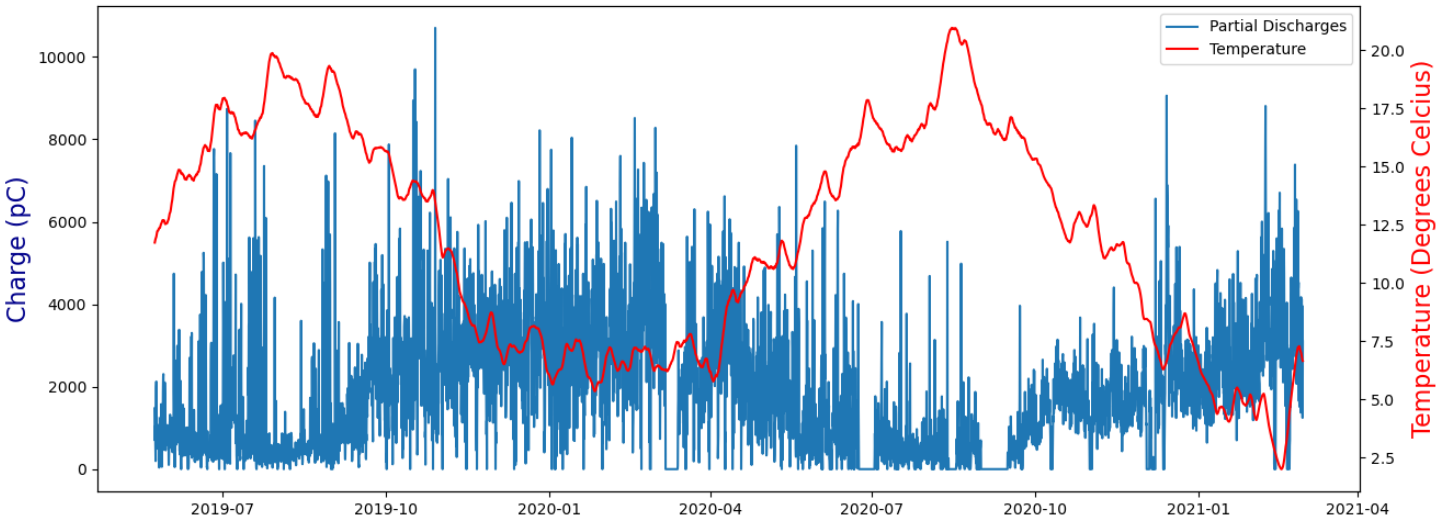


Figure 34: Cluster 1 of circuit 4082: The charge of each hour is summed. Correlation coefficient is -0.52.

6.1.1 Resampling

We can make the curve smoother by resampling the data. By **resampling** we mean bundling the data of each period of 10 days, instead of taking 1 data point each hour. Each period p_i of 10 days gets two data points:

$$q'_{p_i} = \frac{1}{240} \sum_{j=1+240i}^{240(i+1)} q_{h_j} : \text{the mean of the (summed per hour) charge of period } p_i;$$

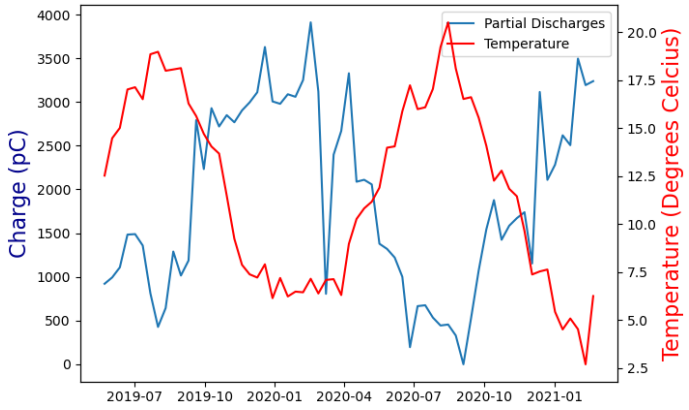
$$t'_{p_i} = \frac{1}{240} \sum_{j=1+240i}^{240(i+1)} t_{h_j} : \text{the mean of the temperature of period } p_i;$$

$$q' = [q'_{p_1}, ..., q'_{p_I}] : \text{The series of the resampled charges;}$$

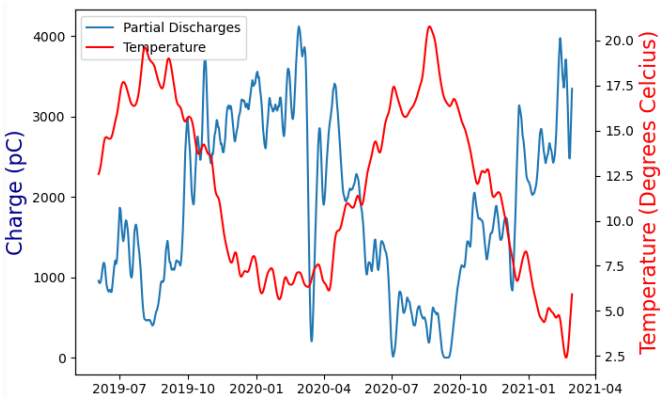
$$t' = [t'_{p_1}, ..., t'_{p_I}] : \text{The series of the resampled temperatures.}$$

A period of 10 days contains 240 hours. q_{h_j} is the sum of the charge measured in hour j, as defined at the previous page. I is the number of periods of 10 days of the relevant cluster.

We could have taken the sum of the hourly partial discharges instead of the mean. This makes no difference to the correlation, so we have chosen the mean because this makes the comparisons in the plots easier. Figure 35(a) shows q' and t' of cluster 1 of circuit 4082. The corresponding correlation coefficient is $\text{Pearson}(q', t') = -0.88$.



(a) **Resampling**: For each period of 10 days the mean of that period is taken, for both the temperature and the charge. This gives a correlation of -0.79.



(b) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. This gives a correlation of -0.78.

Figure 35: Modifications to both the temperature and the measured partial discharges of cluster 1 of circuit 4082

6.1.2 Smoothing

It was needed to take a period as large as 10 days to get a smooth curve for the partial discharges. The downside is that the number of data points shrinks very fast. To overcome this problem we introduce the method of **smoothing**: Taking the mean of the last 10 days for each hour. Then each hour has a value calculated by the 240 previous original numbers:

$$q''_{h_i} = \frac{1}{240} \sum_{j=i-239}^{i} q_{h_j} : \text{the average of the 240 hourly charge values prior to and including } h_i;$$

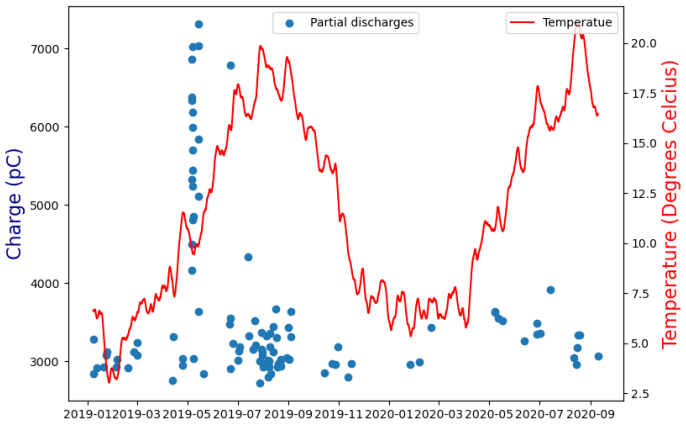
$$q'' = [q''_{h_{240}}, ..., q''_{h_{max}}] : \text{The series of the smoothed charges.}$$

q_{h_j} is only defined for $j \geq 1$, so q''_{h_i} is not defined for $1 \geq i \geq 239$. max is the number of hours in a cluster, so the series q'' contains 239 less data points than the series q. Smoothing is similar to resampling in some sense. In both techniques, we divide the data into smaller chunks and do some kind of aggregation. The difference is that each data point is used only once in resampling. When we use smoothing, each data point is used 240 times (except for the first ones). q'' and t in Figure 35(b) show the effect of smoothing the data of cluster 1 of circuit 4082. The

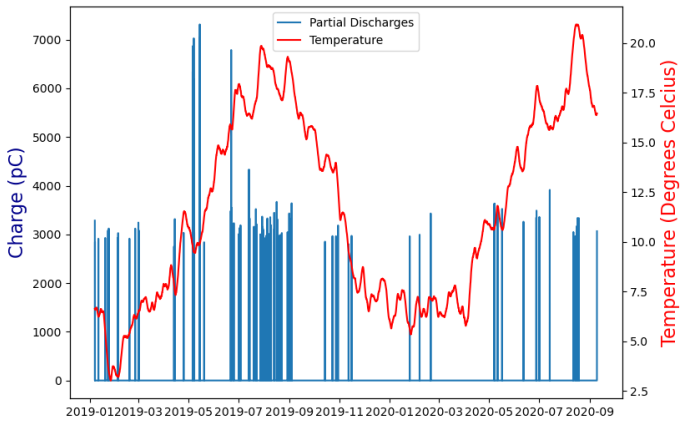
corresponding correlation coefficient is Pearson$(q'', t) = -0.78$. The relation is very clear: much PD in the winters and little PD in the summers. This does not mean that this would hold for all clusters so the next question comes up. Is this a good method to reflect the relation between partial discharges and temperature, for each cluster?

6.1.3 Examples

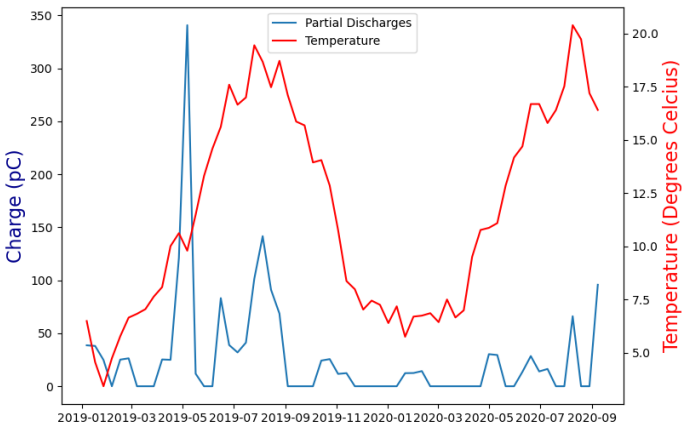
In Figure 36 we see the plots of cluster 8 of circuit 2719. This cluster lives in a period of 21 months: from January 2019 until September 2020. (a) shows the raw data. There seems to be more charge in the warmer periods, especially in the summer of 2019. We expect to see a positive correlation. However there is no correlation between the charge and temperature per hour in (b): the correlation coefficient is 0.03. The plots after using the methods resampling and smoothing are shown in (c) and (d) of Figure 36. These modified data series should represent the relation between the PD and temperature better. They do, albeit slightly: the coefficients are respectively 0.20 and 0.18.



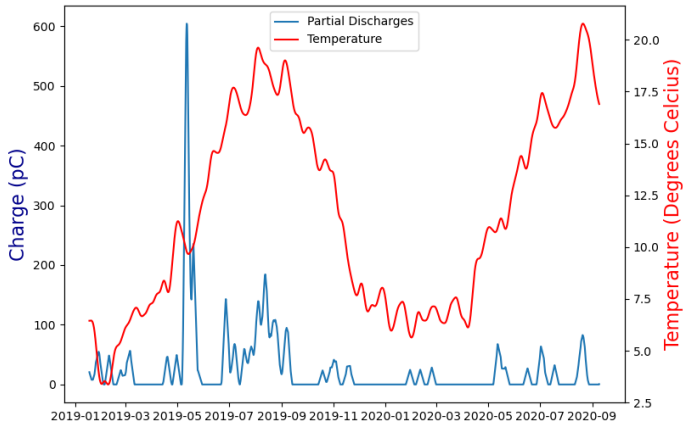
(a) Raw data



(b) For each hour the measured partial discharges are summed. Correlation coefficient is 0.03.



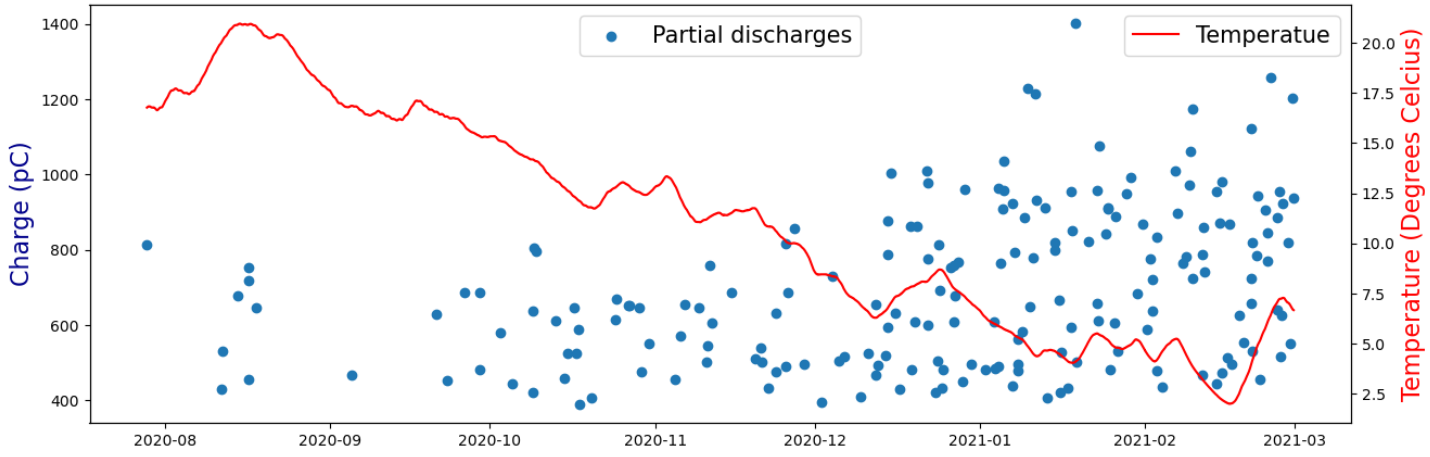
(c) **Resampling**: For each period of 10 days the mean of that period is taken, for both the temperature and the charge. Correlation coefficient is 0.20.



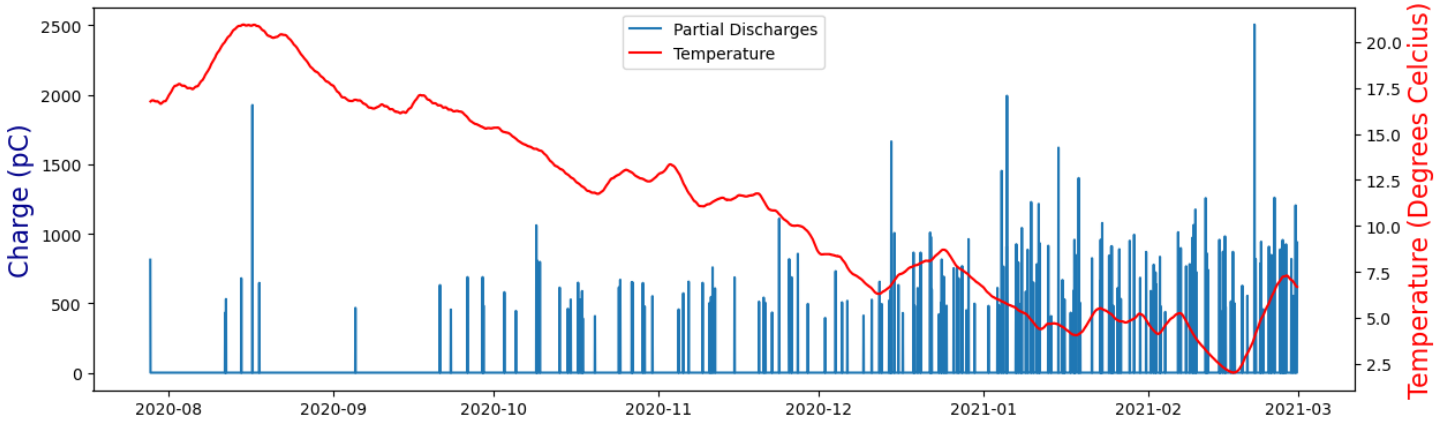
(d) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. Correlation coefficient is 0.18.

Figure 36: Four displays of cluster 8 of circuit 2719 and the soil temperature

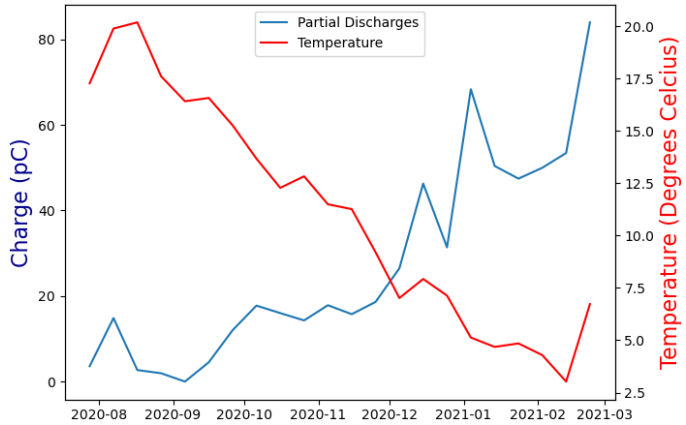
Figure 37 shows the the plots of another cluster of circuit 2719. In plot (a) we see that more PD is measured in early 2021 when the temperature is low. We therefore expect a negative correlation. The data series plotted in (b) confirm this with a coefficient of -0.22. Changing the data amplifies this effect with coefficients of -0.88 and -0.84.



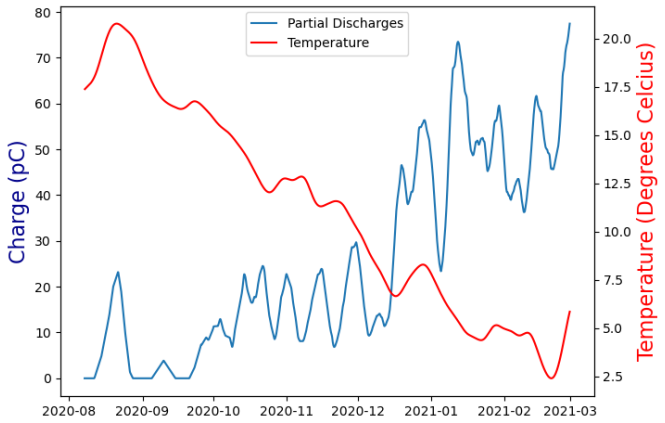
(a) Raw data



(b) For each hour the measured partial discharges are summed. Correlation coefficient is -0.22.



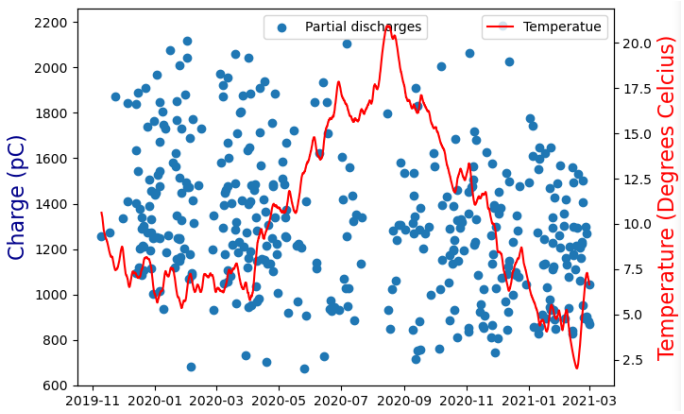
(c) **Resampling**: For each period of 10 days the mean of that period is taken, for both the temperature and the charge. This gives a correlation of -0.88.



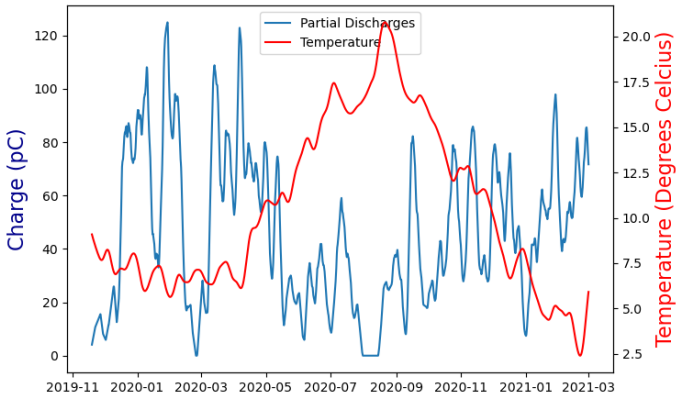
(d) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. This gives a correlation of -0.84.

Figure 37: Four displays of cluster 1 of circuit 2719 and the soil temperature

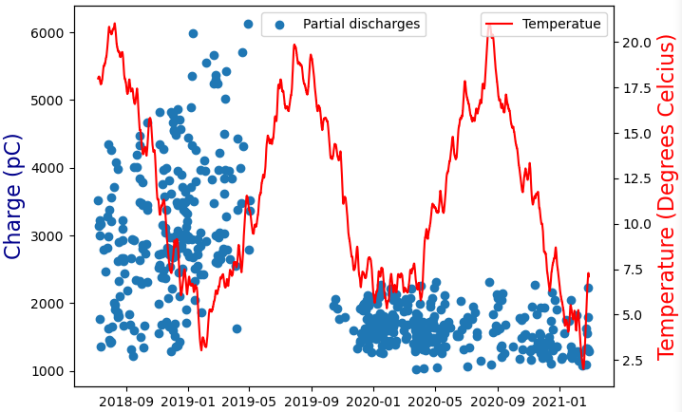
Figure 38 shows two clusters of circuit 2389. In (a) and (c) we see the raw data of respectively clusters 4 and 12. In the summer of 2019 there were no partial discharges in cluster 12 while cluster 4 has some discharges all the time of the cluster. However using the smoothed data in (b) and (d), cluster 4 shows a higher correlation coefficient. This raises the idea of exploring the possibility of quantifying the relationship between partial discharges and temperature by looking at the number of partial discharges per season of the year.



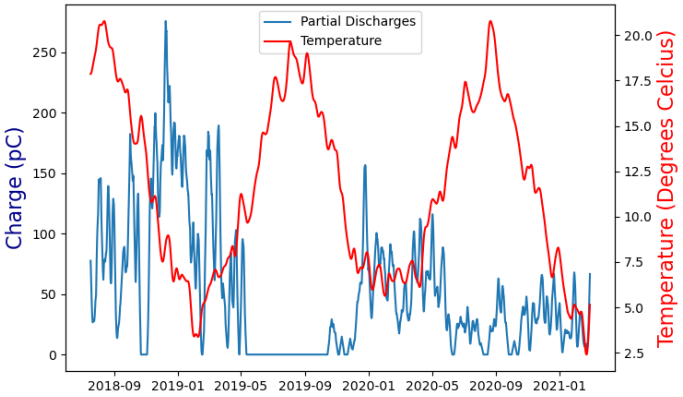
(a) Raw data of cluster 4



(b) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. Correlation coefficient is -0.49.



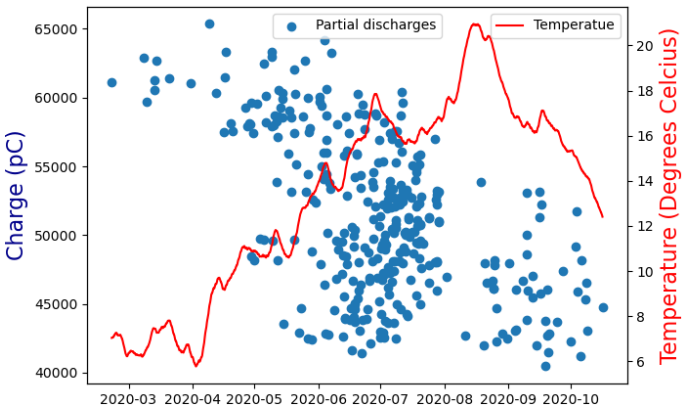
(c) Raw data of cluster 12



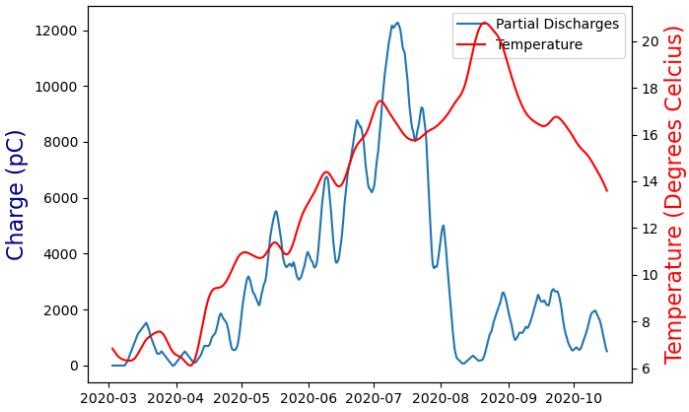
(d) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. Correlation coefficient is -0.32.

Figure 38: Raw and smoothed data of clusters 4 and 12 of circuit 2389 and the soil temperature

There are also many clusters for which the PD only correlates well over a certain period of time. Figure 39 shows an example of this phenomena. After smoothing the data, there is a huge correlation at the start of this cluster. The correlation coefficient for the period until June 2020 is 0.92, while the overall correlation coefficient is 0.21. This gives rise to the idea of investigating whether we can quantify a relationship between PD and temperature by looking at certain periods of the clusters.



(a) Raw data

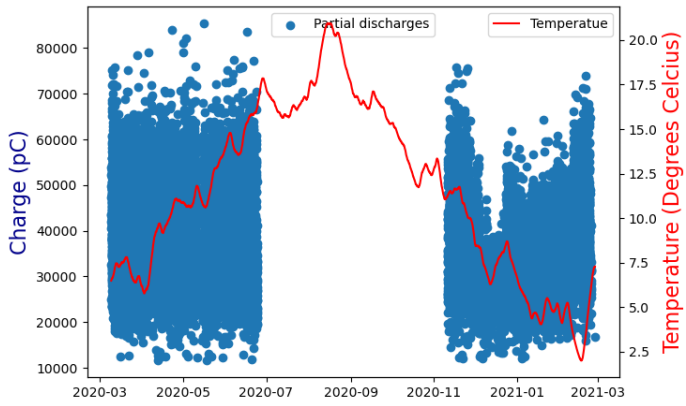


(b) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. This gives a correlation of 0.21.

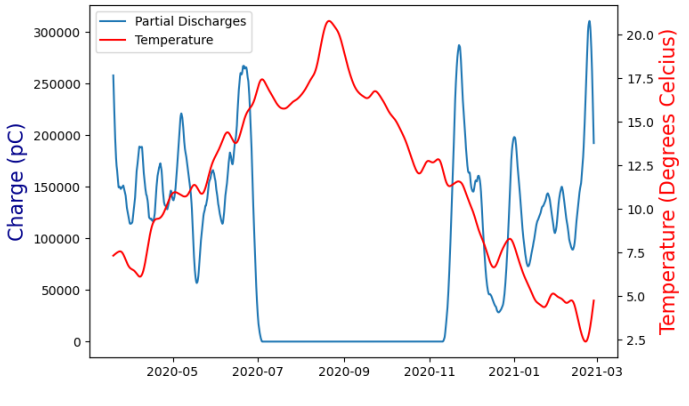
Figure 39: Raw and smoothed data of cluster 26 of circuit 2719 and the soil temperature

Figure 40 is another example of this phenomena. The overall correlation is -0.57 because there is no PD when the temperature is high. However if we zoom in on the first and last period of this cluster, we notice something else. In the period until July 2020 there is a correlation of 0.31 and in the period from November 2020 the correlation coefficient is 0.09. The reason we do not see PD between July and November 2020 is probably not that there did not occur PD. It seems like that there were no measurements during this period. This happens for example when there are maintenance activities on the circuit. We should keep in mind that the data contain gaps, and we could wonder if these features are appropriate for these clusters.

Although this example is not perfect it still shows that correlations of shorter periods could be different, which makes it interesting to explore the correlations of these short intervals.



(a) Raw data



(b) **Smoothing**: For each hour the mean of the previous 10 days is taken, for both the temperature and the charge. This gives a correlation of -0.57.

Figure 40: Raw and smoothed data of cluster 1 of circuit 20830 and the soil temperature

6.1.4 Masterframe feature

The feature *correlation_pd_temp* is added to the masterframe:

- *correlation_pd_temp*: Pearson(q'', t),
  Correlation between the soil temperature and the pd of the cluster.

6.1.5 Conclusion

To compute meaningful correlation coefficients between partial discharges and temperature, the timeseries have to be modified. We investigated several methods. The method of smoothing takes the mean of the temperature and PD of the last 10 days for each hour. This method seems like a good way to modify the data because the curve of the PD gets smoother while the number of data points does not shrink much. This way the correlation coefficient can be calculated well and expresses the relation between the PD and temperature well.

There are many clusters for which this method captures a characteristic of the cluster. So we include this feature to the masterframe. In Chapter 8 we will test the predictive power of this feature by using it in the classification model and see if it contributes to fault prediction.

There are two other interesting methods to quantify the relation between partial discharges and the temperature. In Section 6.2 we investigate the relation between the data in shorter periods of the cluster and in Section 6.5 we look at the relation between the partial discharges and the seasons.

6.2 Dataframe of correlation coefficients of shorter periods

In Section 6.1 the correlation between temperature and partial discharges of a cluster is calculated. The time period was equal to the period of the cluster. However there are clusters of which the partial discharges and temperature relate differently for small time periods. From the masterframe features created in Section 6.1, one can not extract clusters in which there is both a period in which the temperature and the partial discharges correlate very well and a period in which they have a high negative correlation coefficient. The overall coefficient does not show what is happening over time. This is why it could be interesting to look for shorter periods in the cluster in which there is a high positive or negative correlation.

How can we identify clusters where there is only a short period in which partial discharges strongly correlate with temperature?

6.2.1 Daily charge and temperature

We want to examine the effect of soil temperature on PD by looking at the correlation between the soil temperature and the charge. At night there is less PD because of less human activity, also the soil temperature is lower (although not much as we saw in Figure 27). These fluctuations during the nights do not help to find the effect of the temperature on PD. In the previous section we got rid of these fluctuations by taking the mean of 10 days after having hourly values of the temperature and the charge. Now we create daily values to exclude (to some extend) the regular discrepancies on the electricity network:

$$Q_d^c = \sum_{j=d_i}^{d_{24}} q_{h_j} : \text{The sum of the charge of the measured PD of cluster } c \text{ during day } d, \text{ in picocoulomb (pC);}$$

$$T_d = \frac{1}{24} \sum_{j=i}^{24} t_{h_j} : \text{The mean of the hourly temperatures during day } d, \text{ in degrees Celcius;}$$

$$Q^c = [Q_{d_1}^c, ..., Q_{d_{max}}^c] : \text{The series of summed charges per hour in cluster } c;$$

$$T^c = [T_{d_1}, ..., T_{d_{max}}] : \text{The series of temperatures per hour in cluster } c.$$

If $Q_d^c = 0$, there is no PD on day d and we consider Q_d^c as undefined. We use capitals Q and T instead of the q and t of the previous section because these are daily values instead of hourly values. Like in the previous section, the superscripts are omitted if it is clear which cluster is being referred to.

6.2.2 Method to create dataframe of correlation coefficients

To find periods of a cluster in which the correlation is high, we need to calculate the correlation coefficient of all periods. For each cluster we create a rolling window ρ^c consisting of correlation coefficients. For each day of the cluster the correlation of the previous 20 days is calculated:

$$\rho_{d_i}^c = \text{Pearson}([Q_{d_{i-19}}^c, ..., Q_{d_i}^c], [T_{di-19}, ..., T_{d_i}]) :$$
The correlation coefficient of the 20 daily values prior to and including day d_i of cluster c;

$$\rho^c = [\rho_{d_{20}}, ..., \rho_{d_{max}}] :$$
Rolling window of cluster c.

$\rho_{d_i}^c$ is considered as undefined if more than 10 of $\{Q_{d_{i-19}}^c, ..., Q_{d_i}^c\}$ are undefined. Then the correlation coefficient $\rho_{d_i}^c$ would be based on too few data which makes it unreliable. ρ_{d_i} is also not defined for $1 \geq i \geq 19$, because $Q_{d_i}^c$ is not defined for $i \geq 1$. max is the numbers of days in the cluster, so ρ^c contains 19 less data points than Q^c. From this rolling window ρ^c we can see if there are periods with a high correlation.

6.2.3 Example

In Figure 41 is cluster 26 of circuit 2719 again. The temperature and charge are smoother than Figure 39(a) and less smooth than Figure 39(b). The green curve is the rolling window ρ. So each of the three curves has one data point per day. There were some days without PD. These days cause the small holes in the blue curve.

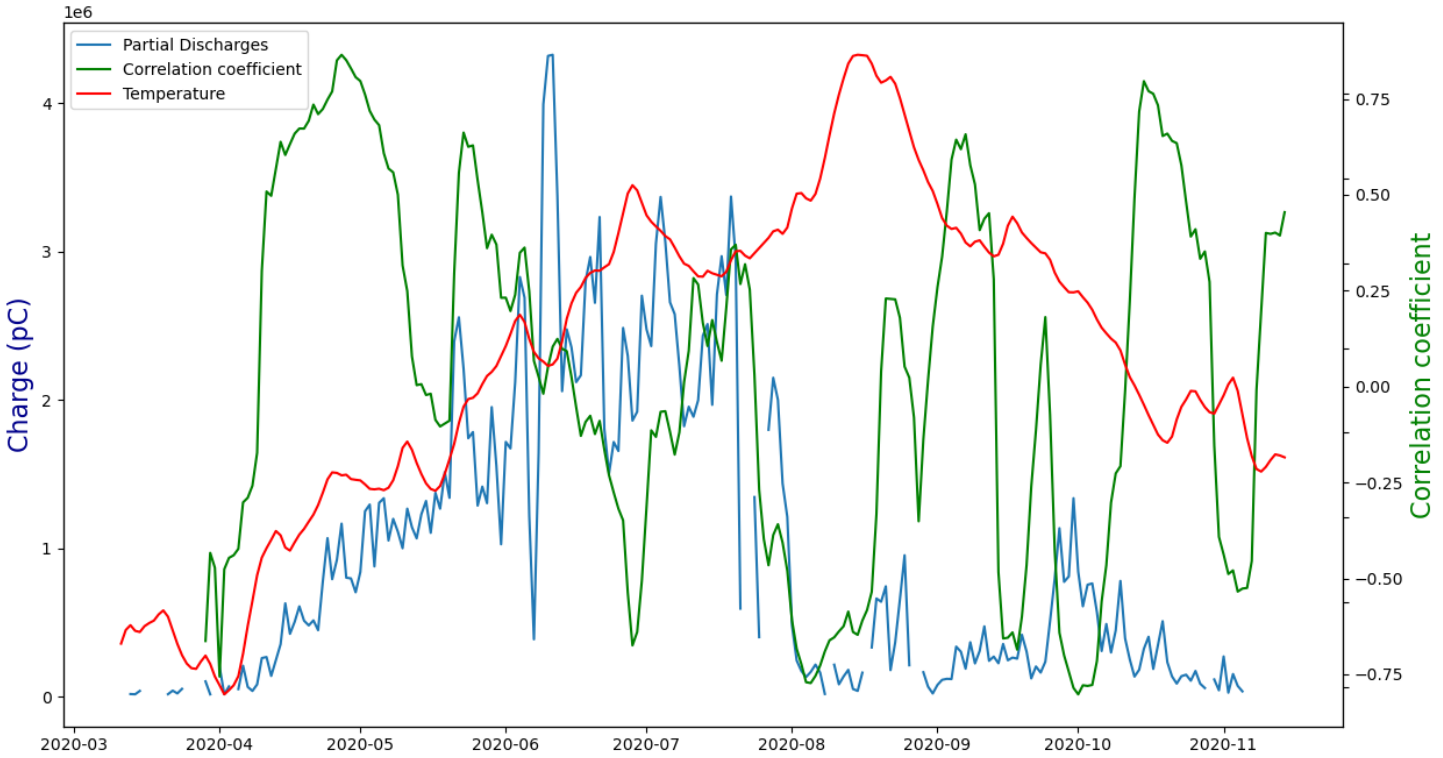


Figure 41: Rolling window consisting of correlations between the temperature and the partial discharges of cluster 26 of circuit 2719

The vertical axis for the temperature is left out for simplicity of the figure. Like Figure 39 the temperature is between 6 and 20 degrees Celsius. During April and May 2020 both the temperature and the charge rises. Therefor the correlation coefficient during this period is very high. During the end of July and August the charge drops while the temperature rises. Therefor the correlation coefficient is negative during this period.

6.2.4 Masterframe features

For the masterframe we need to make features about the cluster. The rolling window is used to extract features about the whole cluster. What is left to do is transforming the rolling window into a single number such that it can be included in the masterframe.

The mean and median of the correlation coefficients of the rolling window are no good indicators because they could be around 0 while there are both periods with high positive correlation and periods with high negative correlation. The maximum and minimum are better indicators. However it is possible that the correlation coefficient coincidentally takes very high values for some days. We see that it fluctuates very much in Figure 41. To make it more robust we introduce two features:

- *max_timeperiod_of_consecutive_positives*: $\max\{\lambda \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+\lambda-1}} \geq 0.7\}$,
  The longest period of the cluster in which the correlation coefficient for each day of this period is at least 0.7.

- *max_corr_which_repeats_12_timeperiods*: $\max\{\mathrm{P} \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+11}} \geq \mathrm{P}\}$,
  The maximum correlation coefficient for which there are at least 12 consecutive days for which the corresponding correlation coefficient is at least this value.

Index i	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Day d_i	d_{10}	d_{11}	d_{12}	d_{13}	d_{14}	d_{15}	d_{16}	d_{17}	d_{18}	d_{19}	d_{20}	d_{21}	d_{22}	d_{23}
Correlation coefficient ρ_{d_i}	0.9	0.3	0.9	0.9	0.6	0.6	0.7	0.7	0.6	0.8	0.8	0.8	0.7	0.6

Table 9: Rolling window of correlation coefficients of a fictitious cluster

Note that a correlation coefficient of a day is based on the past 20 days. See Table 9 for an example that clarifies the two features. This cluster consists of 23 days. Note that the first 9 days are not shown in the table because they do not have a corresponding correlation coefficient. (Their coefficient would be based on data of less than 10 days.) So we see the last 14 days of this cluster with their coefficients. $\rho_{d_i} \geq 0.7$ for $19 \leq i \leq 22$ and there is no longer period in which the minimum correlation is at least 0.7. So the value of the feature *max_timeperiod_of_consecutive_positives* is 4.

$\rho_{d_i} \geq 0.3$ for $10 \leq i \leq 21$, but there is another period of 12 days in which the minimum correlation is higher: $\rho_{d_i} \geq 0.6$ for $12 \leq i \leq 23$. So the maximum correlation which lasts at least 12 days is 0.6. So the value of the feature *max_corr_which_repeats_12_timeperiods* is 0.6.

A negative correlation could also be a good predictor so we also add the opposites to the masterframe:

- *max_timeperiod_of_consecutive_negatives*: $\max\{\lambda \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+\lambda-1}} \leq -0.7\}$,
  The longest period of the cluster in which the correlation coefficient for each day of this period is at most -0.7;

- *min_corr_which_repeats_12_timeperiods*: $\min\{\mathrm{P} \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+11}} \leq \mathrm{P}\}$,
  The minimum correlation coefficient for which there are at least 12 consecutive days for which the corresponding correlation coefficient is at most this value.

In Figure 42 we see a cluster for which the feature *max_timeperiod_of_consecutive_positives* gives 16. So there is a period of 16 days for which the correlation coefficient is at least 0.7. That is the period at the end of July 2019. The temperature drops first and rises after that. The charge does the same. So despite the few days for which there were no partial discharges, the correlation coefficient was high for quite a while. During this period the coefficient was above 0.84 for at least 10 days, which gave the feature *max_corr_which_repeats_12_timeperiods* the value of 0.84.

In Figure 43 is a cluster for which the feature *max_timeperiod_of_consecutive_negatives* gives 22. In July and August 2019 there is a long period for which the correlation coefficient is at most -0.7. This period includes 10 consecutive days for which the coefficient was -0.88 at most, so *max_corr_which_repeats_12_timeperiods* has value -0.88. There are partial discharges measured from August 2018 until September 2019, otherwise there would not be a blue line. However the charge seems insignificant low. The charge and the corresponding correlation coefficients in the period from September 2019 seem much more relevant, so one idea to improve the used method is to exclude the charge below a certain threshold.

Cluster 26 of circuit 2719 in Figure 41 is a good example where the features of this section related to the shorter periods of the cluster are much more expressive than the feature *correlation_pd_temp* of the last section. The overall correlation of this cluster is 0.21 while
max_timeperiod_of_consecutive_positives $= 13$; *max_timeperiod_of_consecutive_negatives* $= 8$;
max_corr_which_repeats_12_timeperiods $= 0.71$; *min_corr_which_repeats_12_timeperiods* $= -0.61$.
These features well express the period of April and May 2020 in which both the temperature and charge rise, and they also express that there is a significant period where the PD behaves opposite to the temperature.

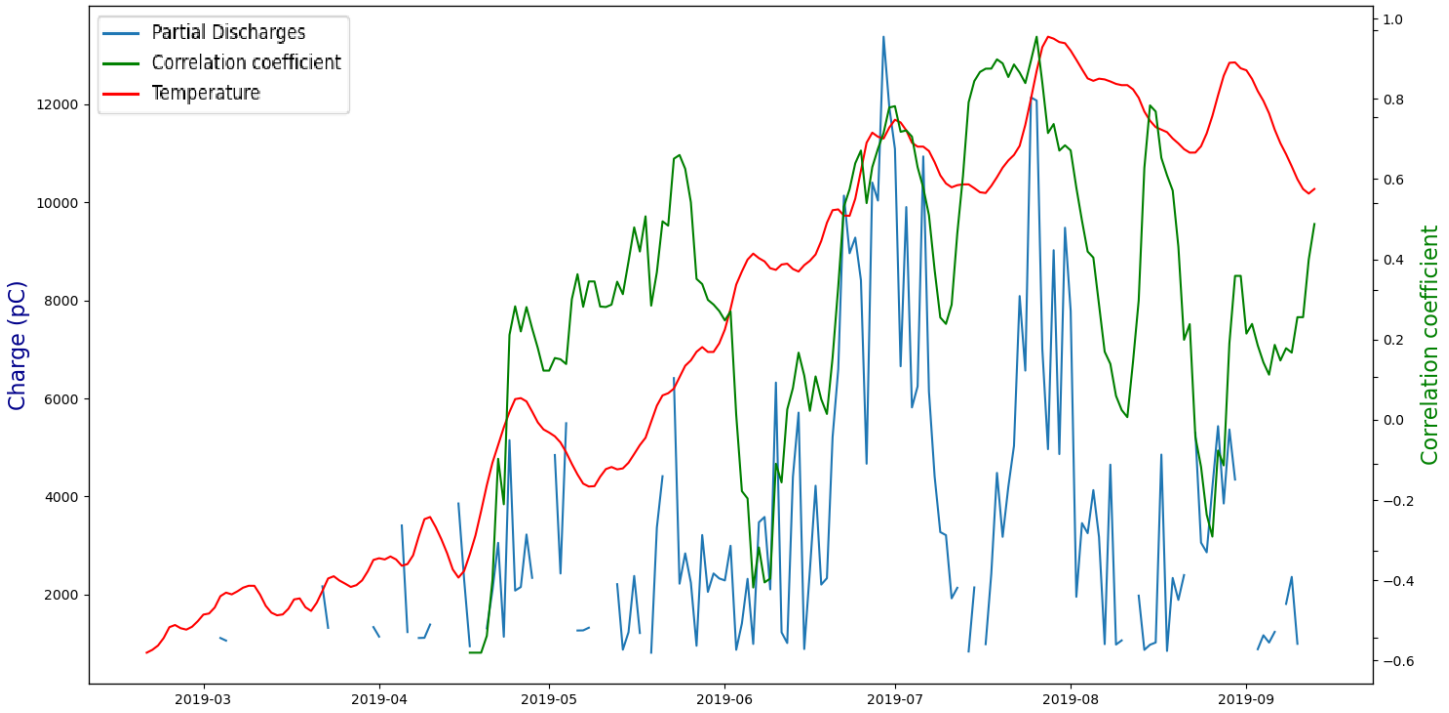


Figure 42: Rolling window of correlations between the temperature and the partial discharges of cluster 98 of circuit 1883: *max_timeperiod_of_consecutive_positives* = 16; *max_corr_which_repeats_12_timeperiods* = 0.84.

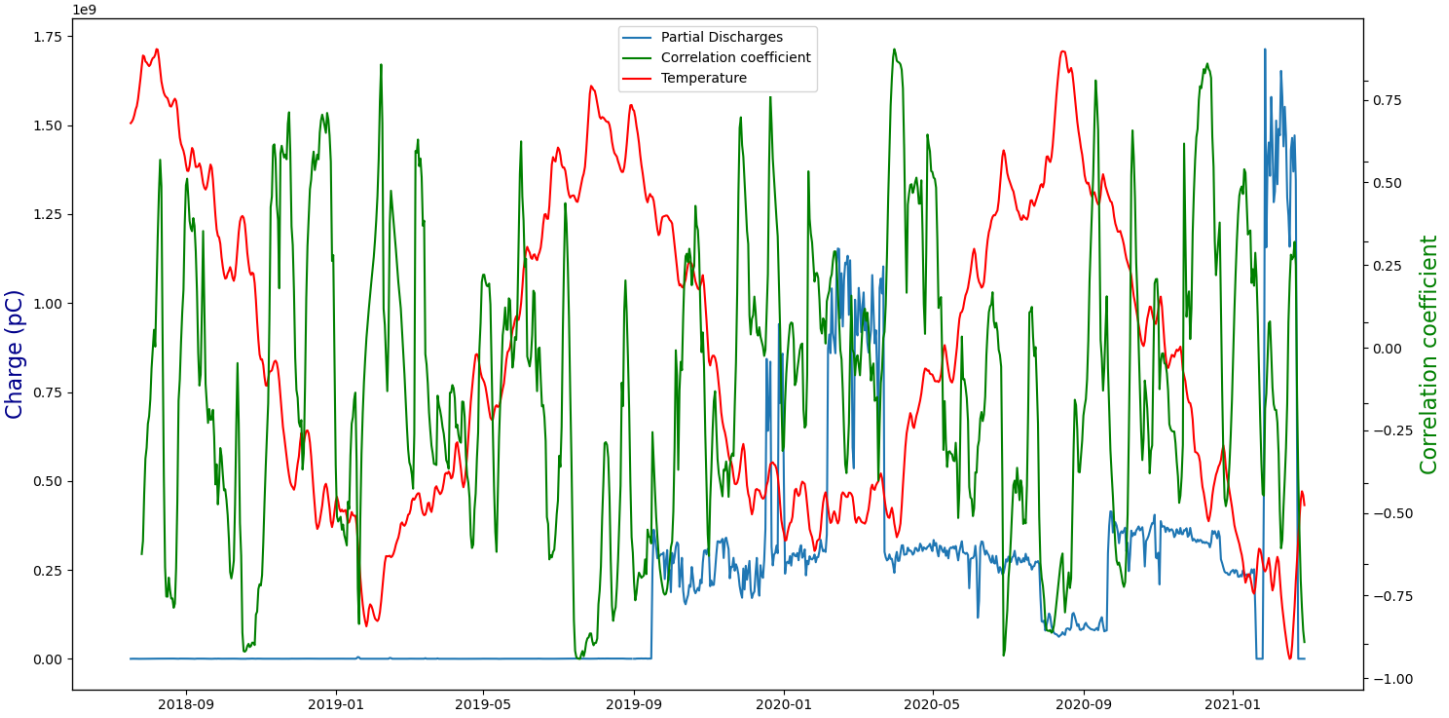


Figure 43: Rolling window of correlations between the temperature and the partial discharges of cluster 8 of circuit 1883: *max_timeperiod_of_consecutive_negatives* = 22; *max_corr_which_repeats_12_timeperiods* = −0.88.

6.2.5 Choice of the parameters

Time period of 20 days
For each day the correlation of the past 20 days is calculated. If we would choose a bigger period, we could miss small periods of for example 7 days in which there's a huge correlation. It could be flattened by the other days of the period. If we choose the periods to be much smaller than 20 days, then the correlation is not reliable anymore because of the uncertainty of the temperature. Remember that we use the temperature of De Bilt while we do not know where the circuits are.

Minimum 10 days of PD to calculate a correlation coefficient
The chosen length of the period of 20 days can not be too close to the chosen minimum number of days of measured PD, which is 10. If these are too close, there would be many days in which the correlation would not be calculated and consequently many empty cells in the rolling window. So the minimal number of days in which there is PD, can not be too large. But it can not be too small either because then the correlation could be based on this few numbers which makes the correlation unreliable.

One data point each 24 hours
Sum of PD and mean of temperature of 24 hours is chosen because we want to examine the direct effect of temperature on PD. To illustrate, at night it's colder than during the day so if we see less PD there is a high correlation but maybe there is little PD because people use less electricity at night, instead of the lower temperature. To exclude (to some extend) the regular discrepancies on the electricity network, we will be looking for a daily relation.

Correlation coefficient lasts for at least 12 days
To explain why we choose 12 days in which a coefficient has to last for both the feature *max_corr_which_repeats_12_timeperiods* and its negative equivalent, imagine a period of 30 days for which there is PD measured only during day 11 until day 20. Then the correlation which is calculated for day 20 until day 29 is only based on days 11 until 20 because the other days have no PD value. So if we would have chosen for 10 days for which the coefficient would last, this could be based on only 10 consecutive days. By choosing the period to be 12 days, the value *max_corr_which_repeats_12_timeperiods* could be based on only 12 values but these 12 values can not be exactly the same. This way we exclude the cluster value can not be based on only one coefficient, which makes it more robust.

Correlation coefficient at least 0.7
For the feature *max_timeperiod_of_consecutive_positives*, we choose the threshold 0.7. In a period in which all days show this coefficient, one can clearly see the relation in the plots. If the threshold is much higher, we would exclude periods in which there is a clear relation but its correlation coefficient is just not sufficient to be detected by this feature. If the threshold is too low, the feature would become less valuable because the variance of the coefficients, what the feature is build with, could be very high.

6.2.6 Conclusion

We created the features *max_timeperiod_of_consecutive_positives*, *max_corr_which_repeats_12_timeperiods*, *max_timeperiod_of_consecutive_negatives* and *min_corr_which_repeats_12_timeperiods*. These four features mark short periods in which the PD correlates very well with temperature. The examples showed that the features take on high values for clusters in which a clear correlation is visible in the graphs. The clusters for which the features take high values can be easily distinguished from the clusters without a period of strong correlation. The next section investigates the four features to get more insight into them.

6.2.7 Discussion

There are several ways to investigate if it is possible to improve the construction of the features in this section:

- The five parameters used for the features are based on common sense and could be tweaked.
- Clusters that have a longer existence have a higher chance to have a period in which the temperature correlates well with the PD. An idea to improve the method used in this section is to normalize the lengths of the clusters.
- Instead of taking the sum of the PD each day it could also be interesting to take the number of times PD is measured per day.

6.3 Relevance of the features about correlation of shorter periods

Four masterframe features are created in the last section:
max_timeperiod_of_consecutive_positives, *max_corr_which_repeats_12_timeperiods*,
max_timeperiod_of_consecutive_negatives and *min_corr_which_repeats_12_timeperiods*.
These features should represent to what extend there is a correlation between partial discharges and temperature in shorter periods. This section investigates them further to get more insight in the reliability of them.

To what extent are the values of the 4 features of Section 6.2 based on coincidence?

6.3.1 Heatmap

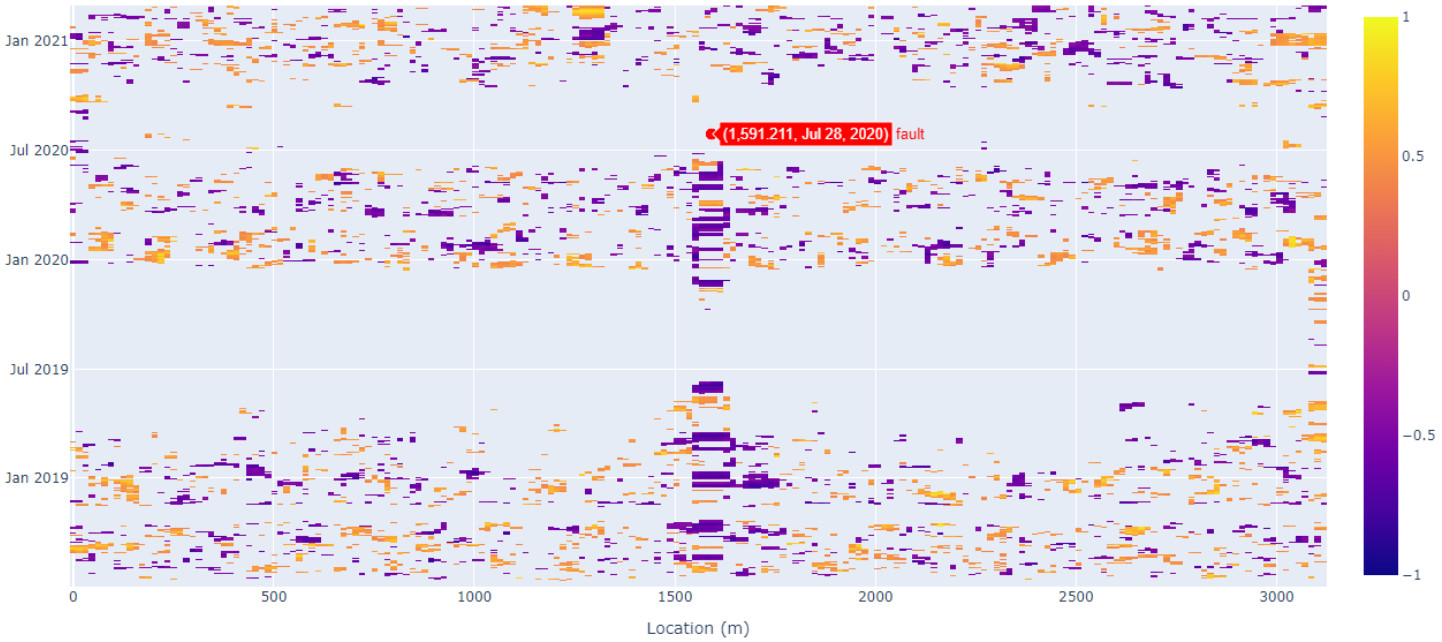


Figure 44: Heatmap of the rolling windows of correlation coefficients of several locations between the temperature and particles of circuit 2389. Correlations between -0.4 and 0.4 are discarded. A fault occurred 2020-07-28 at 1591 m.

Figure 44 shows a so called heatmap. For 200 equally distributed locations of circuit 2389 series of correlations are created. The color in the figure shows the level of the correlation. All coefficients between -0.4 and 0.4 are omitted to get a clearer view. There are many purple spots at approximately 1600m from the end of 2018 until the summer of 2020. So at this location there are many periods in which the partial discharges correlate well with the temperature. Right after this string of purple spots is a detection of a fault at 2020-07-28 at 1591 meters. This indicates that the correlation coefficients calculated for the 4 features are relevant. However it makes sense to see more spots before a fault because in general more partial discharges are detected before a fault, and this is only one example.

6.3.2 Substitute datasets

To investigate if many numbers of the features are based on coincidence, we create two artificial datasets which substitute the data of the temperature. To see if it is purely coincidental, that a cluster has periods for which the temperature correlates well with the partial discharges according to the four features, we create a random set of 'temperatures'. This set is built from the initial temperature data. The only difference is that the dates are randomly shuffled.

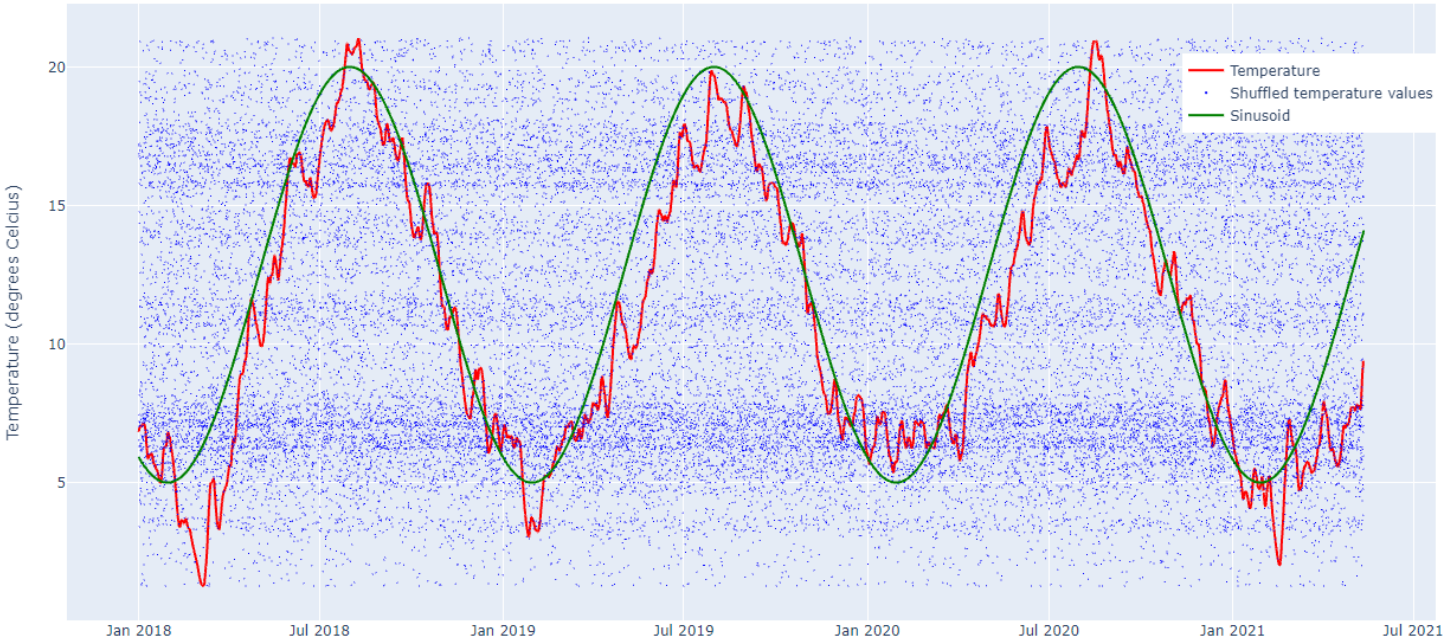


Figure 45: Soil temperature and two synthetic timeseries from January 2018 to May 2021

To see what the influence of the seasons is to the coefficients a sinusoid is created. This sinusoid is meant to get a similar shape as the actual temperature, except for the fluctuations. For this experiment there is no need to use a 'best' formula for the sinusoid to get an idea of the values of the four features. The formula of the siusoid we use is

$$12.5 + 7.5 * \frac{2\pi \sin(x)}{365 * 24},$$

where x is an integer based on the time of the actual temperatures: the temperature of 2018-01-01 00:00:00 gets the integer 0, the next value (2018-01-01 01:00:00) gets 1, and so on.

Both the shuffled set and the sinusoid are plotted next to the original temperature in Figure 45. A random 100 clusters are used to test the influence of the two substitutes, in comparison to the actual temperature. For each substitute the maximum, minimum, absolute value of the mean and absolute value of the median are calculated for all 100 clusters. The means of these 100 numbers are shown in Table 10. Also the mean of the four features is calculated and shown in Table 11.

↓ Data / Features →	Maximum	Minimum	|Mean|	|Median|
Temperature	0.74	-0.71	0.06	0.07
Shuffled temperature values	0.62	-0.62	0.04	0.05
Sinusoid	0.75	-0.75	0.04	0.05

Table 10: Four properties of the rolling window and their means over 100 clusters using the temperature and 2 substitute datasets

↓ Data / Features →	*max_timeperiod_of consecutive_positives*	*max_timeperiod_of consecutive_negatives*	*max_corr_which_repeats 12_timeperiods*	*min_corr_which_repeats 12_timeperiods*
Temperature	2.99	2.15	0.41	-0.37
Shuffled temperature values	0.61	0.63	0.35	-0.35
Sinusoid	2.90	2.67	0.38	-0.35

Table 11: Four properties of the rolling window and their means over the 100 clusters using the temperature and 2 substitute datasets

The mean of all clusters of both the maximum and minimum correlation coefficients per cluster drops a bit if the shuffled temperature values are used, as can be seen in Table 10. However the means of the maxima and minima are still significant high: many clusters have a period of 20 days in which the partial discharges correlate well with the temperature. This suggests that these values are not relevant, as we already expected in Section 6.2 Also the means of the absolute values of the means and the medians are not relevant

Table 11 shows the means of the four masterframe features, created in Section 6.2. The means of the first two features are significantly lower when the temperature values are shuffled, meaning that there are not many long periods in which the correlation coefficients are at least 0.7. This indicates that it is no coincidence that there are periods in which the partial discharges correlate well with the actual temperature. Looking at the last 2 features however, there is not much of a difference. The numbers of the shuffled temperatures are a bit lower though. Maybe 12 days is too long to see an actual difference. This suggests that a shorter period is better to remove some of the coincidence.

The numbers of the sinusoid are very similar to the actual temperature, so it seems that the four features are heavily influenced on the seasonal effects to the temperature, instead of just the fluctuations.

6.3.3 Conclusion

The four masterframe features (*max_timeperiod_of_consecutive_positives*, *max_corr_which_repeats_12_timeperiods*, *max_timeperiod_of_consecutive_negatives* and *min_corr_which_repeats_12_timeperiods*) are based on the rolling window of correlation coefficients of the relevant cluster of partial discharges. The rolling window of the correlations has a predictive value, as we saw in the heatmap. The four masterframe features created out of the rolling window are investigated by comparing them to the features that result from replacing the data of the temperature to first random numbers and second a sinusoid.

We can conclude that the first two features are not much based on coincidence, because the features of the random numbers are very different. We can not conclude this about the other two features. The difference between the features when using the temperature and using the random numbers is very small. However this small difference could be caused by very few clusters in which there are periods of 12 days in which the correlation coefficients are high. So maybe these two features are very useful to detect a few exceptional clusters. This could be investigated further.

The features when using the sinusoid suggest that the features of the temperature are strongly influenced by the seasons, rather than just the fluctuations. Ideally we would split the temperature into a trend (seasonal effect) and the residue (fluctuations). Section 6.4 looks at the residual values of the temperature while Section 6.5 and 6.6 are about the seasons.

6.4 Residue of temperature

We have seen that the seasonal effects have a huge influence on the 4 features we created in Section 6.2. To see what the correlation is between the fluctuations and the partial discharges, we somehow have to get rid of the seasonal effects on the temperature. The seasonal effects are approximately the same each year. It is a trend. The fluctuations of the temperature is the residue: difference between the trend and the actual temperature.

How can we capture the trend of the temperature?

6.4.1 Catching the trend by the best sinusoid

The trend can be described by a sinusoid. A sinusoid is defined by the parameters mean, amplitude, phase and frequency:

$$mean + amplitude * \frac{2\pi \sin(x - phase)}{frequency}.$$

The sinusoid which describes the trend the best is defined by the parameters for which the sum of the squares of the differences is minimal. The differences between the 29455 data points of the temperature and the sinusoid are used.

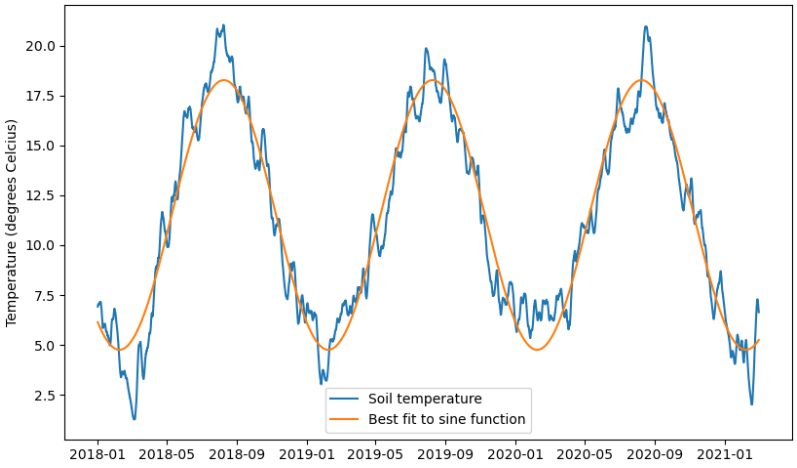


Figure 46: Siniusoid that best describes the trend of the temperature in De Bilt btween January 2018 and February 2021

The frequency is $365 * 24$ because we have a data point for each hour. The best phase happens to be 2018-05-08 04:00:00, the starting point of the sinusoid. The best mean and amplitude happen to be 11.41 and 6.79 respectively. The resulting sinusoid is plotted in Figure 46. The initial sinusoid of Section 6.3 is quite close to this one because there the phase, mean and amplitude are 2018-05-01 00:00:00, 12.5 and 7.5 respectively.

Looking at the figure it seems that the sinusoid does not have the right shape to represent the winters and summers well. The difference between the actual temperature and the sinusoid is quite large at the peaks. So it is worth trying to find a better way to show the trend of the temperature.

6.4.2 Using data from 1981 to determine the trend

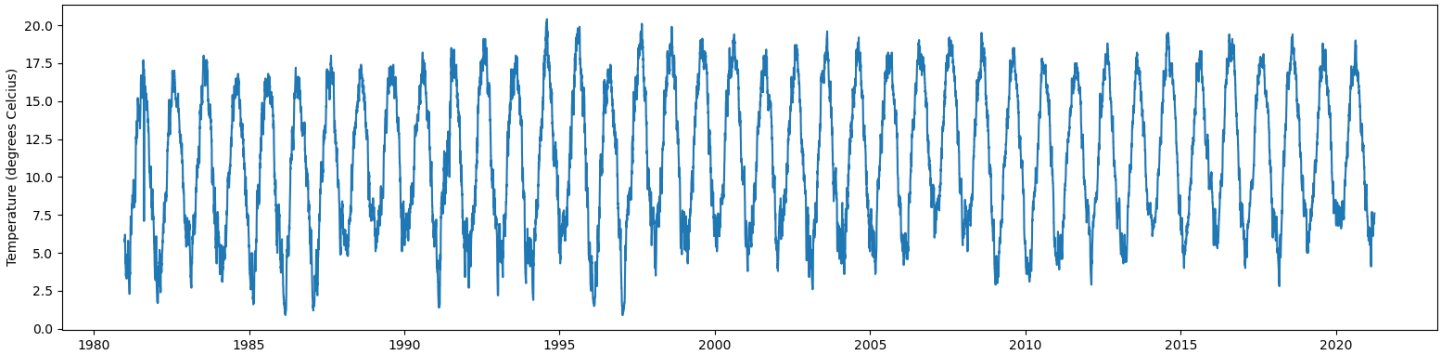


Figure 47: Soil temperatures measured each 6 hours at a depth of 50cm from 1981

The fluctuations are different each year but the trend of the seasons is the same each year. So the trend can be determined by taking the mean of previous years. To get robust means we need the temperature values of many years. The soil temperature we previously used is only available from January 2018, unfortunately. However there is more data of the soil temperature available [21]. These temperatures are measured at a depth of 50 cm every 6

hours from 1981. This data is way better for this experiment than the previously used temperatures measured at 28-100cm every hour from 2018. The soil temperatures $t^y_{p_i}$ we use in this section are plotted in Figure 47.

$t^y_{p_i}$: The temperature of period p_i in year y;
$p_1 =$ 01-01 00:00:00, $p_2 =$ 01-01 06:00:00, $p_3 =$ 01-01 12:00:00, etc.
$t^{2018}_{p_{1460}}$ is the temperature at 2018-12-31 18:00:00.

Note that 2020 is a leap year which complicates the matter. We define the temperature of 2020 in such a way that $t^{2020}_{p_i}$ and $t^{2018}_{p_i}$ are about the same date of the year for each $i \in \{1, ..., 1460\}$. The remaining four temperatures of 2020 are defined as follows:

$t^{2020}_{p_{1461}}$: The temperature of 2020-02-29 00:00:00;
$t^{2020}_{p_{1462}}$: The temperature of 2020-02-29 06:00:00;
$t^{2020}_{p_{1463}}$: The temperature of 2020-02-29 12:00:00;
$t^{2020}_{p_{1464}}$: The temperature of 2020-02-29 18:00:00.

The temperatures of the periods of the leap days of the other leap years between 1981 and 2020 are defined similarly. The idea is to use all of this data to determine the trend by taking the mean of each year. The temperature is measured each 6 hours so for each period of 6 hours between January 2018 and February 2021 the trend is calculated by taking the mean of the same period of 6 hours of all 40 years (1981-2020). For example the value of the trend which belongs to 2018-01-01 00:00:00 is the mean of the measurements of 1981-01-01 00:00:00, 1982-01-01 00:00:00, ..., 2020-01-01 00:00:00.

6.4.3 Compensating for global warming

There is one problem with this method, and that has to do with the global warming. It is not clear from Figure 47, but in Figure 48 are the means of the years and the linear regression, and this shows a rising trend.

To overcome the problem we compensate this rise by normalizing the values of each year. To determine the trend of 2018 we first take the average temperature of each year and calculate the difference with the mean of 2018. Then these differences are added to all values of the corresponding years. For the trend of 2019 and 2020, we do the same. The trends of these three years differ slightly because the means of these years differ slightly.

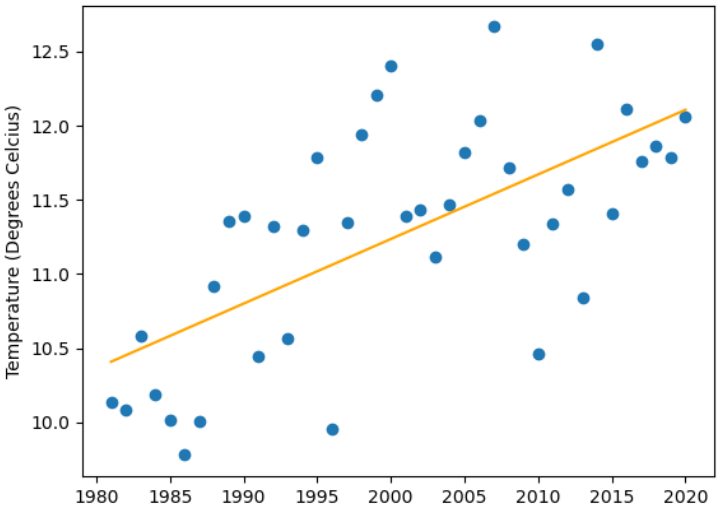


Figure 48: The means of the soil temperatures at a depth of 50cm and the linear regression. In general, the temperature rises over time.

$$M_{2018} = \frac{1}{1460}\sum_{j=1}^{1460} t^{2018}_{p_j} : \text{The mean of 2018;}$$

$$M_{2019} = \frac{1}{1460}\sum_{j=1}^{1460} t^{2019}_{p_j} : \text{The mean of 2019;}$$

$$M_{2020} = \frac{1}{1464}\sum_{j=1}^{1464} t^{2020}_{p_j} : \text{The mean of 2020.}$$

$\forall i \in \{1, ..., 1460\}\ \forall y \in \{1981, ..., 2020\} : u^y_{p_i} = t^y_{p_i} + M_{2018} - M_y,$ The compensated temperatures compared to 2018;
$\forall i \in \{1, ..., 1460\}\ \forall y \in \{1981, ..., 2020\} : v^y_{p_i} = t^y_{p_i} + M_{2019} - M_y,$ The compensated temperatures compared to 2019;
$\forall i \in \{1, ..., 1460\}\ \forall y \in \{1981, ..., 2020\} : w^y_{p_i} = t^y_{p_i} + M_{2020} - M_y,$ The compensated temperatures compared to 2020.

The four periods of the leap day of 2020 are treated separately:

$$\forall i \in \{1461, ..., 1464\}\ \forall y \in \{1981, ..., 2020\} \cup \{\text{leap years}\} : w_{p_i}^{y} = t_{p_i}^{y} + M_{2020} - M_y,$$

The compensated temperatures compared to 2020.

Note that $u_{p_i}^{2018} = t_{p_i}^{2018}$, $v_{p_i}^{2019} = t_{p_i}^{2019}$ and $w_{p_i}^{2020} = t_{p_i}^{2020}$.

Only after normalizing the measurements, we calculate the mean for each period of 6 hours. So for example to get the value of the trend for 2018-01-01 18:00:00 we first add the difference between the mean of whole 2018 and the mean of 1981 to each measurement of 1981, and the difference between the mean of whole 2018 and the mean of 1982 to each measurement of 1982, and so on. Afterwards the mean of all the resulting values of January first 18:00:00 is taken to get the value of the trend for 2018-01-01 18:00:00.

$$\forall i \in \{1, ..., 1460\} : U_{p_i} = \frac{1}{40} \sum_{j=0}^{39} u_{p_i}^{1981+j} : \text{The compensated temperatures of 2018;}$$

$$\forall i \in \{1, ..., 1460\} : V_{p_i} = \frac{1}{40} \sum_{j=0}^{39} v_{p_i}^{1981+j} : \text{The compensated temperatures of 2019;}$$

$$\forall i \in \{1, ..., 1460\} : W_{p_i} = \frac{1}{40} \sum_{j=0}^{39} w_{p_i}^{1981+j} : \text{The compensated temperatures of 2020;}$$

$$\forall i \in \{1461, ..., 1464\} : W_{p_i} = \frac{1}{40} \sum_{j=0}^{9} w_{p_i}^{1984+4j} : \text{The compensated temperatures of 2020-02-29.}$$

The union $\bigcup_{i \in \{1,...,1460\}} \{U_{p_i}, V_{p_i}, W_{p_i}\} \cup \{W_{p_{1461}}, W_{p_{1462}}, W_{p_{1463}}, W_{p_{1464}}\}$ forms the trend which represents 2018-2020, plotted in Figure 49. The trend of January and February 2021 can not be produced the same way so we take a copy of January and February 2020 for this. We see the same outliers as we have seen when fitting the best sinusoid: very high temperature in the summer and very low in the winter. We can conclude from this that we have had warmer summers and colder winters in the past three years. The residual values are the differences between the actual temperature and the trend.

$$\forall i \in \{1, ..., 1460\} : r_{p_i}^{2018} = t_{p_i}^{2018} - U_{p_i} : \text{ The residual values of 2018;}$$
$$\forall i \in \{1, ..., 1460\} : r_{p_i}^{2019} = t_{p_i}^{2019} - V_{p_i} : \text{ The residual values of 2019;}$$
$$\forall i \in \{1, ..., 1464\} : r_{p_i}^{2020} = t_{p_i}^{2020} - W_{p_i} : \text{ The residual values of 2020.}$$

The residue $\mathrm{r} = \bigcup_{i \in \{1,...,1460\}} \{R_{p_i}^{2018}, R_{p_i}^{2019}, R_{p_i}^{2020}\} \cup \{R_{p_{1461}}^{2020}, R_{p_{1462}}^{2020}, R_{p_{1463}}^{2020}, R_{p_{1464}}^{2020}\}$ is also plotted in Figure 49.

6.4.4 Masterframe features

Using the residue instead of the temperature we create four new cluster features for the masterframe. They are exactly the same as their equivalents in Section 6.2 except that the residue is used instead of the temperature.

- *max_timeperiod_of_consecutive_positives_residue*: $\max\{\lambda \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+\lambda-1}} \geq 0.7\}$;
- *max_timeperiod_of_consecutive_negatives_residue*: $\max\{\lambda \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+\lambda-1}} \leq -0.7\}$;
- *max_corr_which_repeats_12_timeperiods_residue*: $\max\{\mathrm{P} \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+11}} \geq \mathrm{P}\}$;
- *min_corr_which_repeats_12_timeperiods_residue*: $\min\{\mathrm{P} \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+11}} \leq \mathrm{P}\}$.

with

$$R_d = \frac{1}{4} \sum_{p \in d} r_p : \text{ The mean of the 6-hourly residues during day } d\text{, in degrees Celcius;}$$

$$\rho'_{d_i} = \mathrm{Pearson}([Q_{d_{i-19}}^{c}, ..., Q_{d_i}^{c}], [R_{di-19}, ..., R_{d_i}]) :$$
The correlation coefficient of the 20 daily values prior to and including day d_i of cluster c;

$$\rho' = [\rho'_{d_{20}}, ..., \rho'_{d_{max}}] :$$
Rolling window of cluster c.

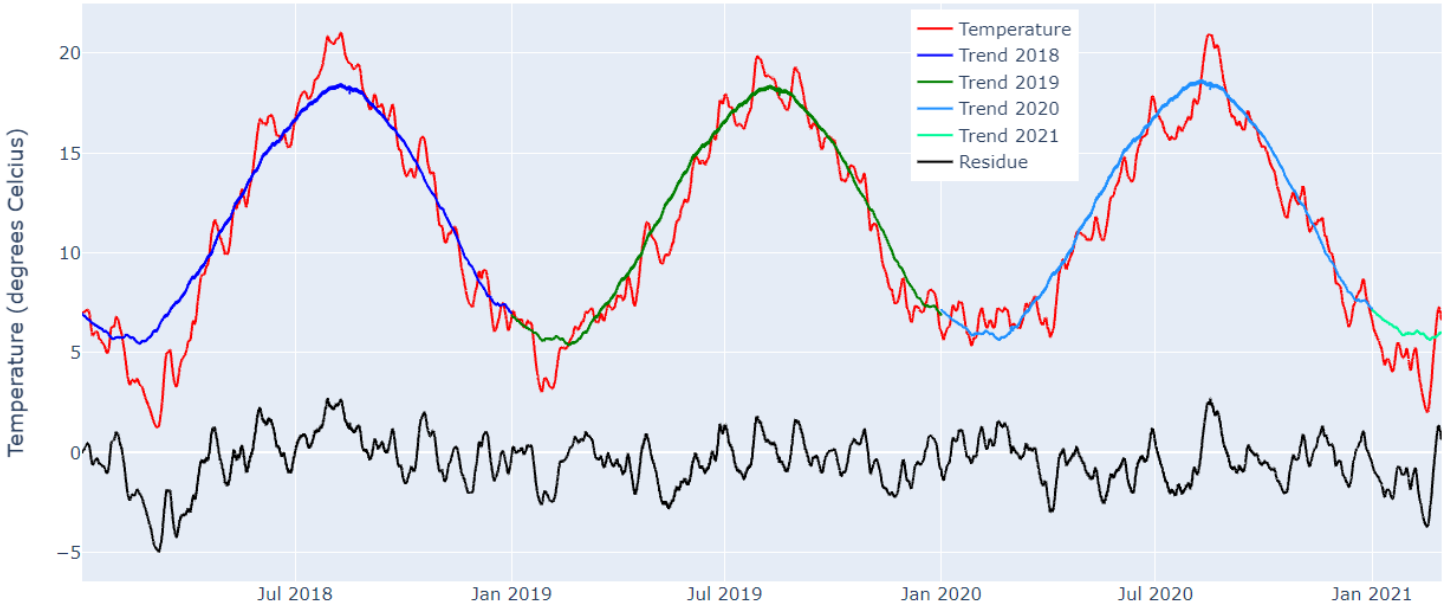


Figure 49: The trend of the temperature is determined by the data of 1981-2020. The residue is the difference between the actual temperature and the trend.

6.4.5 Example

In Figure 41 of Section 6.2.3 we saw the rolling window of cluster 26 of circuit 2719. Figure 50 shows the rolling window of the same cluster based on the residue instead of the temperature. The axis of the residue is left out for simplicity of the figure. The residue is between -3 and 3 between 2020-03 and 2020-11. The residue features are: *max_timeperiod_of_consecutive_positives* $= 0$; *max_timeperiod_of_consecutive_negatives* $= 7$; *max_corr_which_repeats_12_timeperiods* $= 0.31$; *min_corr_which_repeats_12_timeperiods* $= -0.68$.

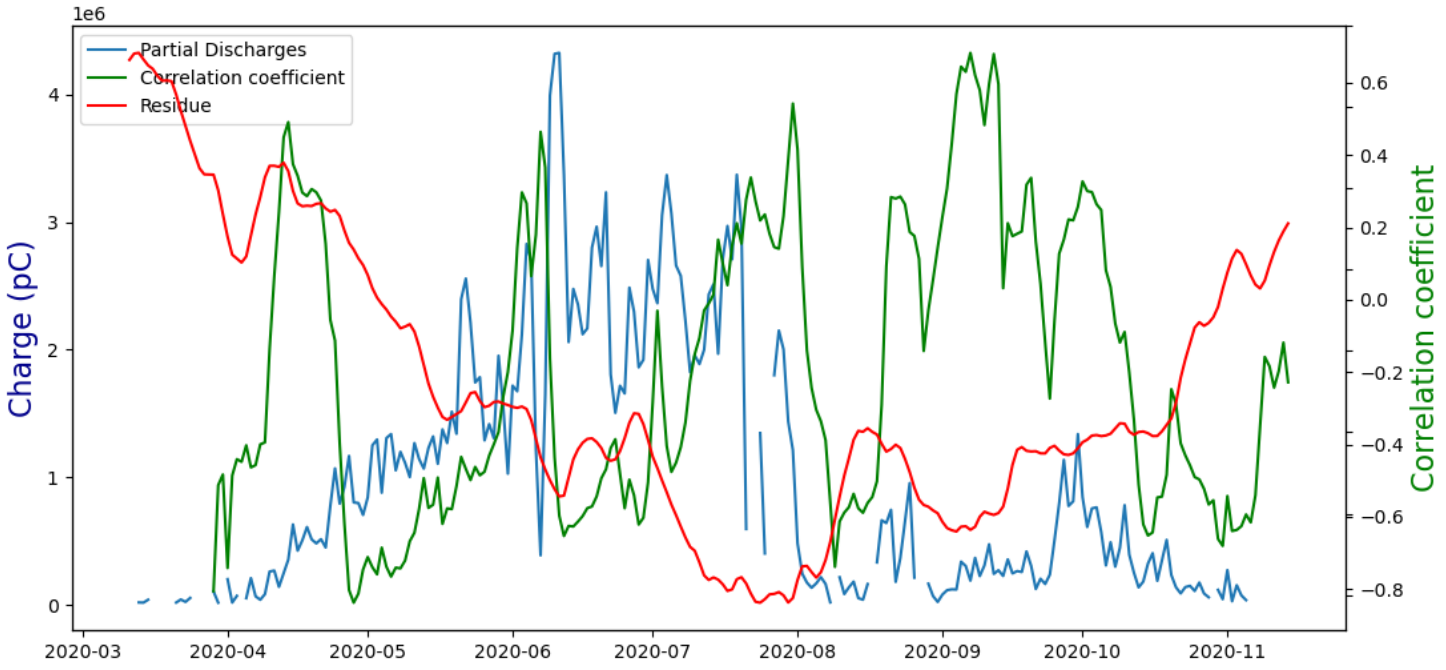


Figure 50: Rolling window of correlations between the **residue** and the partial discharges of cluster 26 of circuit 2719: *max_timeperiod_of_consecutive_positives* $= 0$; *max_timeperiod_of_consecutive_negatives* $= 7$; *max_corr_which_repeats_12_timeperiods* $= 0.31$; *min_corr_which_repeats_12_timeperiods* $= -0.68$.

6.4.6 Conclusion

Two methods are inspected to determine the trend of the temperature. The first uses an sinusoid to mimic the data and the second uses data from previous years to predict the trend. Although they both faced the problem of not representing the winters and summers very well, the second method is considered to be better because it is a more natural way to represent the trend. The trend resulting from the use of historical temperature data is used to get the residual. The residual values represent the fluctuations of the temperature and are therefore used to create 4 new cluster features that are added to the masterframe.

6.4.7 Discussion

The initial idea is to see if the temperature rises or drops and see what the effect is on the partial discharges. In general the residue represents this quite nicely. However there are periods in which the temperature drops very fast while the residue stays positive or vice versa. During these periods, the residue does not accurately reflect the behavior of the temperature. For example right after the peaks in the summer of the last three years, the temperature starts to drop while the residue stays above 0 for a short while. An idea to fix this issue is taking the derivative of the temperature instead, or looking at the difference between the temperature and the temperature one day before for example.

Another idea to determine the trend is by taking the mean of the temperature of the surrounding 20 days. For example for 2018-01-01 18:00:00 one takes the temperatures of 18:00:00 of the next 10 days and the previous 10 days. This still gives a rather smooth curve. This is because the data points of the trend of for example 2018-01-01 18:00:00 and 2018-01-02 18:00:00 are based on many of the same data points. This is a simpler method and it catches a peak if there would be a very hot summer one year. It also observes an effect if the temperature suddenly drops. The rolling window is based on the 20 previous days, so in this matter it could be better to let the trend also be based on the previous 20 days instead of the surrounding 20 days. The resulting residue is in this case the drop or rise in comparison with the same period for which the correlation coefficients are being calculated. In this sense, it seems much more logical to create the trend like this.

6.5 Partial discharges when the temperature is low

There have been a lot of faults at locations of joints of the type resin in August 2020. To search for clues to predict similar faults in the future we looked at many plots of the partial discharges and temperature at these locations. It was quite remarkable that there were a lot of plots of circuits with almost all partial discharges during the cold periods of the year. Figure 51 shows a nice example in which there is a lot of activity around January 2019 and January 2020 while there is almost no PD in the warmer periods of the year. In this section we will investigate the next hypothesis:

Partial discharges leading to faults in resin joints appear more often during the cold periods than during the hot periods.

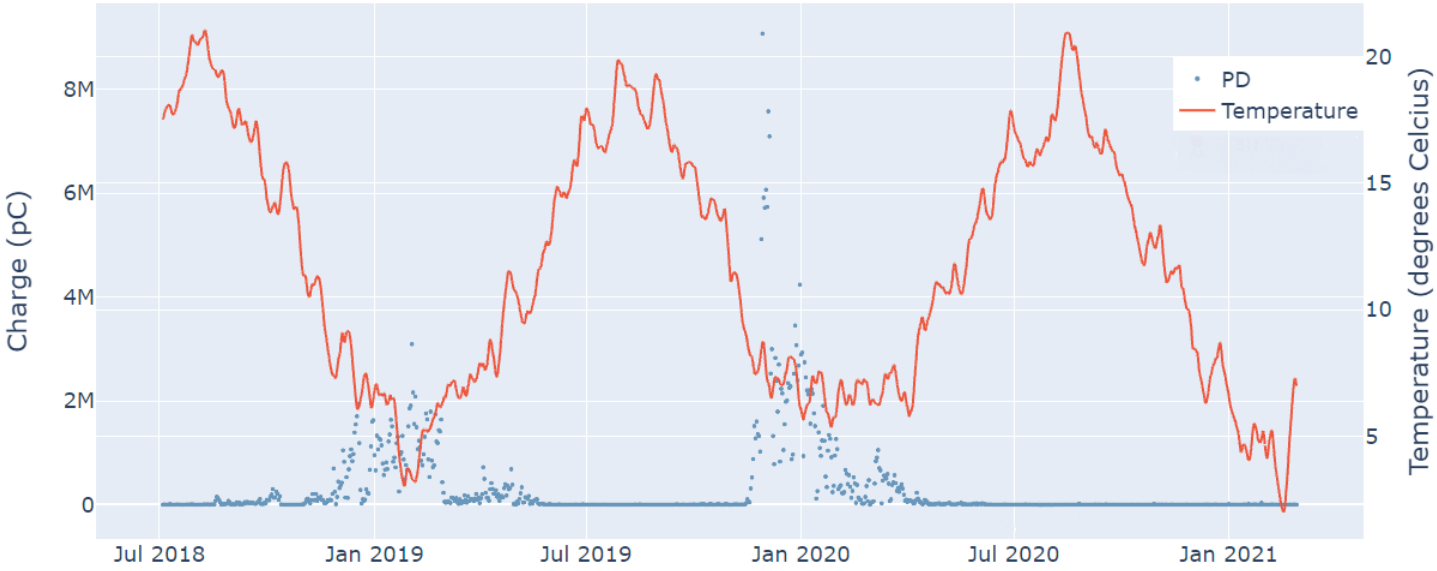


Figure 51: Partial discharges around 780 meters of circuit 2389 compared with the temperature

6.5.1 Method

The ground is coldest approximately between November and April. This is exactly half of the year, so we investigate whether there are more partial discharges in these 6 months than in the other months, for three different sets:

1. Clusters for which the nearest joint to the median location of the cluster is of the type resin.
2. Clusters for which the nearest joint to the median location of the cluster is of the type resin, and the cluster caused ≥ 1 fault according to Definition 6. The partial discharges after the first fault occurred are omitted.
3. Partial discharges measured at the locations (with a bandwith of 1% of the circuitlength) of the faults which occurred at a resin joint. The partial discharges after the fault occurred are omitted.

We only take the partial discharges into account which are measured in 2018, 2019 and 2020 to let each month contribute equally. For each of the sets we receive a dataframe like Figure 52. For each element of the set, the dataframe shows the total measured partial discharges, the discharges in the cold period, hot period and the percentages.

To interpret these numbers we calculate three percentages for each set:

(A) Percentage of elements with more charge in the cold period than in the warm period;

(B) Percentage of the sum of all measured charge in the cold period;

(C) Mean of the percentages of charge in the cold period per element.

pd_sum	pd_cold	pd_hot	pd_cold_percentage	pd_hot_percentage
24954393.50000	11934498.00000	13019895.50000	47.82524	52.17476
62822652.00000	34583504.00000	28239148.00000	55.04942	44.95058
24538629.50000	9681965.00000	14856664.50000	39.45601	60.54399
531076.50000	46160.50000	484916.00000	8.69187	91.30813
910781.00000	618086.50000	292694.50000	67.86335	32.13665
2158683.50000	1099459.50000	1059224.00000	50.93195	49.06805
448836.50000	240447.50000	208389.00000	53.57129	46.42871
15509388.00000	5646800.50000	9862587.50000	36.40892	63.59108
3245261.50000	2027536.00000	1217725.50000	62.47681	37.52319
1638016459.00000	657620068.00000	980396391.00000	40.14734	59.85266
9457825.00000	5436359.00000	4021466.00000	57.48001	42.51999
4987638.00000	2086442.00000	2901196.00000	41.83227	58.16773
95074029.50000	43338597.50000	51735432.00000	45.58405	54.41595
105325753.50000	52150476.00000	53175277.50000	49.51351	50.48649
3407180.00000	1714429.00000	1692751.00000	50.31812	49.68188
6232405.50000	6045217.00000	187188.50000	96.99653	3.00347
719825.50000	404697.00000	315128.50000	56.22154	43.77846

Figure 52: Part of the dataframe that contains the sum of the charge of the particles and the distribution across two halves of the year. Each row represents a cluster of PD for which the nearest joint is resin.

6.5.2 Results

Table 12 shows the percentages A, B, C for all three sets. The size of the sets differ a lot as can be seen in the table as well.

Set	Size of dataset	A	B	C
1. Clusters	747	50.7	48.7	49.4
2. Clusters with ≥ 1 faults	17	52.9	80.9	59.6
3. Resin joints at fault locations	63	52.4	44.6	48.4

Table 12: Percentages showing, according to methods A, B, C, for three sets, how much of the measured charge takes place in the coldest half of the year: November until April

Almost all percentages of Table 12 are around 50%. Only the percentage of the charge during the cold period (B) of clusters with ≥ 1 fault is an outlier. However this is based on just 17 clusters and there is one big cluster that contributes far more than the others. The A and C values of clusters with ≥ 1 fault are not that much influenced by this cluster because each cluster contributes equally for these methods.

6.5.3 Conclusion

We tested the hypothesis that most of the partial discharges leading to faults in resin joints appear during the cold half of the year. From the results we can not conclude that the partial discharges leading to faults occur significantly more during the cold periods.

6.5.4 Discussion

To test the hypothesis we have looked at the sum of the charge instead of the number of PD-measurements. This makes sense because a discharge with a high charge is more dangerous to cause a fault.

It could be interesting to test this hypothesis for other types of joints and for other periods of the year. The next section is about choosing other periods of the year and adding the resulting numbers in the masterframe. The type of joint is already a feature in the masterframe so the model that eventually predicts the faults will be able to distinguish the clusters accordingly by itself. Chapter 8 elaborates on this model.

One has to note that the data of the joint types is not up to date as was already mentioned in Section 4.1.1.

The sets of the partial discharges that are investigated in this section are based on the clusters and faults. It can be argued that they do not represent well the PD causing faults. Many of the faults, that were not prevented, are not preceded by PD. This would make the dataset of the faults less suitable for this investigation. Also the clusters represent the PD not perfectly because many of the clusters are actually noise. It may be better to investigate the partial discharges at the location and time of the manually assigned warnings. Next section examines when most warnings were assigned.

6.6 Partial discharges in seasons

The previous sections provided indications that the distribution of the partial discharges across the seasons could be of predictive value. We have seen clusters with almost all PD in winter (Figure 51 for example). So this section is about:

Creating features for the masterframe that capture the seasonal effects of the partial discharges.

6.6.1 Defining the seasons

For each cluster the measured PD is categorized per season: for each season we get a value which tells us which percentage of the charge occurred during that season. For this we define the seasons. It seems reasonable to expect that the amount of PD in the warmer periods of the year could be of predictive value, for example by relating it to the type of joint. This brings the idea of dividing the year into two parts: the 6 warmest months of the year and the 6 coldest months of the year. There are also clusters with almost all PD in the months in which the temperature decreases. See examples in Figure 54. So it seems like a good idea to divide the year in half in a different way as well: 6 months in which the temperature decreases and 6 months in which the temperature rises.

Looking at the data of the soil temperature, the coldest 6 months appear to be November until April. We name this period the **cold** period. The **hot** period is May until October. The period in which the temperature rises appears to be the last 3 months of the hot period and the fist 3 months of the cold period: August until January will be called the **rising** period and subsequently February until July is the **decreasing** period.

We also cut the year in the four regular seasons: winter, spring, summer and fall. We define them slightly different than usual because the soil temperature has a small delay in comparison to the regular temperature. We use the intersection of the four aforementioned seasons cold, hot, rising and decreasing.

- **Winter** is the intersection of cold and decreasing: November until January;
- **Spring** is the intersection of cold and rising: February until April;
- **Summer** is the intersection of hot and rising: May until July;
- **Fall** is the intersection of hot and decreasing: August until October.

We end up with 8 seasons that arise by dividing the year in 3 different ways, visualised in Figure 53.

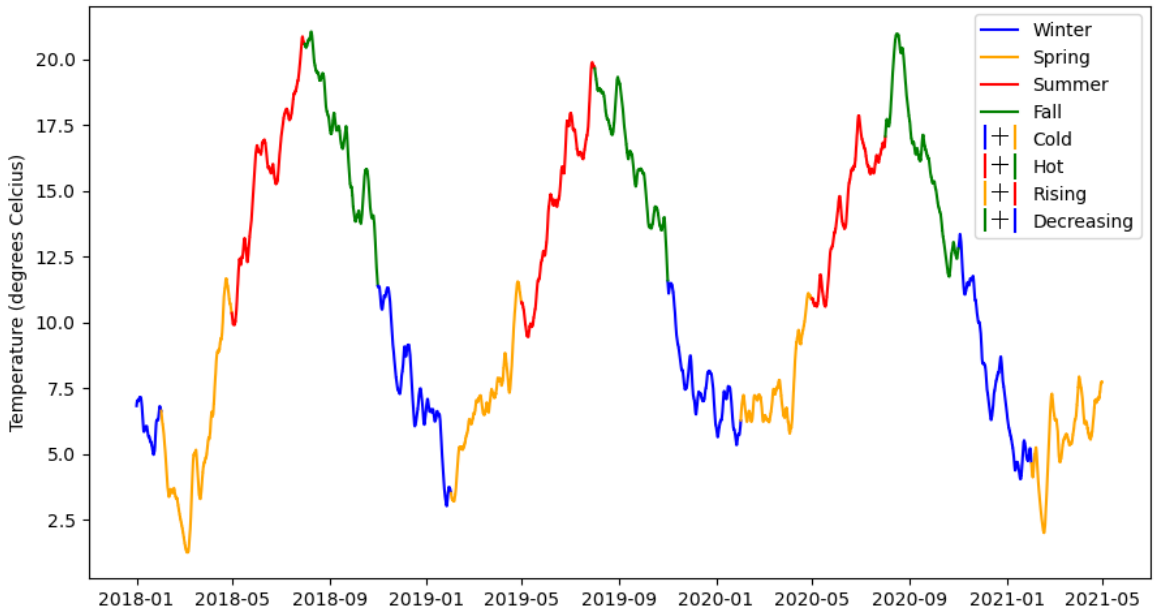


Figure 53: The eight seasons Winter, Spring, Summer, Fall, Cold, Hot, Rising and Decreasing are defined by the time

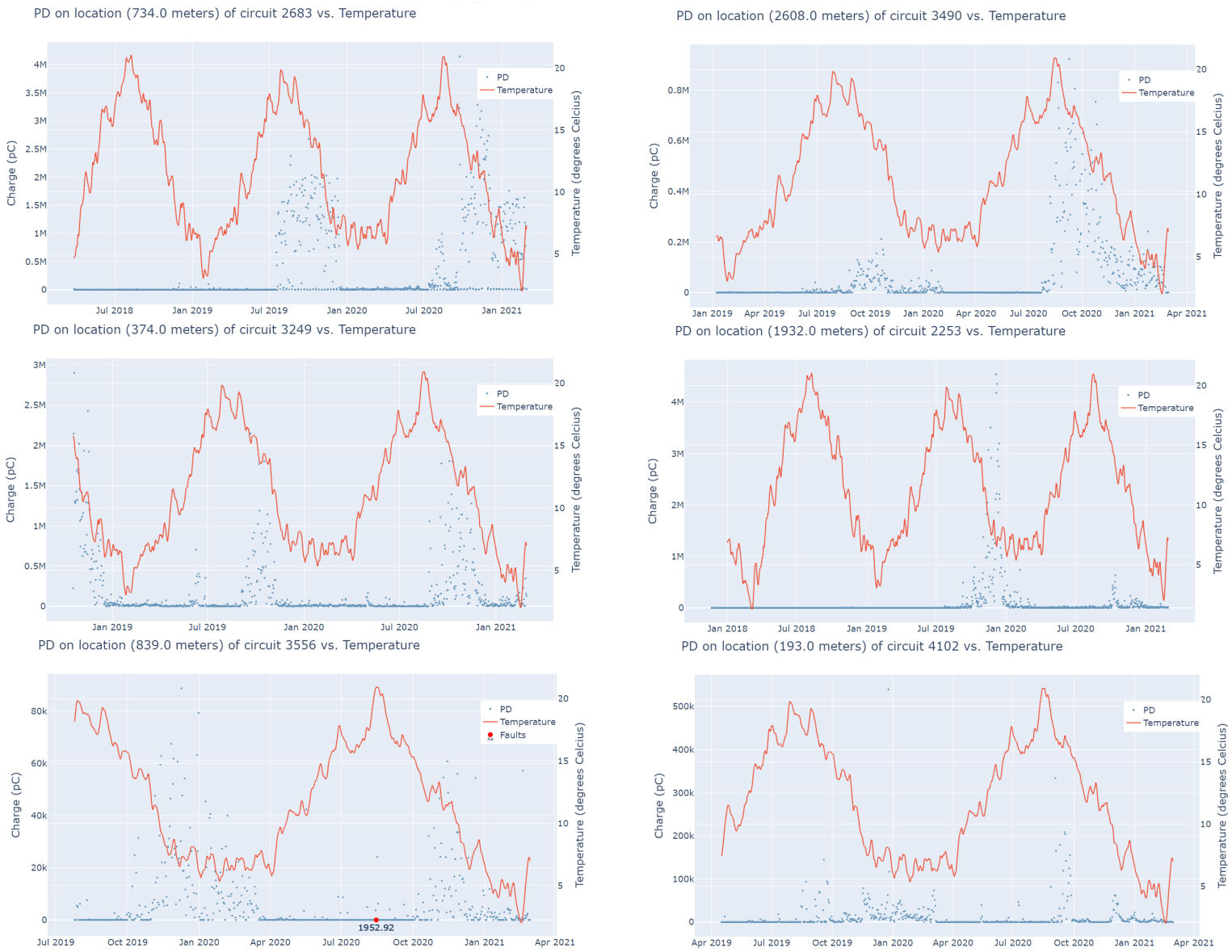


Figure 54: Measured PD at certain locations (with a bandwith of 1% of the circuitlength) for which almost all PD occurred in the **decreasing** season

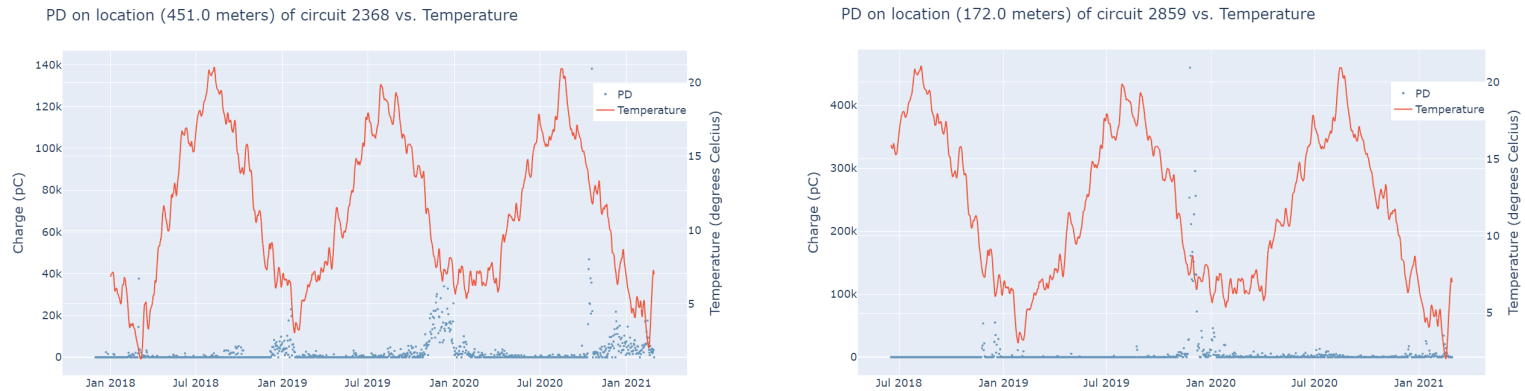


Figure 55: Measured PD at certain locations (with a bandwith of 1% of the circuitlength) for which almost all PD occurred in the **winter** season

6.6.2 Masterframe features

For each season we calculate the percentage of the PD that occurs in that season. We add the next columns to the masterframe:

- *Charge_in_cold_6_months_percentage*
- *Charge_in_hot_6_months_percentage*
- *Charge_rising_temp_percentage*
- *Charge_decreasing_temp_percentage*
- *Charge_in_winter_percentage*
- *Charge_in_spring_percentage*
- *Charge_in_summer_percentage*
- *Charge_in_fall_percentage*

These features tell directly something about the clusters. Some clusters are completely in one season while the total PD at the location of the cluster is divided over multiple seasons. This is possible if the cluster only lasts for one season. Even in this case are the masterframe features informative because there could easily be a reason for this cluster living only in one season.

A cluster could be more dangerous causing a fault if it shows PD divided equally over seasons while most of the PD of that location in the whole existence of the circuit is measured in one season, or vice versa. To catch this effect we introduce new features for the masterframe which for each cluster tell how much of the PD at the location of the cluster is measured during the existence of the circuit, instead of only during the existence of the cluster.

- *Charge_in_cold_6_months_percentage_all_circuit*
- *Charge_in_hot_6_months_percentage_all_circuit*
- *Charge_rising_temp_percentage_all_circuit*
- *Charge_decreasing_temp_percentage_all_circuit*
- *Charge_in_winter_percentage_all_circuit*
- *Charge_in_spring_percentage_all_circuit*
- *Charge_in_summer_percentage_all_circuit*
- *Charge_in_fall_percentage_all_circuit*

The PD to be evaluated for these new features is the PD measured between the fifth percentile and 95th percentile of the location of the cluster. We do this to ignore the outliers. There was no need to do this when we evaluated only the PD of the cluster because the clusters are rounding at the sides, as a result of the cluster algorithm.

In Figure 56 are the clusters of circuit 20133. Two clusters are highlighted by orange and blue. The rest of the clusters are left grey. The vertical orange and blue lines are the fifth and 95th percentiles of the locations of the corresponding clusters. By considering all PD which is between the blue vertical lines, not only most of the blue cluster, but also the entire orange cluster and parts of other clusters are captured. Almost all of these PD occurred in the cold period (November until April). So both the percentage of PD of the blue cluster in the cold period and PD between the blue vertical lines in the the cold period are very high. The second tells that it is not exceptional that the blue clusters live entirely in the cold period. This is an example in which the two features work well together.

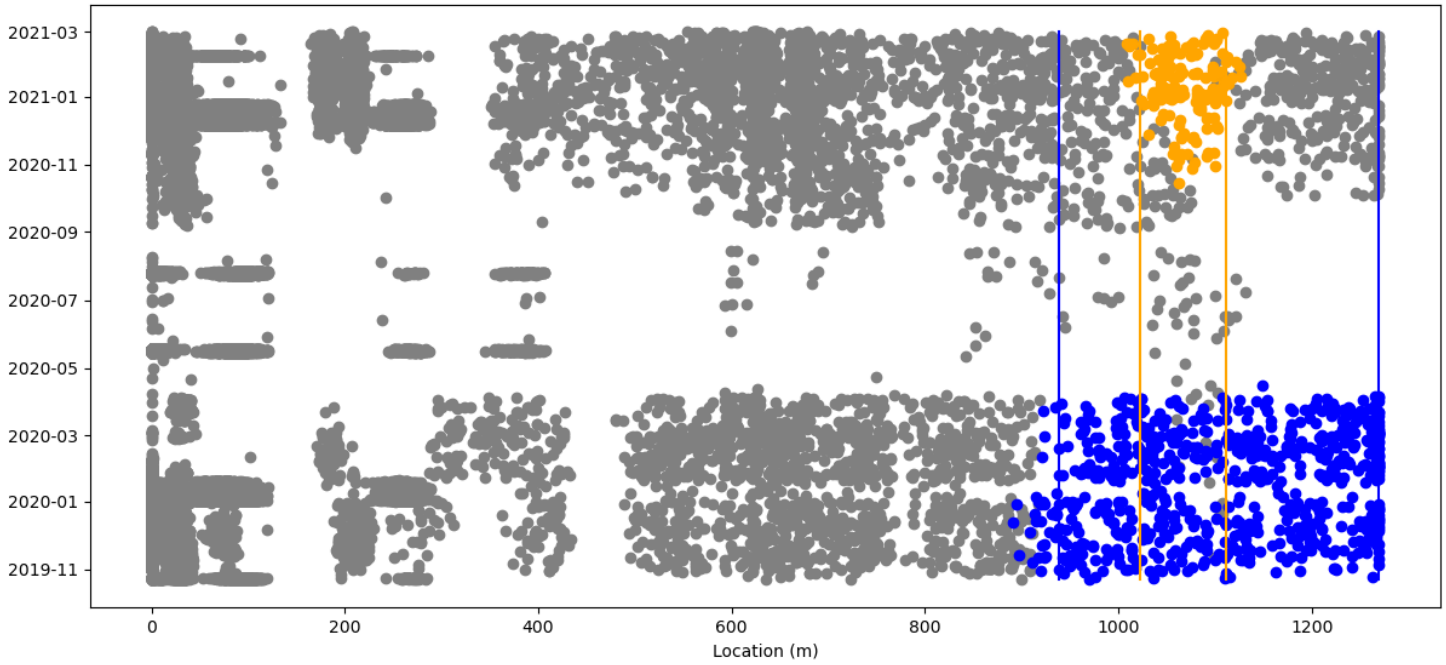
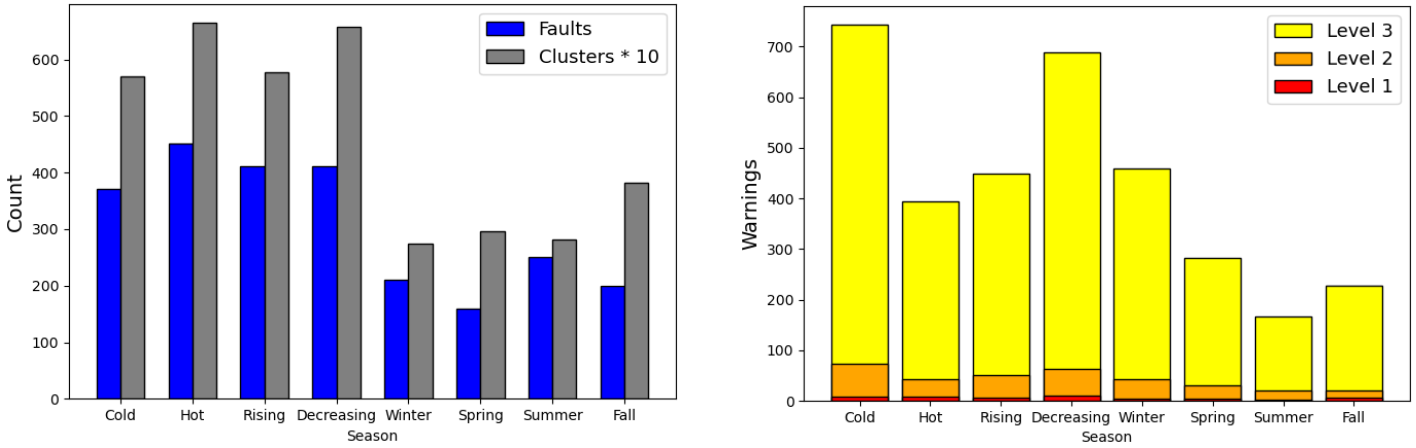


Figure 56: Particles of circuit 20133; clusters 2 and 3 are highlighted with their 5th and 95th percentiles.

6.6.3 Histograms about the seasons

In Figure 57 are distributions of the faults, clusters and warnings over the seasons. Note that the clusters in (a) are listed by 10, so there are more than 10 times as many clusters as faults. In (b) are the warnings per season. There are not many level 1 warnings but it is clear that there are more level 2 en 3 warnings in the cold and decreasing seasons. Therefore also their intersection, the winter, has many warnings. Remarkable is the fact that the faults in (a) occur more often in the hot season and in the summer. One possible explanation for this is that faults usually occur half a year later than the warnings.



(a) Faults and clusters per season

(b) Warnings per season

Figure 57: Distributions of faults, clusters and warnings across the seasons

6.6.4 Conclusion

We defined 8 seasons based on the soil temperature. 16 masterframe are created that describe the amount of charge in the seasons, for both the partial discharges of the cluster and the partial discharges measured at the location of the cluster. Most of the warnings were given during winter and the cold season, although most faults occurred during summer and the hot season.

7 Extract more features from the partial discharges only

One very predictive value of the partial discharges is the charge. The higher the charge, the more dangerous the partial discharges are to cause a fault. This chapter creates cluster features for the masterframe concerning just the partial discharges. So the temperature is not used in any sense.

How can we create masterframe features that quantify the charge of the partial discharges?

One simple but important feature is the total sum of the charge of the cluster. This feature already exists in the current masterframe. So we look for other ways to quantify the charge, and add those quantifications to the masterframe. In Section 6.1 we construct features that highlight clusters that have persistently high values of charge for a while. Section 6.2 is about the distribution of the charge of the PD.

7.1 Persistently high charge

In August 2020, many faults in resin joints occurred. Looking at the plots of the PD at those locations, tells that most of those faults were preceded by much PD. These PD usually started 6 months before the fault occurred and their charge was high. One example is shown in Figure 58. The plot shows the measured particles and the faults around 64 meter (with a bandwith of 1% of the ciruitlength). A fault occurred at 63.55 m at 2020-08-08.

How can we catch the effect of the persistently high charge?

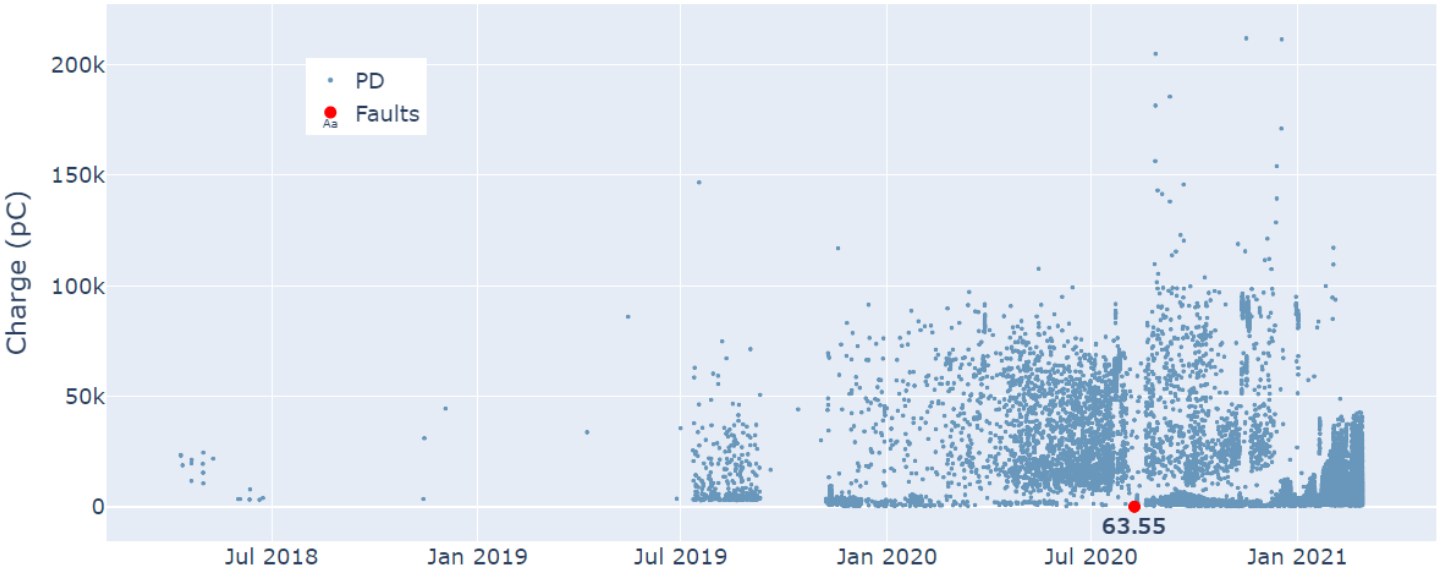


Figure 58: Partial discharges around 64 meter on circuit 2389. A fault occurred at 63.55 meter.

7.1.1 Method

We would like a number that represents the longest period for which the charge is above a certain threshold. The charge is measured every minute and by far most minutes there is no PD, so the charge is 0. As a result, the longest periods when the charge is more than 0 are several minutes. Instead, we are looking at the previous week's charge for each minute. A week is 60 * 24 * 7 = 10080 minutes. Of all these data points, we take the 99.9th percentile. This corresponds to approximately the 10th highest value. We choose this percentile to get the trend of the highest values and avoid the big outliers. The 99th percentile would be 0 too often because of the many minutes without PD.

This method is used to plot the trend of the highest values in Figure 59. A few months before the fault is the curve very high, even above 50k pC. The charge drops rapidly when the fault occurs.

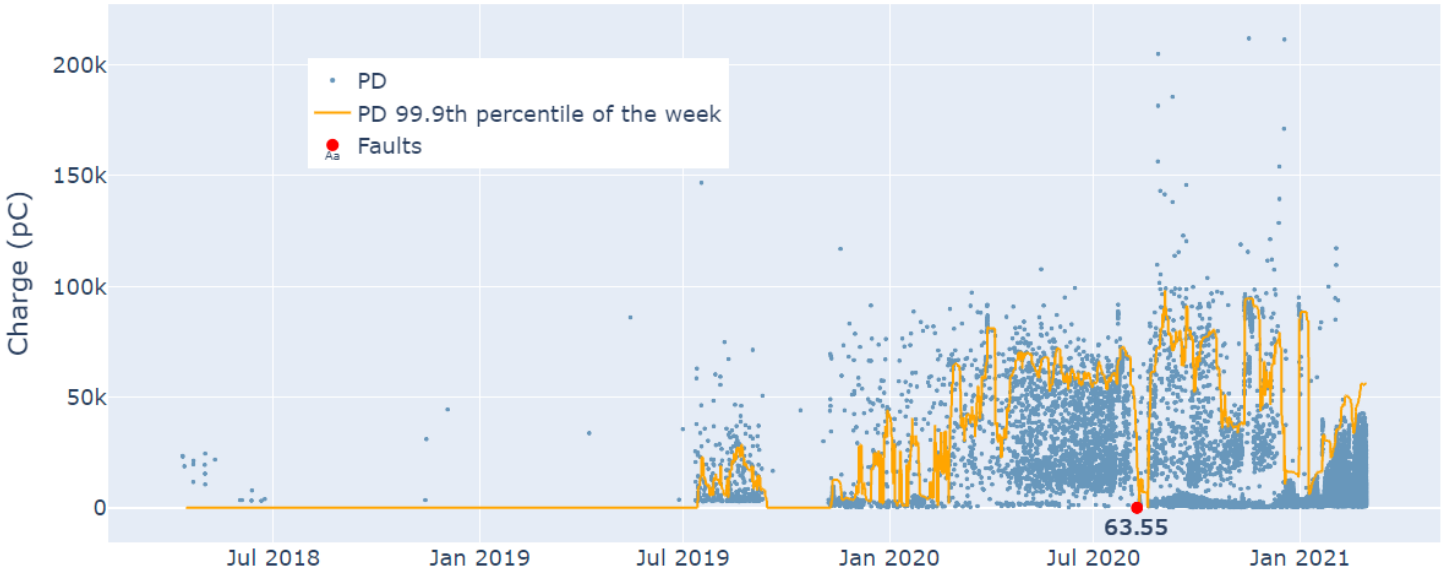


Figure 59: Partial discharges around 64 meter on circuit 2389. The orange line shows the progression of the magnitude of the charge, by taking the 99.9 percentile of the past week for each minute.

7.1.2 Masterframe features

If the threshold is 50000 pC, the longest period in which the curve of Figure 59 is above this threshold is several months. This is significant because it distinguishes clusters in which there is a lot of PD before a fault in a resin joint from other clusters. To be save, we choose a threshold of 20000 pC for the features. This is high enough to still distinguish well, while more clusters are assigned a number higher than 0.

To get a nice number for the master frame, we divide the minutes of the longest period in which the trend is above 20000 pC by $\frac{10080}{7} = 1440$, to get the number of days for which the trend is above 20000 pC. As in Section 6.6, this feature could work very well with the feature about all PD measured at the cluster's location during the circuit's existence. We add two features to the masterframe:

- *number_of_days_more_than_20000_pC*
- *number_of_days_more_than_20000_pC_all_circuit*

7.1.3 Discussion

The method to get the trend of the highest values is only compared to the threshold of 20000 pC. This threshold is not quantitatively substantiated, so it is possible that a different threshold would better distinguish clusters of noise from clusters of actual partial discharges.

In general the trend towards the fault is increasing, while this is not so clear for noise clusters. So a positive slope of a cluster's trend is an indicator that the cluster is not noise. Also the evolution of the density of the particles could be an indicator whether a cluster is noise or PD. In this section, we only looked at the 99.9th percentile. In addition to the persistently high charge, it can also be useful to look at the slope and density or a combination of all of them.

7.2 Fit Weibull distribution to histogram of charge

The distribution of the charge of the PD of a cluster tells a lot about the cluster. This includes the ratio between high charges and low charges, the total sum of charge, the percentiles of the charge and it tells if the cluster contains a bit of noise. Figure 60 shows the distribution of the charge of the particles of a cluster.

How can we quantify the features of the distribution of the charge?

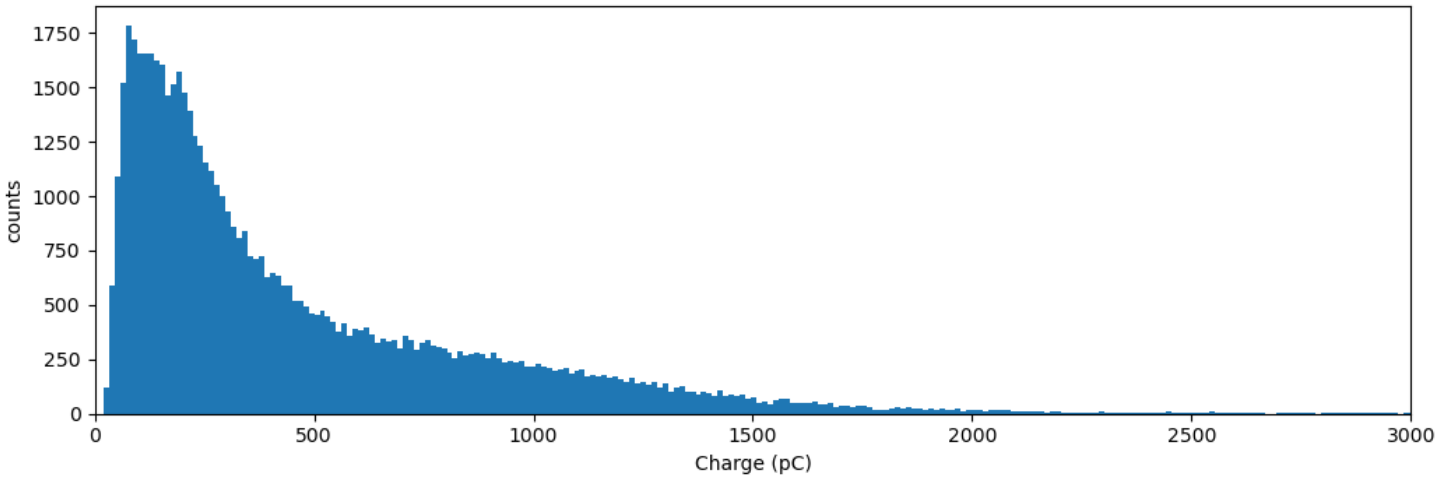


Figure 60: Distribution of the particles of cluster 1 of circuit 4082 across the magnitude of the charge

7.2.1 Method

Figure 60 shows a typical shape of the distribution of the charge of a cluster: a rapid increase in the beginning, followed by a logarithmic decline. Many of the features of the distribution can be quantified by fitting another distribution to it. The Weibull distribution can be used well for this because it is very flexible and non-symmetric. The distribution of the charge of the PD of a cluster is also non-symmetric and can have a variety of forms. The Weibull distribution has the ability to fit on many different shapes due to its versatility. Moreover, the Weibull distribution is forced to be positive and the charge is also always positive. The Weibull distribution is widely used in life data analysis, so this distribution is a good first attempt to fit on the data of the charge. See Section 3.3 for more information about the Weibull distribution. The Weibull distribution is defined by the two parameters α and β:

$$W(x;\alpha,\beta) = \frac{\beta}{\alpha}(\frac{x}{\alpha})^{\beta-1}e^{-(x/\alpha)^{\beta}}.$$

α represents the scale of the distribution and β represents the shape. The method maximum likelihood estimation (MLE) is used to find the best parameters. This method estimates α and β by maximizing a likelihood function, so that the data of the charge is most probable under Weibull(α,β). Figure 61 shows the best Weibull distribution for the data of the cluster in orange. The parameters resulting from the MLE are $\alpha = 507.09$ and $\beta = 1.11$.

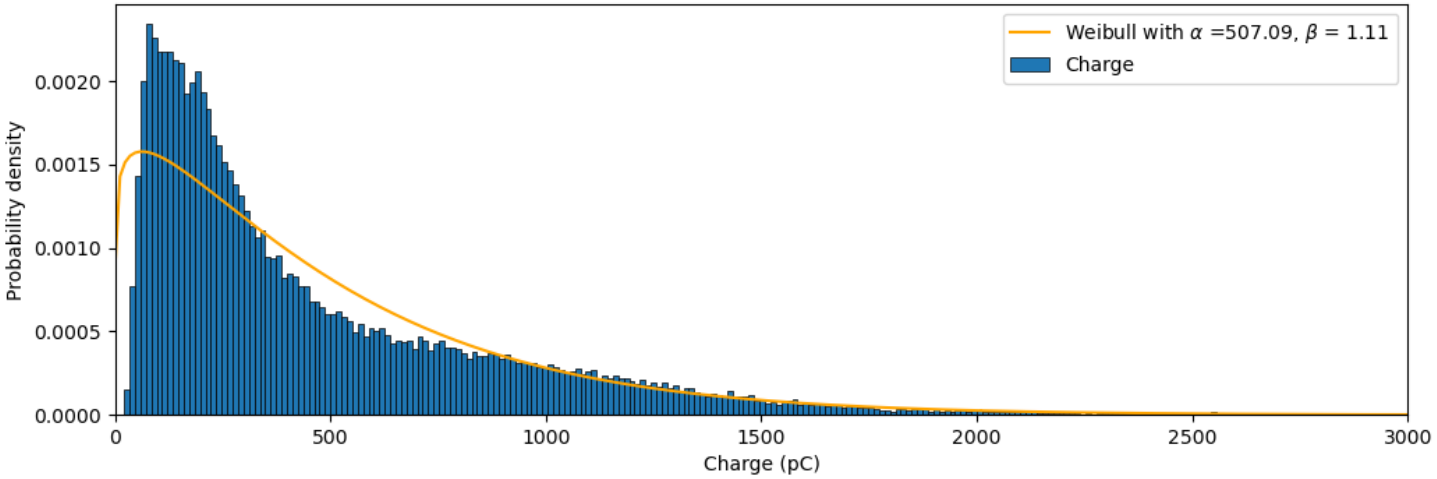
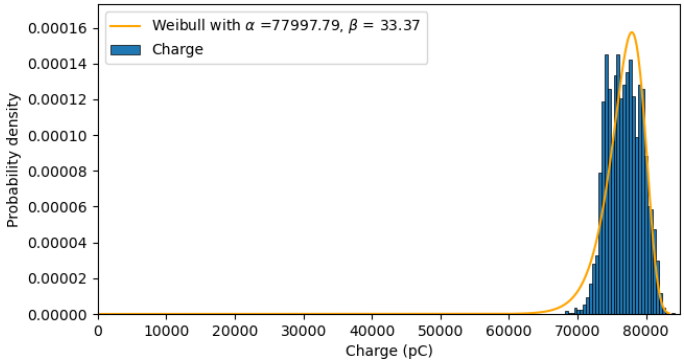
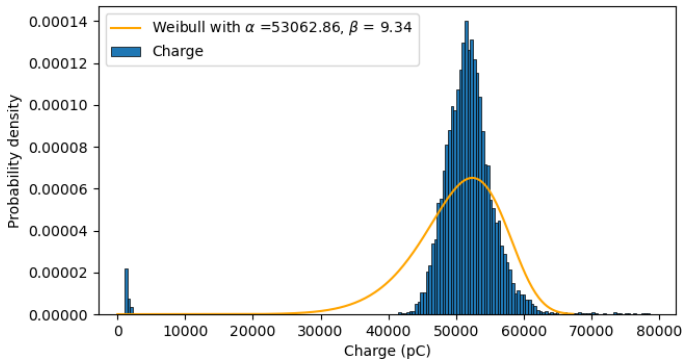


Figure 61: Distribution of the particles of cluster 1 of circuit 4082. The orange line shows the Weibull distribution with parameters $\alpha = 707.09$ and $\beta = 1.11$ that fits best.

The two clusters of Figure 62 are called floating clusters. They have little to no PD with low charge. The Weibull distribution can still fit quite well on these clusters because of the flexibility of the Weibull. The distribution of Figure 62(a) is concentrated around the mode. All partial discharges have approximately the same charge. This s why we see a relative high peak. The shape parameter β describes this because this parameter is higher if the distribution is less stretched out. The scale parameter α is approximately equal to the mode (especially for $\beta > 2$. Thus, this method could also be used to identify floating clusters. However, the cluster in Figure 62(b) has some PD with low charge. This disturbs fitting the Weibull to it.



(a) Cluster 0 of circuit 4082

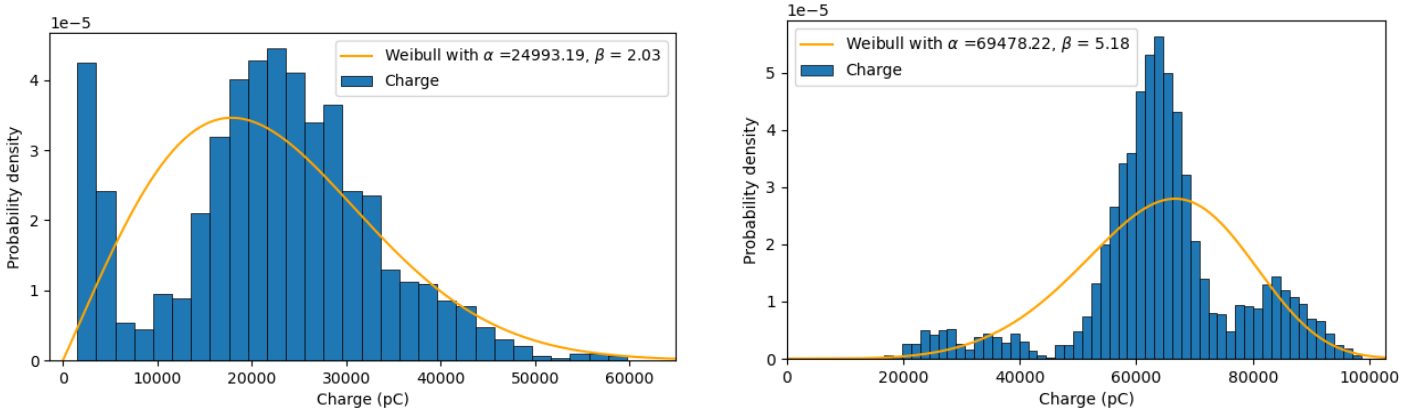


(b) Cluster 17 of circuit 20133

Figure 62: Examples of floating clusters and their best Weibull fits. The flexible Weibull distribution fits well on floating clusters. The low charges thwart.

For some clusters there are no parameters for which Weibull(α, β) fits well on the charges. Figure 63 shows 2 examples. The charge in Figure 63(a) has two different sources. The first two bins are a clear result of noise. Figure 63(b) shows three hills which also suggests there are multiple sources. It is also possible that the behaviour of the PD changed over time, because some clusters span a time period of more than a year. So the hills in Figure 63(b) could be the distributions of three different periods of the cluster.

Anyway, the resulting Weibull(α, β) does not fitt well on these datasets. The obtained parameters α and β for the shape and scale are not very representable. To capture this symptom we determine the goodness of fit. We use four different tests to measure the goodness of fit and eventually find out which one helps the best to predict the faults in chapter 7.



(a) Cluster 0 of circuit 20830

(b) Cluster 12 of circuit 20133

Figure 63: Examples of clusters and their best Weibull fits where noise causes the Weibull distribution to not fit properly.

The first criterion is the log of the maximum likelihood $\hat{L}$, found by the MLE. The likelihood of the charges under Weibull(α, β) gives the maximum likelihood. Taking the log gives a measure of the goodness of fit:

$$\text{Log-Likelihood} = \ln(\hat{L}).$$

The Akaike Information Criterion (AICc) also depends on the number of parameters k:

$$\text{Akaike Information Criterion} = 2k - 2\ln(\hat{L}).$$

A somewhat more complicated value is the Bayesian Information Criterion (BIC). The number of measurements of PD n has its influence:

$$\text{Bayesian Information Criterion} = k\ln(n) - 2\ln(\hat{L}).$$

The Anderson–Darling (AD) test is based on the distance between the empirical cumulative distribution function F_n and the hypothesized distribution $W(\alpha, \beta)$:

$$\text{Anderson-Darling goodness of fit statistic} = n\int_{-\infty}^{\infty} \frac{F_n(x) - W(x;\alpha,\beta)}{W(x;\alpha,\beta)(1 - W(x;\alpha,\beta))}\, dW(x).$$

The Anderson–Darling distance places more weight on observations in the tails of the distribution. So this test is good at detecting the noise as in figure 63.

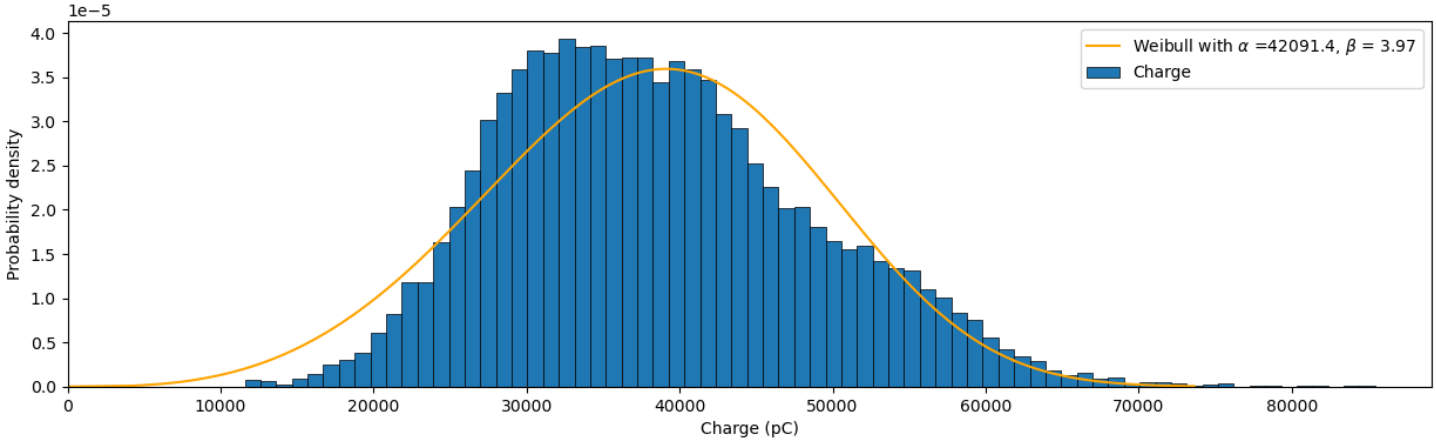


Figure 64: Distribution of the particles of cluster 1 of circuit 20830 and the best Weibull fit. The four criteria of the goodness of fit are: Log-Likelihood = −215504, AICc = 431014, BIC = 431028, AD = 119.

7.2.2 Masterframe features

The method to find the best Weibull distribution for the data of the charge gives the next cluster features for the masterframe:

- *weibull_scale*: The parameter α
- *weibull_shape*: The parameter β
- *goodness_of_fit_loglik*: Log-Likelihood
- *goodness_of_fit_AICc*: Akaike Information Criterion
- *goodness_of_fit_BIC*: Bayesian Information Criterion
- *goodness_of_fit_AD*: Anderson-Darling

7.2.3 Discussion

The Weibull distribution fits the charge better if the noise is not taken into account. Therefore, removing the noise can improve the features. After fitting the Weibull first on all charge, the resulting goodness of fit can help to detect the noise.

All features are fed to a model that ultimately predicts the faults. The features of goodness of fit help the model to determine if the features of the scale and shape are reliable. It is probably a good idea to combine some parameters, such that the model knows immediately if the parameters of the Weibull fit are useful. One suggestion is multiplying the parameter of the scale by the goodness of fit (after normalizing). The resulting feature indicates both the scale and the reliability.

Instead of looking at the distribution of the charge one could investigate features based on the distribution of the locations of the measured PD of a cluster. One advantage is that the density of the PD can easily be extracted from the distribution of the locations.

The Weibull seems a good distribution to compare to the distribution of the charge. However the log-normal distribution also fits well to distributions with a steep slope for the lower values and a flat drop afterwards. The distribution in Figure 64 is similar to a Normal distribution, and there are more clusters with a similar charge distribution. It should be examined whether the features resulting from fitting the log-normal and the Normal distribution contribute to the prediction of faults.

7.3 Conclusion

We created 2 features about the persistently high charge and 6 features about the charge distribution. There is still room for improvement for both methods so that the features can better distinguish PD clusters from noise clusters. Although the examples show that the features created in this chapter can already have predictive value. In the next chapter, the classification model uses all the features to classify the clusters. We will see that especially the features about the charge distribution contribute a lot to the classification.

8 Predicting faults using the features

In Chapters 6 and 7 we have constructed 33 features for the masterframe. In this Section we predict faults using a hierarchical classification model. This model has been created by Alliander and its input is a masterframe. It makes use of the library *XGBoost* to classify the clusters by means of the features. The input and output of this model will be discussed but more details on how the model precisely works are out of scope of this thesis.

Do the newly created features make a positive contribution to the model to predict faults?

8.1 Method

As mentioned in Chapter 1, it is difficult to measure how good the fault predictions are. If a fault is predicted to occur, the joint is replaced and the fault is prevented. So in reality there would be no fault. What we would like to predict are faults when no one would act on the predictions.

We can use the manually assigned warnings of Section 4.4 for this. These warnings are considered to be a very good representative of potential faults: the warnings predict the faults very well when there would be no maintenance work. The clusters are linked to the warnings according to Definition 8. Level 1 and 2 warnings are far more serious than level 3 and noise warnings.

Definition 9. A cluster is **dangerous** if it requires manually inspection by a trained operator. All clusters linked to a warning level 1 or 2 are considered to be dangerous. All clusters to be examined are considered dangerous if a trained operator were to assign a warning level 1 or 2.

We predict the dangerous clusters and consider the results to represent the prediction of the faults.

The classification model needs a part of the masterframe to train and a part of the masterframe to test. We choose the training set to be the data of the masterframe until 2019-12-31. All particles until 2019 are used by the the clustering algorithm to create the clusters. Then all the cluster features are calculated to construct the masterframe until 2019-12-31. The training set is labelled by historically assigned warnings by SCG operators. The model is trained to act as an operator to assign warnings to clusters. The test set is the masterframe of the data between 2020-01-01 and 2020-09-31, constructed in a similar way. The warnings are excluded from the test set, because that is what needs to be predicted.

During the training phase the model uses the training set, including the warnings. It comes up with an algorithm that for each cluster c assigns the **dangerousness** (degree of danger) of the cluster $\delta^c \in [0, 1]$ to it. The input of this algorithm are the features. The dangerousness δ^c tells to what extend a cluster should be considered dangerous. The model uses the warnings as the labels, to create the algorithm. The weight distribution [100, 100, 3, 1] is used to train the model mainly on the level 1 and 2 warnings, and reduce the influence of the level 3 and noise warnings. So the idea is that δ^c is high for clusters with a level 1 or 2 warning.

This resulting algorithm assigns the dangerousness δ^c to each cluster c from both the training set and the test set. The warnings are not included in the test set. The model classifies cluster c to be dangerous if the dangerousness δ^c is higher than a certain threshold. If this threshold is very low, all dangerous clusters are predicted well but also many harmless clusters are unexpectedly considered dangerous. Replacing joints at the locations of the clusters costs money, so we do not want the threshold to be lower than necessary. If the threshold is very high, many clusters will be considered not dangerous when they are, resulting in power outages. So there is an optimal threshold between 0 and 1. The model creates the ideal threshold. This threshold is defined as follows.

Definition 10. The **ideal threshold** $\theta \in [0, 1] = \min\{\delta^c \mid \text{cluster } c \text{ of training set has a warning level 1 or 2}\}$ is the minimum dangerousness of all clusters with warning level 1 or 2 in the training set.

The model classifies cluster c to be dangerous if the dangerousness δ^c is higher than the threshold θ. The other clusters are classified as not dangerous. The ideal threshold is the maximum threshold for which the dangerousness of all clusters in the training set with warning level 1 or 2 is higher than this threshold. Using the threshold on the training set always result in 0 false negatives, so all dangerous clusters are rightly classified as dangerous. On the other hand this would probably also result in false positives. The better the classification model and resulting threshold, the fewer false negatives and false positives there are in the test set.

Definition 11.
TP = True positives: number of clusters correctly classified as dangerous;
TN = True negatives; number of clusters correctly classified as not dangerous;
FP = False positives: number of clusters incorrectly classified as dangerous;
FN = False negatives: number of clusters incorrectly classified as not dangerous.

The ideal threshold may be too high to classify all clusters with warning level 1 or 2 in the test set, as dangerous. There could be some clusters c in the test with a warning level 1 or 2 for which the dangerousness δ^c is lower than the ideal threshold θ. To check the performance of the model, the classification of the clusters into dangerous or not dangerous can be compared to the actual manually assigned warnings.

8.2 Results

We run the model three times, such that the dangerousness of the clusters depend on three different sets of features. First we force the model to only use the 33 features we created in this thesis, create δ^c for each clusters and check the performance. Alliander's features are not taken into account in this test. Next we run the model with only the 44 features that Alliander currently uses to predict faults. The features created in this thesis are not taken into account in this second test. Last we let the model base the dangerousness of the clusters on all clusters: the 33 new features and Alliander's features combined. We call the 33 features of this thesis the **new features**. The 44 features that Alliander currently uses are called **Alliander's features** and the two sets combined are called **all features**.

8.2.1 New features

Based on the 33 new features, the model determined the dangerousness of the clusters in the training set, determined that the **ideal threshold** is 0.73 and determined the dangerousness of the clusters in the test set. Using the ideal threshold to classify the clusters of the test set results in Figure 65.

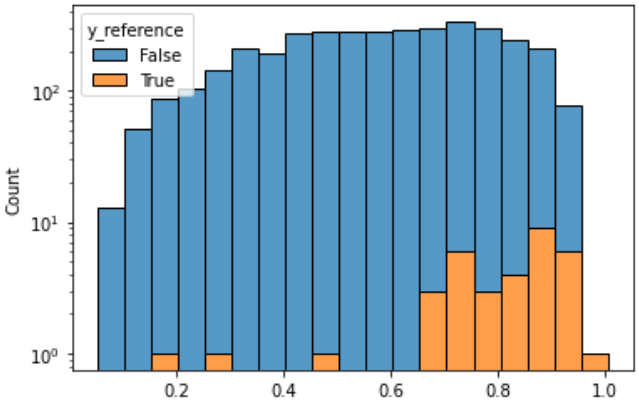
Figure 65(a) shows the confusion matrix of the test set. *pred_neg* is the number of clusters classified as not dangerous and *pred_pos* is the number of clusters classified as dangerous. There are 3660 clusters in the test set for which 1000 are classified as dangerous. So the dangerousness of 1000 clusters is higher than the ideal threshold of 0.73. 2660 clusters of the test set are classified as not dangerous. The manually assigned warnings are also shown in the confusion matrix. The row *overlap_lvl1* shows that there are 7 clusters in the test set that have a level 1 warning. 4 of them are classified as not dangerous and 3 are classified as dangerous. *overlap_lvlN* is the number of clusters with a noise warning. *no_information* tells that the clusters do not have a warning.

Figure 65(b) shows a histogram of the 3660 clusters of the test set. The dangerousness δ is on the horizontal axis. The vertical axis shows the number of clusters in the test set with the corresponding dangerousness. The orange parts of the bars are the clusters that are classified as dangerous. The blue bars show the number of clusters that are classified as not dangerous. Note the logarithmic scale. 3 clusters have a very low dangerousness while in fact they have been assigned a level 1 or 2 warning. In the ideal scenario, the dangerous clusters have a high dangerousness and the other clusters a low dangerousness, resulting in a plot where the orange bars are all on the right and the blue bars on the left. We see that most of the orange bars are on the right, which is a sign that the features used have predictive power.

In Figure 65(a) we see that the ideal threshold of 0.73 classifies 9 clusters as not dangerous while in reality they have been given a level 1 or 2 warning. On the other hand it rightly classifies many clusters as dangerous. So the number of false negatives is 9 and the number of false positives is $1000 - 3 - 23 = 974$. If the threshold would be 0.65 instead, we see in Figure 65(b) that the number of false positives would only be 3. However the threshold can not be chosen after seeing the results of Figure 65 because in reality we would not have access to the right labelling of the clusters. The ultimate goal is to predict the dangerous clusters without knowing it beforehand.

	pred_neg	pred_pos	total
no_information	1915	429	2344
overlap_lvl1	4	3	7
overlap_lvl2	5	23	28
overlap_lvl3	124	246	370
overlap_lvlN	612	299	911
total	2660	1000	3660

(a) Confusion matrix: prediction of all 3660 clusters in the test set and their manually assigned warnings

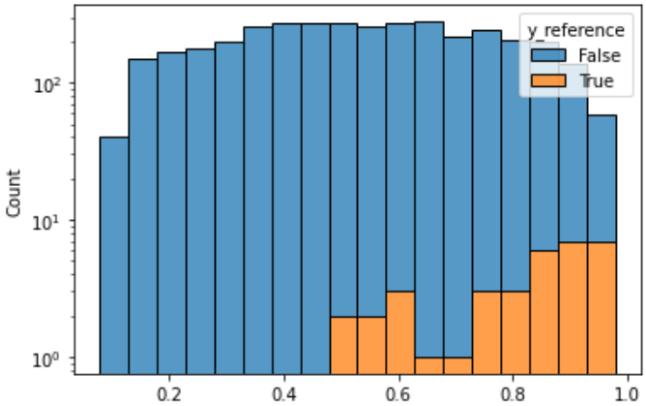


(b) Number of clusters with a warning level 1 or 2 (True, orange) and the clusters with warning level 3, Noise or no warning (False, blue) with their dangerousness on the horizontal axis

Figure 65: Results of the classification model for the 33 new features

	pred_neg	pred_pos	total
no_information	1999	345	2344
overlap_lvl1	0	7	7
overlap_lvl2	9	19	28
overlap_lvl3	137	233	370
overlap_lvlN	605	306	911
total	2750	910	3660

(a) Confusion matrix: prediction of all 3660 clusters in the test set and their manually assigned warnings

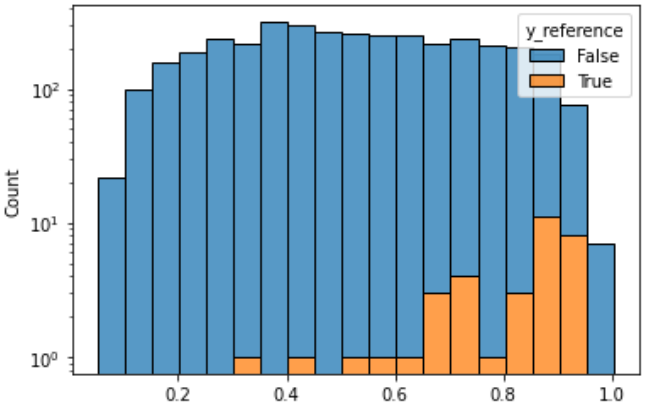


(b) Number of clusters with a warning level 1 or 2 (True, orange) and the clusters with warning level 3, Noise or no warning (False, blue) with their dangerousness on the horizontal axis

Figure 66: Results of the classification model for Alliander's features

	pred_neg	pred_pos	total
no_information	2091	253	2344
overlap_lvl1	1	6	7
overlap_lvl2	10	18	28
overlap_lvl3	152	218	370
overlap_lvlN	677	234	911
total	2931	729	3660

(a) Confusion matrix: prediction of all 3660 clusters in the test set and their manually assigned warnings



(b) Number of clusters with a warning level 1 or 2 (True, orange) and the clusters with warning level 3, Noise or no warning (False, blue) with their dangerousness on the horizontal axis

Figure 67: Results of the classification model for all features

8.2.2 Alliander's features

When the classification model only uses Alliander's features, the ideal threshold is 0.71. The results of the test set are shown in Figure 66. There are 9 false negatives, but the number of false positives is lower this time. For Alliander's features the false positives are $910 - 7 - 19 = 884$.

Figure 66(b) shows that there are no clusters with a warning level 1 or 2 and a low dangerousness (lower than 0.4). This suggests that this is a better result than we see in Figure 65(b). However there are much more dangerous clusters with a dangerousness lower than 0.65. The threshold used to classify the clusters is therefore decisive in determining whether the new features or Alliander's features lead to a better classification.

8.2.3 All features

The model now makes use of the 33 new features and the 44 features of Alliander combined. The resulting ideal threshold is 0.74. Figure 67 shows the results. 11 clusters are wrongly classified as not dangerous. However the total number of clusters that are classified as dangerous is only 729.

We see in Figure 67(b) that all clusters with a dangerousness over 0.95 do not have a warning level 1 or 2 in reality. That in itself is remarkable, but there are only a few clusters that are assigned such a high dangerousness. The orange bars seem more scattered than in Figure 66(b), but the number of dangerous clusters that are assigned a dangerousness lower than 0.65 is only 5, compared to 8 for Alliander's features.

8.3 Feature importance

The model uses the features to classify the clusters. The contribution of the features is also determined by the model. This represents the predictive value of the features. The two features of the parameters of the Weibull fit (*weibull_scale* and *weibull_shape*) of Section 7.2 appear to contribute a lot to the predictions. Also the feature of the correlation between the PD and the soil temperature (*correlation_pd_temp*) of Section 6.1 has a large predictive power. The relative feature importance of all features can be found in Figure 68, 69 and 70 in the Appendix.

8.4 Conclusion

The predictive model classifies clusters as dangerous to cause a fault or not dangerous. The model uses cluster features and DNV's manually assigned warnings to train. The training set used for this model consists of clusters that are labelled by historically assigned warnings by SCG operators. The model is trained to act as an operator to assign warnings to clusters.

We trained the model on three sets of cluster features, and then compared the predictive performance of the model on the test set. First, the features created in this thesis are used, second, Alliander's features, and finally all features combined are used. The test set consists of 3660 clusters of which 35 with a warning level 1 or 2.

When using the 33 features of this thesis, the model performs remarkably well. Almost all clusters with a warning level 1 or 2 received a high degree of danger. The threshold for determining which clusters are classified as dangerous seems a bit high, causing the model to incorrectly classify some clusters as not dangerous. Of the 35 clusters with a warning level 1 or 2, 26 have been identified and classified as dangerous. In total, 1000 of the 3660 clusters are classified as dangerous.

Alliander's features resulted in 910 clusters that were classified as dangerous, of which 26 with a warning level 1 or 2. Alliander's features combined with the 33 features of this thesis resulted in only 729 clusters that were classified as dangerous, of which 24 with a warning level 1 or 2. So by adding the new features, the model has fewer false positives, but also misses more clusters that should have been classified as dangerous.

The model can predict well when only the features of this thesis are used. These features apparently contain a lot of relevant data. Adding these to the model with Allianer's features improves the performance of the model as the number of false positives decreases with approximately 20%. The parameters of the Weibull fit (*weibull_scale* and *weibull_shape*) and the correlation between the PD and the soil temperature (*correlation_pd_temp*) contribute a lot to the predictions. Although the number of false negatives rises sightly when all features are combined, this concerns only individual cases. These cases need to be looked at more closely.

The most important is the significant decrease in false positives, as this reduces the number of clusters that must be manually assessed by a trained operator. Since the warnings represent the faults, we conclude that the newly created features contribute positively to the model to predict faults.

8.5 Discussion

To test the performance of the classification model, we used a training set of clusters until 2019 and a test set of clusters between January and September 2020. These periods were used because this is customary within Alliander for this classification model. However, using more recent clusters leads to a more reliable approximation of the model's performance on future clusters. Our training set consists of clusters between 2014 and 2019. The data from the early years does not represent the future clusters well, so it was also worth considering using only the last years of the training set.

The classification model makes use of the library *XGBoost*. The parameters used to train are the same for all three sets of features. They were pre-selected based on Alliander's features, so that the model performs relatively better on Alliander's features. Ideally we would use the best parameters for all sets of features.

9 Discussion

In Section 4.1 we saw that there are many fault measurements that are caused by human actions. These do not lead to a power outage or cannot be predicted with PD. These faults were included in the data, which could have given a wrong picture. The faults outside circuits should be filtered better. Nowadays Alliander is able to filter those faults.

In Chapter 6, the temperature data measured at De Bilt were used. However these data are related to the PD on the circuits over the entire service area of Alliander. The relations would have been better represented if the temperature of the regions of the respective circuits had been used. We used the data of De Bilt because the locations of the circuits are not easily accessible. Besides the location, the depth of the measurements is also not exactly the same as the depth of the cables. The depth of the joints were assumed to be all between 80-100. We used soil temperature level 3 which are measurements between 28-100cm. Although these depths do not exactly match, these data are the best available temperature data.

In Chapter 6 we related the temperature to the discharge magnitude. Also in Chapter 7 we constructed cluster features based on the magnitude of the charge. This makes sense because a discharge with high charge gives more reason to suspect that a fault will occur. However the number of PD measurements could also be predictive. In addition to the features created based on the magnitude, these features can also constructed based on the number of measurements. The combination of these features gives the classification model even more information. This will likely improve predictive performance even further.

It should be noted that in Chapter 8 no faults were predicted. Instead, it was determined which clusters are dangerous, or should receive a warning level 1 or 2. This is because Alliander's classification model uses the manually assigned warnings as labels for the model to train. It is assumed that level 1 and 2 warnings represent a potential fault and that clusters without such a warning will not cause a fault. This can be questioned, but it should be very accurate.

Ideally, we would like to check to what extent the predictions of the classification model have come true. However, many predicted faults have been prevented of course. The data of the replacements of the joints would offer a solution. However, a lot of work is required to properly match this data to the correct circuits. Alliander has not yet succeeded in this. So unfortunately we were unable to use this data in this thesis.

Many features constructed in this thesis are based on parameters. These parameters are determined so that the feature best represents a character from a cluster. Whether these parameters are actually the best when it comes to predicting faults has not been tested. These parameters can be tweaked according to the performance of the classification model. This takes a lot of time, which makes it beyond the scope of this thesis.

Features that involve PD after the cluster

In Section 7.1 we created the feature *number_of_days_more_than_20000_pC_all_circuit*. In comparison to the feature *number_of_days_more_than_20000_pC* this deals with the PD at the median of this cluster with a bandwidth of the distance between the median and the 5th percentile, during the entire period in which the circuit is being monitored. Likewise, in Section 6.6, we have created 8 features on the seasonal distribution of PD on the cluster median over the entire period of the circuit. For example, *Charge_in_cold_6_months_percentage_all_circuit* concerns the PD in the coldest 6 months of the year. These features were in addition to the features about the PD of the clusters themselves only. For example *Charge_in_cold_6_months_percentage*.

These features concern not only the PD of the cluster and the PD that occurred earlier in the same location, but also the PD that occurred after the cluster. Care should be taken when using these features to test the performance of the classification model.

We want to predict which clusters are a precursor to a fault so that the fault can be prevented by replacing the joint. Of course, to test whether the features can predict these potential faults, we should not use the data obtained after the fault has occurred or the joint has been replaced. Only the data before the fault or replacement of the joint may be used.

It often happens that a fault occurs much later than the end of the cluster causing the fault. Therefore, the particles after the cluster may also be used to make the prediction. However, the particles after the fault/replacement should absolutely not be used when testing the classification model.

This is not a problem when the model will be put into operation because it will then predict, on the basis of PD, whether a joint should be replaced. So then no fault or joint replacement has occurred yet. Although the model should only use the information before the mistakes and joint replacements during training.

The 9 features that concern the PD after the cluster can not be used carelessly during testing. They may only contain the PD that occurs before a fault/joint replacement. Unfortunately, the data of the replacements are not available, so this is unfortunately not possible. However, to ensure that there is no cheating during the testing of the classification model, we should limit these features to the PD up to and including the cluster. Then these features still add something because in addition to the PD of the cluster, they also contain the PD before the cluster.

We assume here that no fault/replacement can take place in the middle of a cluster. This would make things even more complicated because then not even all particles of a cluster would be allowed to be used in testing.

So for both training the model and testing the performance the features need to be tweaked slightly. So the test results of Chapter 8 are not entirely valid. However, only for 9 of the features, extra features have been compiled, and these features (recognizable by the *_all_circuit*-part) have no major impact on the predictive model as seen in Figure 70 in the Appendix. Thus, the conclusion remains that the features created in this thesis contribute to the prediction of faults.

The features involving PD after the cluster can thus be put to good use when the classification model is put into use. Versions should therefore also be made of all other features that not only concern the PD of the cluster, but also the PD before and after the cluster. The combination of a feature about the PD of a cluster and a feature about the PD of a circuit in the same location will work very well together to predict faults because the behavior of the cluster is then compared to what normal PD development is at that location.

10 Conclusion

Alliander is responsible for the maintenance of the medium voltage grid. This includes preventing power outages so that the cable joints can be replaced and customers are not left without power. Faults are short-circuit currents, which usually lead to power outage. Alliander predicts faults using the partial discharges (PD), measured by SCG on the circuits. The contribution of this thesis can be divided into two parts. First, we collected faults data and weather data so that we could thoroughly analyze the faults in relation to the PD and the weather. Second, we searched for predictive features of the PD clusters, quantified them and tested their ability to predict faults.

In the period December 2014 to February 2021, 822 faults occurred, an average of 1.56 faults per circuit. 358 of these faults occurred in 2020 and this number is increasing every year as the number of SCG-systems increases. We added the faults to the masterframe of clusters and only 56 faults in this period can be linked to a cluster. These faults have not been prevented because no warning level 1 or 2 has been issued at the relevant locations. Thus, only 7% of the faults that were not prevented could have been predicted from the PD. Many potential faults have already been prevented by replacing the joints.

Comparison of the faults with the weather conditions suggests that faults can be predicted using weather data. The temperature has the greatest influence on the faults. The soil temperature also affects the PD preceding a fault. Depending on the material of the joint, the PD in a joint correlates positively or negatively with the soil temperature. Most faults occur during the warm seasons of the year, while the warnings and the PD are more likely to occur during the cold seasons, suggesting that faults usually occur about 6 months later than the PD preceding them. A deeper analysis of the faults requires the data of the joint replacements.

Alliander uses a classification model that classifies clusters as dangerous to cause a fault or not dangerous. We have constructed 33 features in addition to the existing features that Alliander has already constructed. These features are used by the model to get better classification results.

25 of these features are based on the relation between discharge magnitude and soil temperature during the discharge. The correlation between them has a high predictive power. Also, the distribution of the PD across the seasons contributes to the prediction of the faults. The distribution of the discharge magnitude is also very informative. The shape and scale of these distributions are two features extensively used by the classification model to predict faults.

While there is still a lot of room to improve the features, they are already effective. When the classification model only uses these 33 features instead of Alliander's features, the model is able to predict faults with similar accuracy. Thus, it is possible to predict faults from the relation between faults, PD and weather conditions. When the features constructed in this thesis are combined with Alliander's features, there is a significant improvement. Then 20% fewer clusters are incorrectly classified as dangerous. Thus, the features constructed in this thesis contribute to the prediction of faults. This helps prevent power outages.

List of features

5 features about the faults

Definition 12. A fault is linked to a cluster if the fault occurred at a location between the 5th percentile and the 95th percentile of the location of the particles of the cluster and the fault occurred after the first particle of the cluster has been detected.

- *fault-count_inside/after_cluster*: Number of faults linked to the cluster;
- *Date/time (UTC)_of_first_fault*: Time of the first fault linked to the cluster;
- *location_of_first_fault*: Location of the first fault linked to the cluster;
- *locationdelta_of_first_fault*: The distance between the location of the first fault and the median location of the cluster;
- *locationdeltarelative_of_first_fault*: The distance between the location of the first fault and the median location of the cluster, divided by the circuitlength.

9 features about the correlation between PD and temperature

The period between the first and last particle of the cluster is used for the feature *correlation_pd_temp*. The soil temperature level 3 (depth of 28-100cm) from De Bilt (CDS) is used. The charge q_h^c is summed for each hour. Then for each hour of this period the mean of the past 10 days (including this hour) q''_{h_i} is assigned to this hour. This way the first 10 days of the period are discarded. Then the correlation between these two series is calculated:

q_h^c : The sum of the charge of the measured PD of cluster c during hour h, in picocoulomb (pC);

$q''_{h_i} = \frac{1}{240} \sum_{j=i-239}^{i} q_{h_j}$: the average of the 240 hourly charge values prior to and including h_i;

$q'' = [q''_{h_{240}}, ..., q''_{h_{max}}]$: The series of the smoothed charges;

t_h : The temperature during hour h, in degrees Celcius;

$t^c = [t_{h_1}, ..., t_{h_{max}}]$: The series of temperatures per hour in cluster c.

- *correlation_pd_temp*: $Pearson(q'', t)$,
  Correlation between the soil temperature and the PD of the cluster.

A rolling window ρ is used for the next 4 features. For each day of the cluster the correlation of the past 20 days (including the day in question) $\rho^c_{d_i}$ is calculated. This way we get a rolling window with a length of the number of days of the cluster and for each day a coefficient. The two series of data which are used to calculate the correlation have 20 data points each. For each day the mean of the temperature T_d and the sum of the charge of the PD Q^c_d are used. No correlation is being calculated for a day if the previous 20 days contain more than 10 days without PD.

$Q^c_d = \sum_{j=d_i}^{d_{24}} q_{h_j}$:

The sum of the charge of the measured PD of cluster c during day d, in picocoulomb (pC);

$T_d = \frac{1}{24} \sum_{j=i}^{24} t_{h_j}$:

The mean of the hourly temperatures during day d, in degrees Celcius;

$\rho^c_{d_i} = Pearson([Q^c_{d_{i-19}}, ..., Q^c_{d_i}], [T_{di-19}, ..., T_{d_i}])$:

The correlation coefficient of the 20 daily values prior to and including day d_i of cluster c;

$\rho^c = [\rho_{d_{20}}, ..., \rho_{d_{max}}]$:

Rolling window of cluster c.

- *max_timeperiod_of_consecutive_positives*: $\max\{\lambda \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+\lambda-1}} \geq 0.7\}$,
  The longest period of the cluster in which the correlation coefficient for each day of this period is at least 0.7.

- *max_timeperiod_of_consecutive_negatives*: $\max\{\lambda \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+\lambda-1}} \leq -0.7\}$,
  The longest period of the cluster in which the correlation coefficient for each day of this period is at most -0.7;

- *max_corr_which_repeats_12_timeperiods*: $\max\{\mathrm{P} \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+11}} \geq \mathrm{P}\}$,
  The maximum correlation coefficient for which there are at least 12 consecutive days for which the corresponding correlation coefficient is at least this value.

- *min_corr_which_repeats_12_timeperiods*: $\min\{\mathrm{P} \mid \exists i : \rho_{d_i}, ..., \rho_{d_{i+11}} \leq \mathrm{P}\}$,
  The minimum correlation coefficient for which there are at least 12 consecutive days for which the corresponding correlation coefficient is at most this value.

For the next 4 features the rolling frame is based on the *residue* instead of the temperature. In short the residue is the difference between the trend of the temperature and the actual temperature. It represents the fluctuations of the temperature.

Residue is being created using the 6-hourly soil temperatures from De Bilt (depth of 50cm). For 2018 the difference between the mean of the year and the other years is being calculated to get $3 * 40$ values. Then the difference between the mean of 1980 and 2018 is added to the 6-hourly temperatures values from 1981. This will be done for each year in 1980-2020. Now the means of all years are equal to the mean of 2018. The mean of all first values (the first measurement which is 01-01 06:00:00) is taken and assigned to the first value of 2018. This will be done for each period of 2018. And the process is repeated for 2019 and 2020. The period January and February 2021 is created by making a copy of January and February 2020. The residue is for each value the difference between this and the actual temperature.

R_d : The residue of day d :
The difference between the trend and the actual temperature;

$\rho'_{d_i} = Pearson([Q^c_{d_{i-19}}, ..., Q^c_{d_i}], [R_{di-19}, ..., R_{d_i}])$:
The correlation coefficient of the 20 daily values prior to and including day d_i of cluster c;

$\rho' = [\rho'_{d_{20}}, ..., \rho'_{d_{max}}]$:
Rolling window of cluster c.

- *max_timeperiod_of_consecutive_positives_residue*: $\max\{\lambda \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+\lambda-1}} \geq 0.7\}$;

- *max_timeperiod_of_consecutive_negatives_residue*: $\max\{\lambda \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+\lambda-1}} \leq -0.7\}$;

- *max_corr_which_repeats_12_timeperiods_residue*: $\max\{\mathrm{P} \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+11}} \geq \mathrm{P}\}$;

- *min_corr_which_repeats_12_timeperiods_residue*: $\min\{\mathrm{P} \mid \exists i : \rho'_{d_i}, ..., \rho'_{d_{i+11}} \leq \mathrm{P}\}$.

16 features about the relation between PD and seasons

The sum of the charge of PD of the cluster in a certain period, divided by the total PD in this cluster. The year is divided into periods:

- *Charge_in_cold_6_months_percentage*: November - April

- *Charge_in_hot_6_months_percentage*: May - October

- *Charge_rising_temp_percentage*: February - July

- *Charge_decreasing_temp_percentage*: August - January

- *Charge_in_winter_percentage*: November - January

- *Charge_in_spring_percentage*: February - April
- *Charge_in_summer_percentage*: May - July
- *Charge_in_fall_percentage*: August - October

The next 8 features are similar. Only the dataset of PD is different: The sum of the charge at the median of this cluster with a bandwidth of the distance between the median and the 5th percentile in a certain period, divided by the total PD in this cluster. This concerns the entire period in which the circuit is being monitored.

- *Charge_in_cold_6_months_percentage_all_circuit*: November - April
- *Charge_in_hot_6_months_percentage_all_circuit*: May - October
- *Charge_rising_temp_percentage_all_circuit*: February - July
- *Charge_decreasing_temp_percentage_all_circuit*: August - January
- *Charge_in_winter_percentage_all_circuit*: November - January
- *Charge_in_spring_percentage_all_circuit*: February - April
- *Charge_in_summer_percentage_all_circuit*: May - July
- *Charge_in_fall_percentage_all_circuit*: August - October

8 features about the charge of the PD

For each minute of the cluster the 99.9th percentile of the charge of the previous 10080 minutes (1 week) is assigned to it. The longest consecutive period in which all minutes have a value higher than 20000 pC determines the feature. The value is the number of minutes of this period divided by $\frac{10080}{7} = 1440$, to get the number of days.

- *number_of_days_more_than_20000_pC*
- *number_of_days_more_than_20000_pC_all_circuit*: This deals with the PD at the median of this cluster with a bandwidth of the distance between the median and the 5th percentile, during the entire period in which the circuit is being monitored.

Fitting the best Weibull distribution to the distribution of the charge of the PD, results in 6 features: 2 parameters of the Weibull distribution and 4 criteria for the goodness of fit. The function Fit_Weibull_2P from reliability.Fitters is used for this. More information can be found on reliability.readthedocs.io.

- *weibull_scale*: Scale of the Weibull distribution
- *weibull_shape*: Shape of the Weibull distribution
- *goodness_of_fit_loglik*: Log-Likelihood
- *goodness_of_fit_AICc*: Akaike Information Criterion
- *goodness_of_fit_BIC*: Bayesian Information Criterion
- *goodness_of_fit_AD*: the Anderson-Darling goodness of fit statistic

List of terms

Circuit
Group of connected electricity cables, joints and RMUs monitored by a SCG-system, by means of a SCG-device on each side of the circuit. The placement of the measuring devices defines the circuits. i, 2, 4, 10–13, 15, 17, 19–23, 25, 28, 30, 32, 73, 75, 79, 80

Circuit number
The number of the circuit. Each circuit has a unique number assigned to it to differentiate them. 10, 11, 13, 14, 22, 79

Circuitlength
The circuitlength is the length of the circuit: the cumulative length of the cables of the circuit. This is equal to the location of the slave unit. See also Figure 9(a). 11, 19, 21, 25, 55, 59, 76, 79

Climate Data Store (CDS)
The CDS provides information about the past, present and future climate, on the global, continental, and regional scale [18]. 15, 18, 76, 80

Cluster
A group of particles, bundled by the cluster algorithm. i, 1, 2, 19, 20, 73–75

Cluster algorithm
The cluster algorithm is an algorithm developed by Alliander which clusters the particles measured by . The particles which are partial discharges and have the same source are put into one cluster. All other particles are filtered and considered noise. i, 19, 60, 79

Det Norske Veritas (DNV)
Det Norske Veritas which translates to "The Norwegian Truth", provides digital solutions for managing risk and improving safety and asset performance for ships, pipelines, processing plants, offshore structures, electric grids, smart cities and more [2]. It provides Smart Cable Guard (SCG) to Alliander to monitor its electricity grid. 4, 11, 12, 25, 27, 71

Fault
Faults are short-circuit currents. These usually lead to circuit breakdown. i, 1–4, 10–13, 15–23, 25, 26, 28, 29, 33, 62, 71, 73–76, 79, 80, 83

Faultcount
Faultcount is the number of large sparks that occur during the minute of the fault. One fault can consist of multiple detections. If a short circuit occurs in a cable, multiple blows (sparks) can occur. SCG combines these per minute. So if there are multiple detections per minute, a fault gets a faultcount of more than 1. 10, 13

Faultgroup
Some faults belong to the same faultgroup. Fault grouping is a way to group multiple faults, based on location and time. We give faults the same *faultgroup* if they have the same circuit number and location is within and including ±2% of the circuitlength and their time is within 180 days.
For example, when a fault occurred on one specific circuit at 2020-10-01 00:00:00, at location 30% and then another fault, on that circuit, occurred on 2020-10-05 00:00:00, at location 32%, then those faults will have a same fault group id. 11, 13

Joint
Joints are the parts of a circuit that connect the cables. The joints are usually the weak spots of the circuits, depending on the material of the joints. i, 1, 3, 4, 33, 73, 75, 79

Koninklijk Nederlands meteorologisch instituut (KNMI)
Dutch national weather forecasting service [23]. 15, 80

Master unit
The master unit is the SCG-device that is placed in the RMU at the start of the circuits. 3, 4, 80

Masterframe
Huge dataframe with all clusters and its features. The masterframe is used to predict faults.. 19–23, 26, 28, 29, 75

Partial discharge (PD)
A partial discharge is a small charge displacement in the cavity or layer of the insulation of a component. PD is a good predictor of faults. i, 1, 2, 4, 10, 19, 20, 22, 23, 28–30, 32, 33, 55, 71, 73–78, 80

Particle
Umbrella term of PD and noise. The observations of SCG when it measures PD. 19–22, 26, 28, 76, 79

Ring Main Unit (RMU)
A ring main unit is a set of switchgear used at the secondary substations. The SCG-devices are placed in the ring main units. 3, 4, 11, 12, 79, 80

SCG-device
A SCG-device is one of the two measuring devices of a SCG-system. It is either the master unit or the slave unit. 3, 4, 79, 80

SCG-system
A SCG-system consists of two SCG-devices: the master unit and the slave unit. These SCG-device are placed in the RMUs at the end of a circuits, to measure the faults and PD on the circuits. 1–4, 10, 12, 16–18, 75, 79, 80

Slave unit
The slave unit is the SCG-device that is placed in the RMU at the end of the circuits. 3, 4, 79, 80

Smart Cable Guard (SCG)
Smart Cable Guard is a sensor-based digital monitoring platform that puts owners in control of their medium voltage cable network. Combining patented technology with 24/7 monitoring and support, it detects and locates faults and weaknesses in underground cables [3]. i, 1, 10, 12, 17, 19, 30, 68, 71, 75, 79, 80

Substation
A substation is a station that transforms the power from high voltage to medium voltage. A secondary substation transforms the power from medium voltage to low voltage. 3, 80

Weather application programming interface
An interface to retrieve data about the weather from both the KNMI and the CDS [15]. 10

References

[1] Alliander N.V. *Annual report 2019 - working together on transition*. Accessed: 2020-01-27. URL: https://2019.jaarverslag.alliander.com/FbContent.ashx/pub_1037/downloads/v200320083558/Alliander_Annual_Report_2019.pdf.

[2] DNV. Accessed: 2021-06-17. URL: https://www.dnv.com/about/index.html.

[3] DNV. *SCG*. Accessed: 2021-05-25. URL: https://www.dnv.com/power-renewables/services/scg/technology.html.

[4] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[5] Plotly Technologies Inc. *Collaborative data science*. Accessed: 2021-7-12. 2015. URL: https://plot.ly.

[6] Wagenaars P. *Integration of online partial discharge monitoring and defect location in medium-voltage cable networks*. PhD thesis, Department of Electrical Engineering, 2010. DOI: doi.org/10.6100/IR656994.

[7] Van Minnen F. Harmsen D. and Wagenaars P. *On-line Fault Detection to Prevent Network Failure*. Accessed: 2021-6-29. 2018. URL: https://www.tdworld.com/intelligent-undergrounding/article/20971385/online-fault-detection-to-prevent-network-failure.

[8] Pearson K. "Notes on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58 (1895), pp. 240–242.

[9] Kendall M. "Note on the distribution of quantiles for large samples". In: *Supplement to the Journal of the Royal Statistical Society* 1.7 (Aug. 1902). ISSN: 83-85. DOI: 10.2307/2983633.

[10] Galton F. "The most suitable proportion between the values of first and second prizes". In: *Biometrika* 1.4 (Aug. 1902). ISSN: 0006-3444. DOI: 10.1093/biomet/1.4.385.

[11] Galton F. "Some results of the Anthropometric Laboratory". In: *Journal of the Royal Anthropological Institute* 14 (1885). ISSN: 275-287. DOI: 10.2307/2841985.

[12] pandas. *pandas quantile*. Accessed: 2021-6-15. URL: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Series.quantile.html.

[13] Weibull W. "A statistical distribution function of wide applicability". In: *Journal of Applied Mechanics* 18.3 (1951), pp. 293–297.

[14] Reliability HotWire. *Characteristics of the Weibull Distribution*. Accessed: 2021-7-29. URL: https://www.weibull.com/hotwire/issue14/relbasics14.htm.

[15] Alliander. *Weather application programming interface*. Accessed: 2021-6-29. URL: https://weather.appx.cloud/api/v2/docs.

[16] KNMI. *Daggegevens*. Accessed: 2021-5-5. URL: https://www.knmi.nl/nederland-nu/klimatologie/daggegevens.

[17] KNMI. *Uurgegevens*. Accessed: 2021-5-5. URL: https://www.knmi.nl/nederland-nu/klimatologie/uurgegevens.

[18] Climate Data Store. Accessed: 2021-5-5. URL: https://cds.climate.copernicus.eu.

[19] KNMI. *Dagwaarden neerslagstations*. Accessed: 2021-5-5. URL: https://www.knmi.nl/nederland-nu/klimatologie/monv/reeksen.

[20] KNMI. *Metadata De Bilt*. Accessed: 2021-5-5. URL: http://projects.knmi.nl/klimatologie/metadata/debilt.html.

[21] KNMI. *Bodemtemperaturen*. Accessed: 2021-5-5. URL: https://www.knmi.nl/nederland-nu/klimatologie/bodemtemperaturen.

[22] KNMI. *Temperatuur*. Accessed: 2021-5-5. URL: https://www.knmi.nl/kennis-en-datacentrum/uitleg/temperatuur.

[23] KNMI. *Klimatologie*. Accessed: 2021-5-5. URL: https://www.knmi.nl/nederland-nu/klimatologie.

[24] Data-Flair. *T-table*. Accessed: 2021-7-29. URL: https://data-flair.training/blogs/t-test-in-qlikview/.

Appendix

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Table 13: t-table to determine the p-value for the t-test [24]

	Feature	Absolute 12-2014	Relative 12-2014	Absolute 8-2017	Relative 8-2017	Absolute 1-2018	Relative 1-2018
0	T10N	0.29	0.05	0.34	0.41	0.4	0.44
1	temperature_min	0.29	0.05	0.35	0.41	0.4	0.44
2	precipitation_max	0.27	0.06	0.38	0.4	0.43	0.43
3	temperature	0.29	0.04	0.34	0.4	0.36	0.42
4	temperature_max	0.28	0.03	0.32	0.38	0.34	0.39
5	soil_temperature_level_1	nan	nan	nan	nan	0.27	0.39
6	soil_temperature_level_2	nan	nan	nan	nan	0.27	0.39
7	2m_dewpoint_temperature	nan	nan	nan	nan	0.28	0.39
8	2m_temperature	nan	nan	nan	nan	0.26	0.39
9	humidity_min_hour	0.13	0.01	0.17	0.3	0.19	0.35
10	soil_temperature_level_3	nan	nan	nan	nan	0.25	0.35
11	EV24	0.24	0.01	0.29	0.35	0.26	0.34
12	air_pressure_min_hour	0.14	-0.08	0.28	0.27	0.33	0.33
13	global_radiation	0.22	0	0.26	0.31	0.22	0.29
14	surface_solar_radiation_downwards	nan	nan	nan	nan	0.14	0.28
15	surface_solar_radiation_downward_clear_sky	nan	nan	nan	nan	0.15	0.28
16	VVX	0.22	-0.13	0.23	0.27	0.19	0.26
17	wind_gust_max_hour	0.18	-0.21	0.22	0.22	0.23	0.24
18	soil_temperature_level_4	nan	nan	nan	nan	0.19	0.22
19	sunlight_duration	0.2	-0.03	0.22	0.26	0.16	0.22
20	total_precipitation	nan	nan	nan	nan	0.04	0.2
21	precipitation	0.04	0.01	0.09	0.11	0.2	0.19
22	wind_direction	-0.06	-0.08	-0.03	0.11	0.05	0.18
23	VVXH	-0.02	-0.01	0.04	0.13	0.03	0.14
24	VVN	0.16	-0.18	0.12	0.18	0.03	0.12
25	temperature_min_hour	-0.05	0.18	0.01	0.05	0.08	0.12
26	percentage_of_max_possible_sunlight_duration	0.15	-0.04	0.15	0.16	0.07	0.11
27	precipitation_max_hour	0.03	0.01	0.13	0.04	0.2	0.09
28	temperature_max_hour	0.05	-0.02	0.12	0.08	0.09	0.09
29	wind_speed_max_hour	0.05	-0.12	0.09	0.08	0.07	0.07
30	10m_u_component_of_wind	nan	nan	nan	nan	0.04	0.06
31	100m_u_component_of_wind	nan	nan	nan	nan	0.04	0.04
32	10m_v_component_of_wind	nan	nan	nan	nan	0.12	0.02
33	100m_v_component_of_wind	nan	nan	nan	nan	0.12	0.02
34	T10NH	-0.07	0.17	-0.01	-0.05	0.05	0
35	air_pressure_max_hour	-0.02	0.09	-0.02	-0.03	0.01	0
36	wind_gust_max	0	-0.16	0.03	0.02	0	-0.01
37	humidity_max_hour	0.02	0.01	0.09	-0.01	0.09	-0.03
38	humidity_max	-0.19	0.16	-0.14	-0.11	-0.04	-0.04
39	air_pressure_min	-0.1	0.08	-0.03	-0.05	-0.05	-0.07
40	mean_sea_level_pressure	nan	nan	nan	nan	-0.12	-0.09
41	surface_pressure	nan	nan	nan	nan	-0.12	-0.09
42	air_pressure	-0.13	0.06	-0.07	-0.08	-0.09	-0.1
43	cloud_cover	-0.03	0.01	-0.14	-0.13	-0.08	-0.1
44	air_pressure_max	-0.17	0.04	-0.12	-0.12	-0.14	-0.14
45	volumetric_soil_water_layer_4	nan	nan	nan	nan	-0.14	-0.15
46	precipitation_duration	-0.19	0.07	-0.16	-0.2	-0.1	-0.15
47	wind_speed_max	-0.11	-0.12	-0.07	-0.09	-0.14	-0.15
48	wind_speed_min_hour	-0.11	0.02	-0.14	-0.21	-0.09	-0.18
49	humidity_min	-0.2	0.08	-0.2	-0.23	-0.14	-0.19
50	wind_speed	-0.12	-0.13	-0.1	-0.14	-0.16	-0.2
51	humidity	-0.23	0.11	-0.23	-0.26	-0.16	-0.21
52	wind_speed_min	-0.14	-0.1	-0.15	-0.18	-0.18	-0.22
53	VVNH	-0.17	0.09	-0.12	-0.25	-0.1	-0.23
54	volumetric_soil_water_layer_2	nan	nan	nan	nan	-0.17	-0.26
55	volumetric_soil_water_layer_3	nan	nan	nan	nan	-0.17	-0.26
56	volumetric_soil_water_layer_1	nan	nan	nan	nan	-0.16	-0.27
57	FHVEC	-0.17	-0.12	-0.17	-0.23	-0.23	-0.29

Table 14: Correlation coefficient between absolute/relative number of faults and several weather variables for three periods

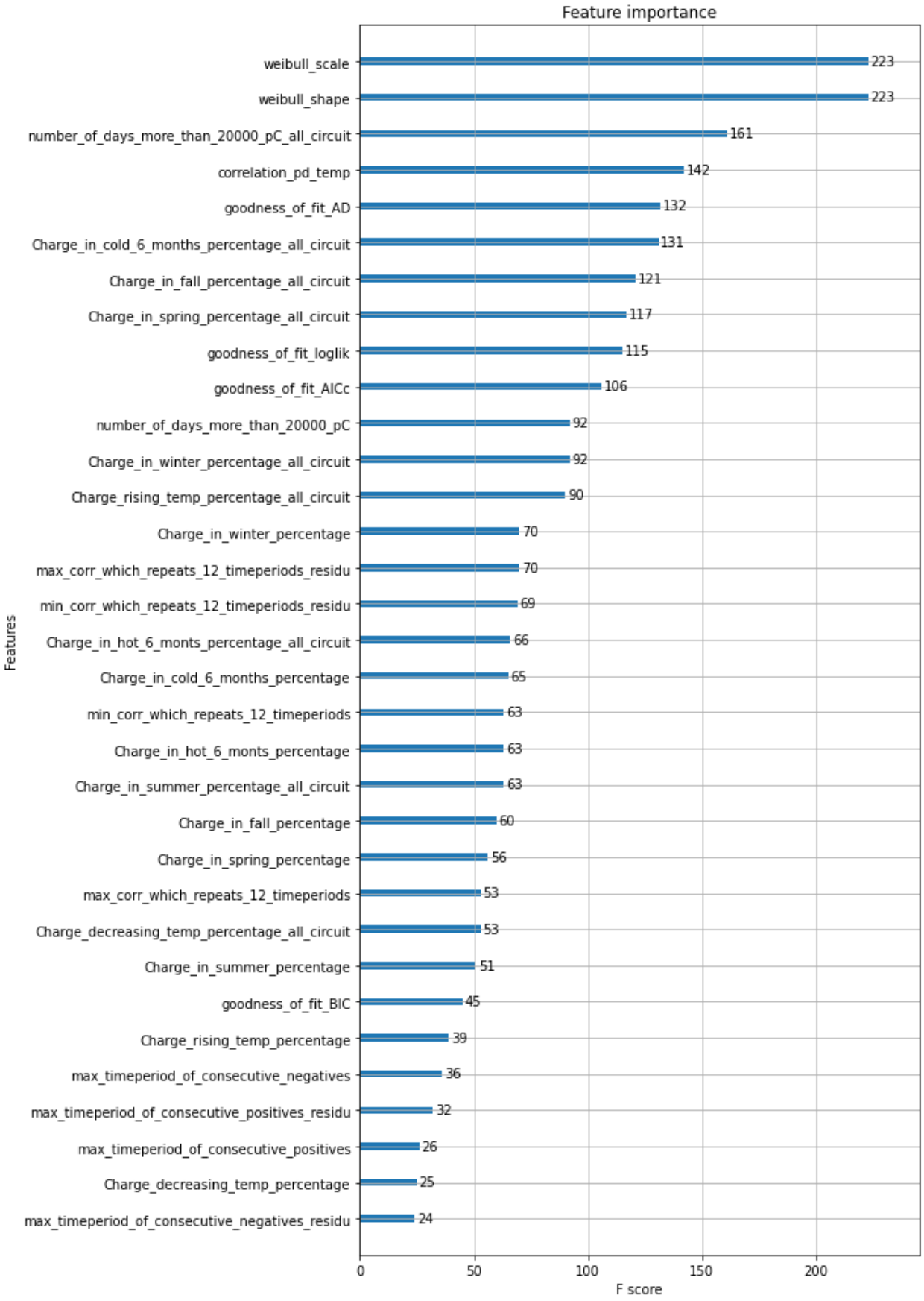


Figure 68: Feature importance of the 33 new features

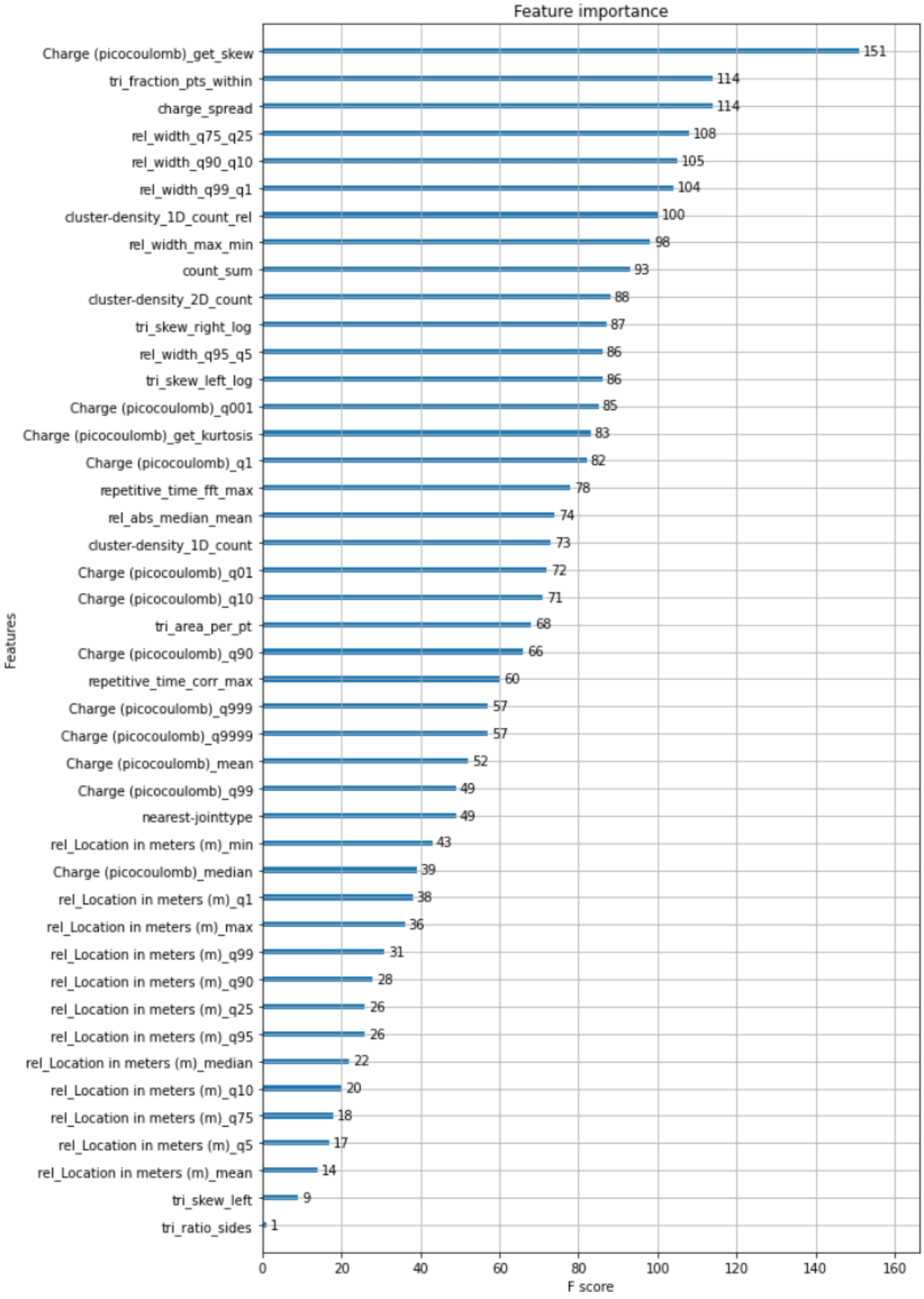


Figure 69: Feature importance of Alliander's features

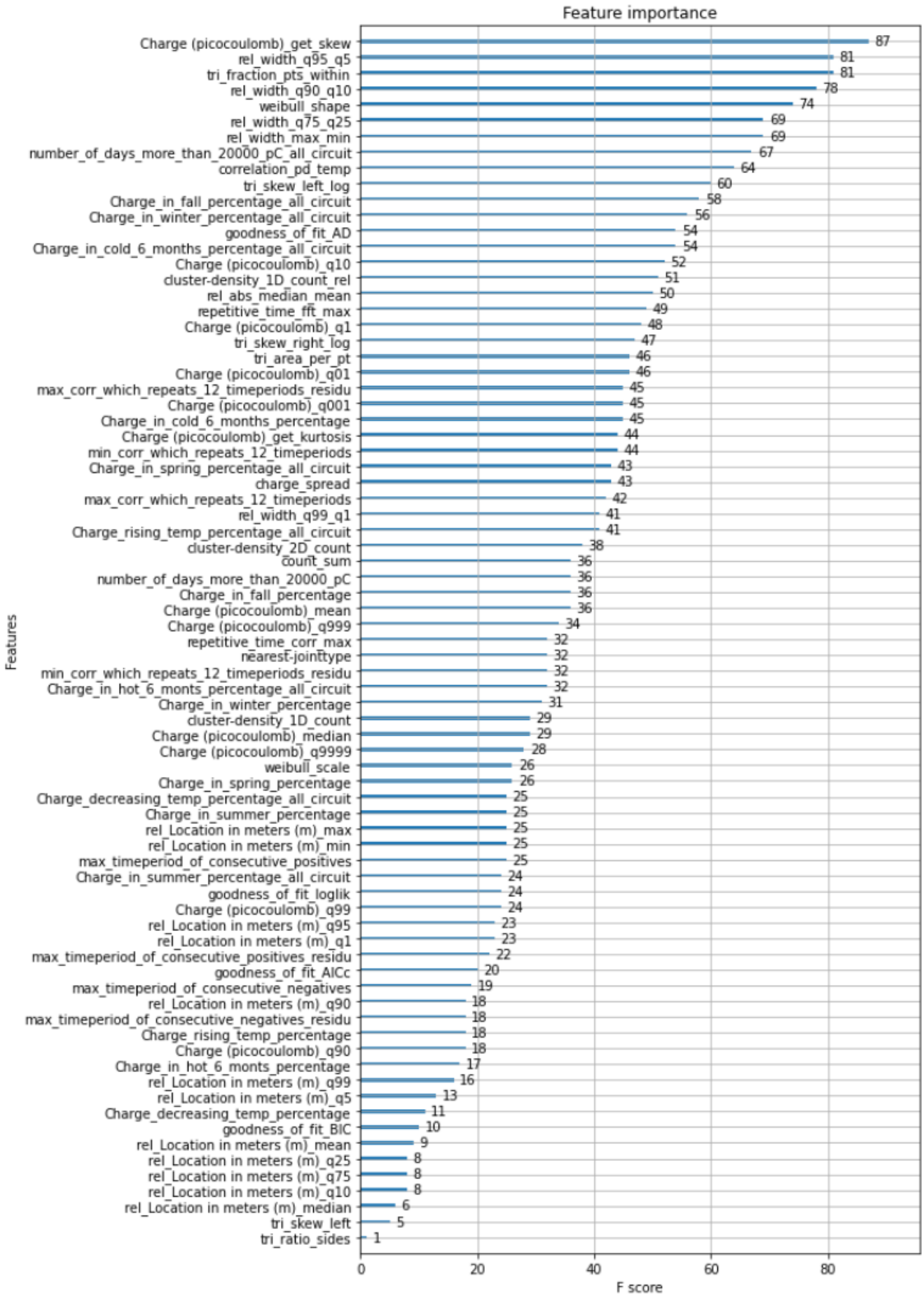


Figure 70: Feature importance of all features