Unsupervised Learning to Locate Weak Spots in the Medium Voltage Grid

Master's Thesis

Devendra Kulkarni

Thesis Committee:

Prof. Dr. Peter Palensky	Professor
Dr. Simon Tindemans	Daily Supervisor
Dr. Armando Rodrigo Mor	External Committee
Dr. Sander Rieken	Company Supervisor

Delft, 18 November 2020

Abstract

The Medium Voltage (MV) network in the Netherlands is almost entirely composed of underground cables. The failure statistics show that the interruptions in the medium voltage have a high contribution in the mean outage time per year per customer. The interruptions in the medium voltage grid are often caused by failure of circuit components mainly the cables, joints, and terminations which account for 73% of the failures. The occurrence of these interruptions and their duration can be limited by proper maintenance measures.

Partial discharge diagnostics provide a way for the condition assessment of the circuit insulation. The Smart Cable Guard (SCG) systems from DNV GL help in continuous monitoring of the medium voltage cable circuits during the circuit operation and aid the network operators in maintaining the MV grid. With its non-intrusive monitoring characteristic, we see that the SCG systems provide the measurement of partial discharges (PD) occurring in the cable circuit. Moreover, the partial discharge measurements consists of the information about the location of each observed discharge event along with the time of its occurrence and its corresponding discharge magnitude observed in picocoulombs for every minute. The partial discharge events occurring in the circuit are subjected to evaluation from an expert at DNV GL control rooms who qualitatively assigns a warning level (Level 1, 2, 3, or Noise) to the observed pattern of discharge events. This manual process of evaluation of PD events and assignment of warnings to them is a laborious task and the network operator has to rely on the availability of the expert and his/her accurate assessment. To reduce this dependence for the network operator, Alliander in collaboration with the DNV GL experts is developing an automated decision support tool to identify the partial discharge events and to aid the operator and the expert in evaluating the condition assessment of the cable network.

This thesis proposes a clustering methodology using the ST-DBSCAN density-based technique for identifying high-density discharge events or 'areas of interest' in the PD data obtained from the SCG systems. The clusters identified from the method are further evaluated by extracting their features or characteristics using the PD data attributes as well as describing their characteristics based on the context of the circuit. The performance of the clustering method is validated using the DNV GL warnings by formulating performance indicators and metrics to measure the performance. The clustering method along with the features extracted from the cluster contribute towards the development of the automated decision support tool.

Acknowledgements

First and foremost, I would like to thank Dr. Simon Tindemans for supervising me and supporting me during the course of this thesis. I would also like to thank Dr. Sander Rieken and the SCG Analytics team at Alliander for allowing me to conduct my graduation thesis at the company and supporting me with my technical queries.

I would like to thank Prof. Dr. Peter Palensky and Dr. Armando Rodrigo Mor for agreeing to be a part of my graduation committee.

I would like to especially thank my friends Yashasvi and Sneha for their constant support in this thesis and tolerating me during my frustrations and also accompanying me during my bursts of excitements. I really will cherish all the enriching conversations that we have had during this time.

This thesis started at the onset of the dreadful coronavirus and we were all doomed and locked up in our rooms for what seems like an eternity now. But these months of monotony and lockdown were made cheerful by friends through video calls who I am immensely grateful for and would like to thank them for supporting me throughout these months.

Last but not the least, I would like to thank my parents and sister: Aai, Baba and Mugdha for inspiring me, guiding me and supporting me in all my endeavours.

Contents

Chapter 1

Introduction

Following the 1998 Electricity Act [1], the electric power system in the Netherlands underwent several changes - from the decoupling of the transmission and distribution operations to market liberalization and introduction of a regulatory authority to monitor the operations and the market and cater to the consumer interests. This resulted in separation of vertically integrated companies and allowed the consumer to exercise freedom of choice in selecting the electricity supplier thus encouraging competition of supply and efficiency.

Although the consumer could have a choice in deciding their energy supplier, it was not possible for the consumer to have a choice in deciding their grid operator. The grid operators are therefore a natural monopoly [2]. The electric power industry is divided into three areas of operation, namely, generation, transmission and distribution, and supply of electricity. The transmission responsibility now lies with Tennet, Transmission System Operator (TSO) of the Netherlands, which transmits energy from generation points to regional distribution network operators. There are a total of eight distribution network operators with Enexis, Liander, and Stedin providing services to most of the country. The division of responsibilities between the transmission and distribution operators were done on different grid voltage levels with the Extra High Voltage (EHV), Ultra High Voltage (UHV) and High Voltage (HV) network was maintained and operated by the Transmission System Operator (TSO) and Medium Voltage (MV), Low Voltage (LV), and part of the HV network was handled by the Distribution System Operators (DSO).

The Authority for Consumers and Markets (ACM) operates as the regulatory authority with its main tasks of ensuring that all suppliers have access to the transmission and distribution infrastructure with similar financial and technical conditions so as to achieve a fair and secure functioning of the network alongwith exercising tariff regulation for the usage of the infrastructure. The duties of both Transmission and Distribution (T&D) operators include (i) operation and maintenance of the network, (ii) guaranteed transport of electricity in a safe and reliable manner, (iii) construction, repair and expansion of networks, (iv) planning and maintaining sufficient capacity, (v) access to new connections without discrimination, and (vi) promoting safe use of electricity [3].

Reliable operation of power systems has always been of utmost importance to ensure there is uninterrupted and secure supply of energy to consumers. Utilities often have to deal with aging infrastructure and need to maintain and replace old equipment every now and then. Due to high cost of components, utilities are constantly striving to optimize the use of resources for maintenance activities while safeguarding system reliability within adequate limits [4]. The regulator expects utilities to provide acceptable levels of service for lowest possible rates and also expects data-driven spending budget reports to ensure credibility of the operator [5]. To comply with the standards of the regulation, utilities adopt an Asset Management (AM) framework

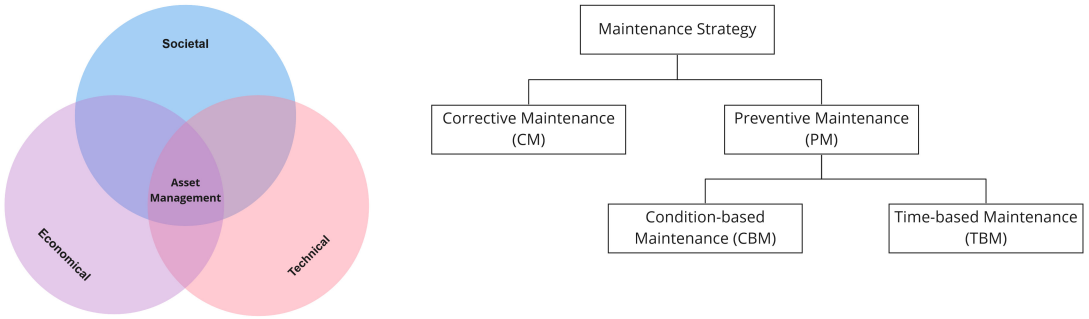

Figure 1.1: Three categories of information aspects of electrical asset management and Asset Maintenance Strategy [6]

to achieve a balance in the technical, financial and societal aspects of the engineering and infrastructure needs for a reliable energy supply.

In [6], a model is presented that describes the societal, technical, and economical information aspects that go into assessment of the status of asset (Figure 1.1). The decision process for asset management is divided into three categories :-

- *Technical Aspect* takes into account the asset related parameters such as the physical condition of the asset and maintenance activities for the ageing of components, failure probability of the component, and inventory and maintenance activity planning.
- *Economical Aspect* takes into the financial aspect of the maintenance cost and other costs related to the replacement and procurement of components and costs due to failure.
- *Societal Aspect* evolved from the need to take into account the failure acceptability of the society wherein some institutions like hospitals and public places would suffer gravely due to unreliable supply of electricity and hence the social and economical impact would be high.

Considering the technical aspect of asset management it can be further extended for employing maintenance strategies for condition assessment of components. The maintenance strategy for the performance of electrical assets is categorized into *Corrective Maintenance (CM)* which aims to restore performance after a failure through repairing the faulty component, and *Preventive Maintenance (PM)* which aims to reduce the failure probability of the component by inspecting or replacing the component before it fails. In [7], a framework is developed based on preventive maintenance called reliability centered maintenance (RCM) for asset management providing statistical relationship between PM of assets and total maintenance costs for distribution systems.

PM activities can be planned as scheduled or time-based maintenance activity (TBM) for a routine inspection of the components or as a condition-based maintenance (CBM) where the activity is performed based on the condition of the component monitored through sensors. An advantage with the TBM activity is that all the components are checked at regular time schedules but poses a risk if the time schedules of inspection are not optimized which can lead to failure of the component. CBM activity involves the continuous collection and interpretation of data of the components and determining the initiation of failure modes and subsequent time of failure to update the decision process of maintenance strategy such as initiation of repair activity and priority of the activity [8].

Condition based maintenance includes the process of (i) data collection through condition monitoring of the component and historical data available for the component, (ii) handling of the acquired data into meaningful abstract representation, (iii) utilizing the knowledge for failure prediction, (iv) and finally decision making (repair, replace or run-to-fail) based on the assessment of the failure risk. With the introduction of smart sensing technology being installed by utilities, collecting data has become increasingly important in supporting organizational decisions along with the decision making process. The development of data analytics in the form of data mining, statistical modelling and machine learning techniques help in better failure prediction and can bridge the gap between short-term corrective work and long-term capital planning for utilities [9].

1.1 Challenge

Owing to the increase in demand of high capacity connections of solar farms and data centres has accelerated the need for updating the network and its expansion to accommodate these demands and avoid failure of the grid. With these updates in line, the network operator also has to ensure a reliable supply of electricity at all times. Several reliability indices are calculated to evaluate the interruptions observed in the grid. In Figure 1.2, the annual outage duration due to interruptions in the LV, MV, HV, and EHV network is shown. The annual outage duration is a commonly used indicator to evaluate the degree of reliability. It is evident that the interruptions in the medium voltage network have a larger share in contributing the annual outage time. The interruptions are mainly caused due to failure of components in the network. In the Netherlands, the MV network is almost entirely composed of underground cables.

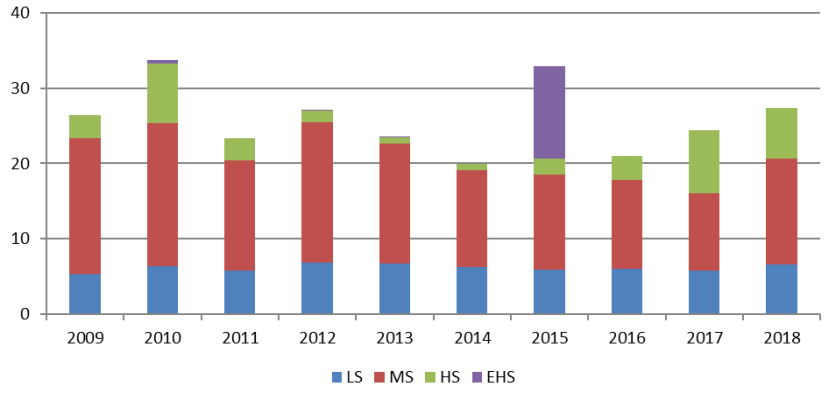

Figure 1.2: Annual outage time per customer per network area [min/year], (2009 – 2018) [10]

Figure 1.3 illustrates the failure statistics in the MV grid per component. It is evident from the figure that majority of the interruptions in the MV grid are caused due to failure in the circuit of the MV grid which comprises of the cables, joints, and termination components. The bar chart (Figure 1.3) illustrates the cause of failure in the circuit components. The causes of failures are mainly due to ageing of the component, damages due to digging activities, and internal defects in the insulation of the component. The proportion of failure in joints due to internal defect is prominent and contributes to the majority of interruptions in the MV network.

Aging mechanisms like thermal degradation, electrical stress, mechanical stress, and environmental conditions are often the cause of failure in cable networks and alter its insulation property. These changes in the insulation property observed over time help in identifying the

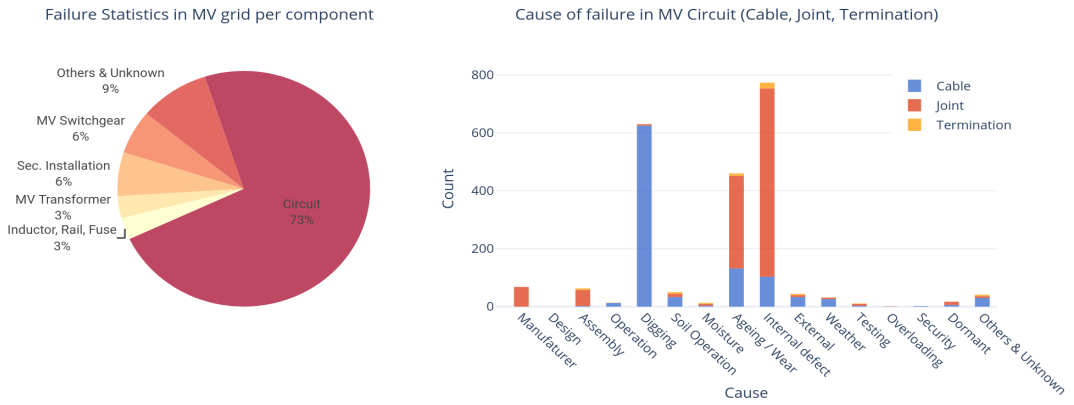

Figure 1.3: Failure statistics in the MV Grid [10]

process and state of deterioration. Partial Discharge (PD) activity is one such mechanism associated with degradation of the insulation integrity of the component. Due to thermal aging, cracks may appear in the insulation which may lead to a rise in PDs. Several tools have been developed to monitor partial discharge activity.

1.2 Company Profile

This thesis was carried out with the collaboration and support of Alliander, a Dutch distribution system operator (DSO). Alliander is among the biggest distribution network operators of the Netherlands and extends its service in the operation and maintenance of gas and electricity grid across the country's provinces of Amsterdam, Friesland, Gelderland, North Holland, Flevoland, and parts of South Holland (Figure 1.4).

With the total electricity grid length of 91,000 km, Alliander serves close to 5.8 million customers in the Netherlands. The operator supports several types of loads and feed-in generation on the grid. Typical loads can be categorized into residential, industrial, and data centres. With the growing awareness about climate change, the need for energy transition has led to increase in local energy generation and consumption. Residential rooftop solar and solar/wind energy parks provide a cleaner and affordable alternative. There has also been an increase in the adoption of electric transport and its subsequent need for charging infrastructure. This increase in the local generation capacity and increased demand causes bottlenecks (or congestion) in the grid and the operator needs to solve these bottlenecks by upgrading the network and applying innovative solutions such as the digitalisation of the grid for improved operation.

One such digital upgrades undertaken by Alliander to improve the operational efficiency of the network is the installation of Smart Cable Guard (SCG) for medium voltage cable networks to continuously monitor and detect partial discharge activity in the cable network. The SCG system detects weak points in the underground cable network before they might lead to outages. Alliander has over 1500 SCG systems monitoring close to 6000km of the MV cable network that contain approximately 35,000 joints. The SCG system is described in detail in the following chapter.

Figure 1.4: Geographical overview of the service area for electricity network of Alliander [11]

1.3 Research Objective

1.3.1 Problem Definition

Partial discharge activity in cables is a widely known diagnostic method to assess the condition of circuit components. For a long time, the condition of insulation of the components due to incipient partial discharges was identified through offline investigation of the cable. With the introduction of SCG systems in the cable network, online monitoring of the cable condition has become possible and can provide early warnings in the deterioration of cable components due to PD activity. To assess the condition of the cable, the utility heavily relies on the analysis of human experts due to their extensive knowledge gained through the PD behaviour observed from offline tests. The knowledge rules derived through years of offline investigations are often not transferable for assessing the data collected from online systems and are also labour intensive to provide accurate assessment [12].

Continuous monitoring through SCG provides a vast data stream for every circuit which needs to be correctly interpreted in order to extract useful information on the status of the circuit and its components. Often the data collected from the online monitoring tool contains a lot of noise detected from neighbouring radio signals or proximity to industry or solar/wind parks. The partial discharge events occurring in the circuit are subjected to evaluation from an expert at DNV GL control rooms who qualitatively assigns a warning level (Level 1, 2, 3, or Noise) to the observed pattern of discharge events. This manual process of evaluation of PD events and assignment of warnings to them is a laborious task and the network operator has to rely on the availability of the expert and his/her accurate assessment. To reduce this dependence for the network operator, Alliander in collaboration with the DNV GL experts is developing an automated decision support tool to identify the partial discharge events and to

aid the operator and the expert in evaluating the condition assessment of the cable network.

This research focuses on the interpretation of the data acquired from the SCG monitoring tool through elimination of noise points and identifying 'hubs' or clusters of high density discharges in the circuit. Further statistical analysis on the identified clusters would help in preparing knowledge rules through meaningful features of the clusters. For the scope of this thesis, two main research objectives were identified:-

RO 1: *How can we identify high-density discharge events in continuously monitored, noisy PD activity?*

The SCG systems connected at both ends of the circuit, continuously measure the discharges occurring at various locations in the circuit. Disturbances from surroundings of the circuit such as proximity to industry or solar/wind parks, construction activities, railway tracks running along the circuit etc. are also sometimes picked up by the SCG as discharges. PD activity is also intermittent in nature and can occur during high loading cycles, increase in temperature in the conductors etc. This research objective aims at identifying high-density activity occurring in the circuit and filter the noise measurements from the received SCG data.

RO 2: *How can we describe relevant events from background noise?*

The identified high-density activity is evaluated by calculating features which will help in describing the activity as well as help in understanding partial discharge events from noisy events. The attributes of the SCG data such as the location of the activity, timestamp of the occurrence of the activity and discharge magnitude are used to describe the high-density activity identified.

1.3.2 Research Outline

For the above research objectives the following tasks were formulated:-

- Implementation of a clustering method to identify high-density discharge events in the SCG data
  - Identify high density clusters using the spatial and temporal attributes of the data.
- Evaluate performance of the clustering method
  - Validate the clusters identified with the DNV GL warnings assigned for the circuit.
- Extract quantifiable features from the identified clusters
  - Density of the cluster, charge distribution of the cluster, calculating the inter-arrival times of the discharge events occurring inside the cluster
  - Calculate the statistical parameters of the cluster properties such as the mean, median, skew, and kurtosis.

1.4 Thesis Outline

This thesis report is structured in the following manner:

In Chapter 2, a detailed background of the medium voltage network and the causes of failure in the network is provided. A brief account of the partial discharge behaviour and the various monitoring methods for PD activity is described. An elaborate account of the Smart Cable Guard (SCG) system is described in detail.

In Chapter 3, we explore the project data used for the scope of the research. A high level description of the data is provided and preliminary descriptive analysis is performed on the data. This is followed by the description of the Automated Warning System developed at Alliander and the contribution through this thesis towards its development.

In Chapter 4, we discuss the clustering methods and elaborate on the density based clustering methods for the project data. The clustering process implemented is described in detail and the performance indicators to evaluate the clustering method are introduced.

In Chapter 5, we discuss the experiments for our evaluation and elaborate on the results.

In Chapter 6, conclusions and some possible future improvements to this work are discussed.

Chapter 2

Background

This chapter provides a brief description of the medium voltage network structure and its main components. The common causes of failure of the components and the degradation mechanisms associated with the components is briefly summarized. For the scope of this thesis, the basics of partial discharges and its monitoring methods are described along with the advantages of online monitoring. The concept and working of the Smart Cable Guard (SCG) online monitoring tool is elaborately discussed.

2.1 Medium Voltage (MV) Grid Network

In the Netherlands, the network is divided into four different grid voltages [10]:- (i) Low Voltage (rated voltage $\leq 1kV$), (ii) Medium Voltage (rated voltage $> 1kV$ and $\leq 35kV$), (iii) High Voltage (rated voltage $\geq 35kV and \leq 150kV$), (iv) Extra-High Voltage (rated voltage $> 150kV$ and $\leq 380kV$).

The extra-high voltage grid is maintained by the TSO (Tennet) and helps in transporting electricity over large distances in the Netherlands mainly through overhead lines. The high voltage grid connects the extra-high voltage grid to the distribution networks. The voltage at the power station is transformed to high voltage level (HV: $> 35kV$). Power plants, energy intensive industries, and large wind and solar parks are connected to the HV grid. It consists of partly overhead lines and underground cables. In the HV substations, the voltage is transformed down to the MV level (1kV - 35kV). Typically 10 - 30 MV feeders leave the substation to distribute the power to customers in the region [13]. These feeders usually are comprised of underground cables in densely populated regions such as the Netherlands. A feeder is characterized by a number of shorter sections of underground cable that are interconnected by ring-main-units (RMUs) above ground. The length of a cable between two RMUs can vary from 100m to 8-12km depending on the topography [14]. Due to local topological changes, most of the cables between two RMUs consist of multiple shorter cables sections that are connected together by underground joints [15]. Typically, a RMU contains a busbar that connects multiple, usually two, incoming MV cables which can be switched on and off the busbar. The busbar connects to a transformer that transforms the medium voltage to low voltage (LV: $< 1kV$). The low voltage leaving the RMU is distributed over cables or overhead lines to deliver power to customers in the neighborhood.

The structure of most medium voltage grids is ring-shaped or meshed. The ring-main-unit derives its name to the fact that RMUs act as nodes in this ring. The ring structure offers the advantage to network operators more options to feed consumer via an alternative route in the unlikely event of a malfunction or during maintenance and repair hence providing better reliability. Therefore, the medium voltage grid is designed in such a way that it is possible to switch cables to restore the energy supply after a failure has occurred. In that case, the grid operator locates the fault, isolates the fault location and restores the energy supply via another

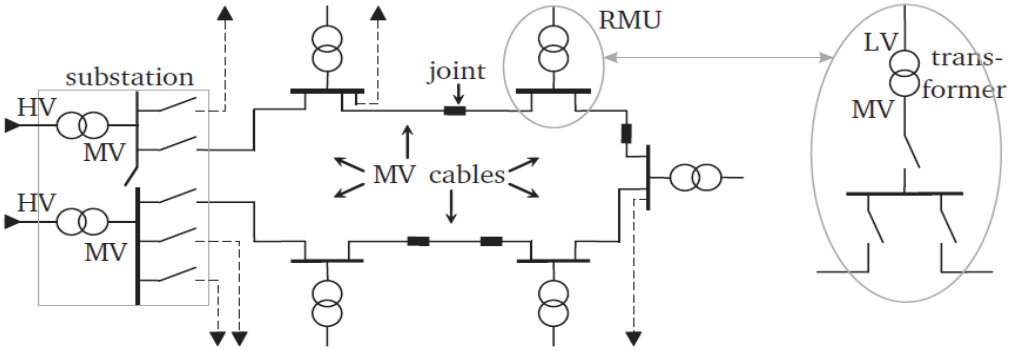

Figure 2.1: Structure of the Medium Voltage (MV) grid in ring configuration [13]

part of the medium voltage grid.

2.1.1 Causes of Failures and Degradation mechanism

A MV grid network comprises of several components, mainly the circuit which consists of the cable, joints which connect parts of the cable, and termination at the end of the cable connection. There are number of different types of cable, joints, and terminations based on topographical conditions. Each type provides different properties based on its construction and insulation material. Failure statistics shown in Figure 1.3 indicate that the MV grid circuit contributes towards 73% of breakdowns in the MV grid, especially 36% from cable joints [10].

The circuit components are connected in a distributed manner meaning that in the case of failure, the issue can be resolved by replacing the section of the failed cable and as well as the failed joint and thereby reducing the time of interruption in power supply. For instance, if part of a cable connection shows strong degradation, it is removed and replaced. Such repairs may not only result in an exchange with a similar component, but it also could involve replacement by another type of the same component or even a modification of the topology of the circuit. Each of these components may have different insulating media. Cables can be either paper insulated lead covered (PILC) type or cross-linked polyethylene (XLPE) type [16]. Joints and terminations can have insulation types such as mastic, resin, paper, oil-filled, premoulded and hot/cold shrink [16]. A schematic representation of the circuit components is shown in Figure 2.2.

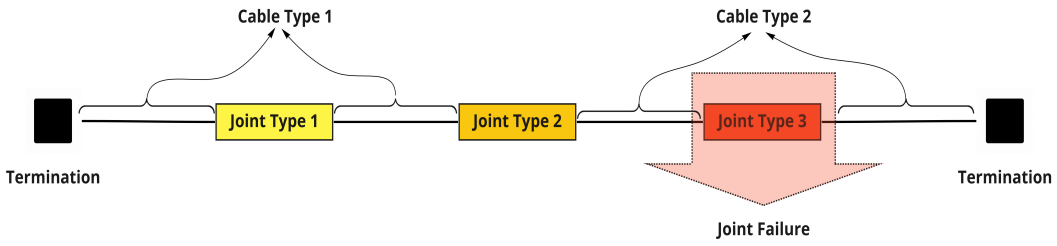

Figure 2.2: Representation of the circuit with its components (cables, joints, terminations)

The causes of defects leading to a breakdown in insulation that occur in the circuit com-

ponents can be seen from the Figure 1.3. The main contributing factors can be characterized into three main classes [16, 15]:

Manufacturing faults
Impurities induced during the manufacturing process lead to defective components. Cavities and other defects in insulation emerge due to poor treatment of the insulation material during the design and engineering phase. Incorporating factory acceptance tests (FAT) and production tests for examining the quality can help in identifying cavities or inclusion of impurities or bad finishing of the components.

Incorrect Handling
Substandard handling of the circuit components during transportation can result in damage to the protective layers of the cable and can lead to initiation of sites of degradation in the insulation. Unprofessional workmanship or lack of sufficient assembling guidelines by the manufacturer can cause several problems such as misplacement of conductors in a joint, rough treatment of components during installation etc. Such violations are prone to occur during high-pressure situations or bad weather conditions increasing the probability for inefficient handling.

In-Service defects
Defects and stress induced during the operation of the components can be classified into the operational conditions of the system, environmental condition surrounding the components and external forces acting on the circuit.

- Operational condition - Aging of the component leads to deterioration of the chemical or mechanical properties of the component insulation thereby reducing its dielectric or structural strength consequently leading to a breakdown. Overloading of the cables can cause thermal and mechanical stresses and lead to damage in the insulation.

- Environmental condition - The surrounding conditions upon laying the circuit components in the ground affect the lifetime of the insulation. Interaction between the moisture content of the soil and the component insulation strongly impact the aging process. Soil operation can lead to thermal and mechanical stresses on the insulation such as a drop in soil water level combined with increase in soil temperature due to overloading can result in deterioration of the insulation. Loosened soil can lead to sinking of the cable causing mechanical stress affecting the water tightness of the joint creating a path for moisture penetration into the insulation resulting in corrosion of the cable.

- External forces - Excavation activities due to repair work or nearby construction activities directly damage the cable insulation. These activities can also cause vibrations in the soil which may cause misplacement of conductors in the joint thereby reducing the water tightness of the joint.

A single factor or a combination of above mentioned factors can result in the acceleration of the degradation mechanism of the component. The main degradation mechanisms associated with medium voltage levels are summarized [17]:

- Thermal degradation: Low dielectric loss in new equipment help prevent overheating of the component hence avoiding insulation breakdown. However, the dielectric losses may increase gradually over time due to aging of the component. The aging process is accelerated during heavy service conditions resulting in high temperatures thereby reducing dielectric strength.

- Partial discharges (PD): Emergence of discharges is evident in the cavities or cracks in insulation when the electric field of the insulation defect exceeds its inception voltage [18]. This process is often accelerated due to high temperatures, chemical deterioration etc. making PDs both a cause and a symptom of deterioration.

- Electrical treeing: The process of electrical treeing is considered as the last step in the insulation degradation before a breakdown occurs. Corrosion of the insulation due to partial discharges leading to an emergence of carbonized channels indicates the inception of electrical treeing.

- Water treeing: The process of water treeing is caused due to a combination of water and electrical stress leading to a degradation in polyethylene materials [19]. The electric breakdown strength of a polyethylene insulating material is reduced due to water trees. Although it takes many years for water trees to develop, it can lead to formation of PDs and electrical trees in its final stages resulting in further deterioration and eventual breakdown.

2.1.2 Partial Discharge Behaviour

The IEC 60270 standard [20], defines partial discharge as, *"localized electrical discharge that only partially bridges the insulation between conductors and which can or can not occur adjacent to a conductor"*. These discharges are a result of local electrical stress concentrations in the insulation that appear as pulses with a duration of much less than $1\mu s$.

Partial discharges can be categorized into three types namely internal, surface, and corona discharges [16]. *Internal discharges* occur in gas-filled voids or cavities within solid dielectrics which may lead to electrical treeing if the frequency of discharges increases leading to full breakdown (Figure 2.3 (b)). *Surface discharges* occur along the dielectric interface. The conductive path formed along the surface of the insulation leads to tracking which eventually propagates into electrical treeing causing a complete breakdown Figure 2.3 (a)). *Corona discharges* are caused due to electrical overstress at sharp edges in gas or liquid medium due to the ionization of air surrounding the conductor leading to inhomogeneous field.

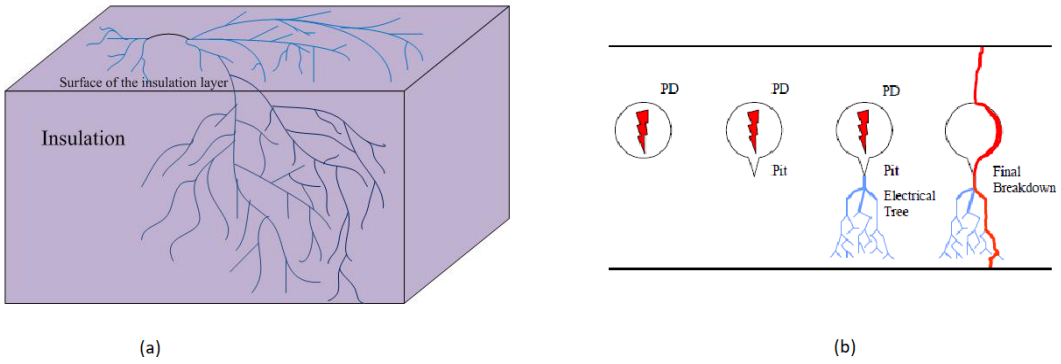

Figure 2.3: (a) Formation of tracking from surface discharge leading to inward electrical treeing [16], (b) Growth of damage in cavity from PD to formation of pit and to breakdown [21].

2.1.3 Partial Discharge Monitoring

Several diagnostic methods are present to identify the deterioration of the insulation in the circuit components and help in the determination of the circuit condition and ultimately its lifetime. Testing procedures at each stage of the cable lifecycle - development, type testing, pre-qualification, production, installation, service, and eventual failure - are available [22].

PD detection methods provide useful information about the cable insulation and its accessories [23]. High frequency pulses propagate towards the cable ends during partial discharge activity. The PD detection technique helps provide a reliable way to detect these high frequency pulses and also helps in locating defective sites or regions along the cable length.

Partial discharge monitoring can be carried out either offline or online. Offline monitoring of partial discharges in medium voltage cables involves disconnecting the cable from the grid. This allows adjustment of test parameters to achieve an optimal outcome of the diagnostic technique. The cable under test requires a separate power generator to energize the cable and testing needs to be carried out in a controlled environment i.e. with low disturbance levels and on a voltage level of choice. High frequency pulses originate upon excitation of partial discharges at sites along the cable which propagate in both directions and are reflected at cable ends. The location of the PD site is identified by calculating the difference in the time of arrival of the first pulse and the reflected pulse at the far end of the cable [24].

Although offline monitoring techniques provide flexibility in test settings for better detection sensitivity to PD defects, measurement of PDs while the cable is in-service i.e. online over a longer time periods prove to be beneficial in understanding the development of defects in the cable circuit. The advantages of online monitoring techniques over offline monitoring techniques are enumerated below:

- Online monitoring does not require discontinuation of power supply unlike offline techniques which requires disconnecting the cable from the grid to isolate the circuit leading to increased risk in loss of supply and increased time of testing.

- Offline methods pose a limitation in the number of cable sections that can be tested at one time to avoid the risk of serious outage. This is not the case with online monitoring methods.

- Online monitoring reduce the need for personnel efforts for testing thereby reducing the cost of operation for utilities.

- The cable is monitored under real operating conditions which include the effect of temperature and humidity, overvoltages, and variation in loads which can better insight into the cable condition during actual conditions.

- Online monitoring provides continuous data registration that gives a possibility of observing the trend of PD activity over time and its effect on ageing and insulation degradation. This also helps in capturing PDs occuring before a failure. In case, a PD activity occurs only for a short period of time and disappears this trend can be intercepted by online monitoring methods [25].

2.2 Smart Cable Guard (SCG) System

During the period 2001-2005 [17, 13, 26], extensive research was conducted to investigate and develop a cost-effective solution for online PD detection and localization for MV cable systems. This led to a presentation of a proof-of-concept called as PD-OL (Partial Discharge testing On-line with Location) [27]. The proof-of-concept was made commercially available under the product name Smart Cable Guard (SCG) by DNV GL [28].

The Smart Cable Guard (SCG) system consists of two measurement units which are installed at the cable circuit ends terminated at the substation or RMU(s). The measurement unit comprises of a Sensor/Injector Unit (SIU) and a Controller Unit (CU). The measured data is communicated to the server via the internet [29].

The SIU consists of a sensor that measures pulses from the cable and an injection device to inject pulses into the cable for time synchronization of the sensors. The sensor/injector unit is of inductive type and hence placed either around the cable or the earth lead of the cable. This makes it possible to mount the device without disconnecting the cable. The controller unit operates as the heart of the locally installed SCG system and is connected to the sensors

through an optical fibre cable. The CU is also connected to DNV GL's control centre server through internet via GPRS or LAN connection. The DNV GL control centre server collects the measurement data and provides a web platform to visualize the PD and fault data 24/7 [30]. In the event of fault the server sends out a warning to the network operator via email and SMS including the location of the fault and the timestamp of the occurrence of the fault.

2.2.1 SCG Monitoring Methodology

The Smart Cable Guard monitoring methodology for the detection and location of PDs can be defined by following process steps:

- The sensor/injector units clamped at the ends of the cable measure pulses propagated along the cable length. The sensor installed at one end of the cable acts as a master and the sensor at the other end will act as a slave. Every minute, the master unit injects a pulse to the other end of the circuit where it is received and detected by the slave unit. This patented pulse injection technique helps in time synchronization of the internal clocks of the inductive sensors with an accuracy of 100ns and therefore perform accurate fault/defect localization.

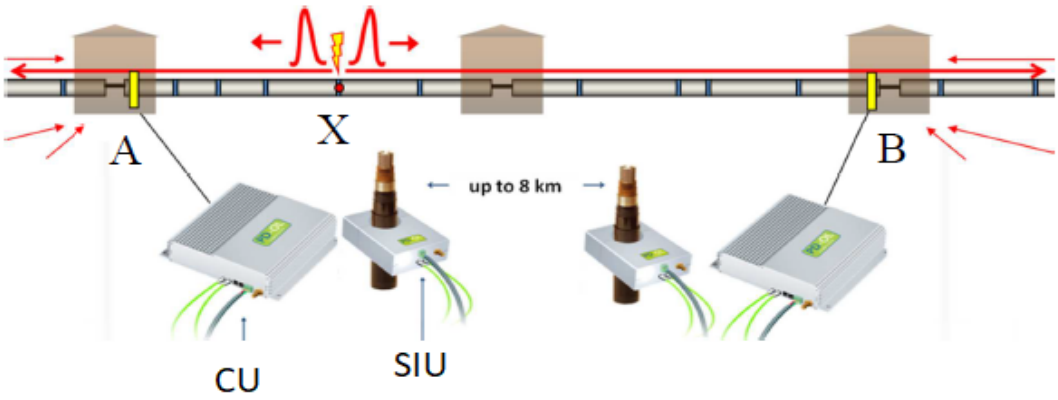

Figure 2.4: SCG Measurement Setup [14]

- From a defect spot X (Figure 2.4) between positions A and B, electromagnetic waves from PD pulses travel along the cable in two directions, away from the defect spot X. Each of the two SCG sensors detects the PD pulse passing. The travelling wave amplitude alongwith the time of arrival of the pulse is stored. The difference in the time of arrival of pulses detected at the cable ends helps the SCG system to accurately locate the origin of the PD using the equation [31]

$$l_{pd} = \frac{L}{2}\left(\frac{\triangle t}{\triangle T} + 1\right) \tag{2.1}$$

where:

l_{pd} : defect location where the PD pulse originated from
L : total cable length
$\triangle t$: difference in arrival time at both cable ends of the PD pulses coming from the same origin.
$\triangle T$: cable propagation time [order of nanoseconds]

- The Controller Unit (CU) connected to the SIU controls the measurement sequence, data collection, signal processing and the data communication to the Control Centre Server. PD signals are impeded by noise and interference from the surrounding signals such as radio broadcasts, broadband background noise, and finite-energy interference such as thyristor pulses, switching transients or PDs from adjacent HV systems. This makes signal processing a crucial part for noise suppression. Techniques like matched filtering help in detecting PD pulses in the presence of noise. Matched filtering technique is used for detection of deterministic pulses in the presence of noise. Hence, a matched filter is optimized such that the average signal to noise ratio (SNR) is maximized at the filter output provided, the signal waveform and noise properties are known or can be estimated. PD signal waveforms can determined by the signal propagation paths of the cable system and therefore matched filters can be designed for PD signal extraction [26].
- The injected pulses from the SIU can also be used for calibration of the measuring system. The pulses injected have a known pulse shape which helps in determining the transfer impedances at the ends of the cable. This information can help in calculating the PD charge from the measured PD pulse shape and amplitude using the matched filters. Matched filters are scaled to the measured PD pulse. This provides a linear relation between the PD pulse measured and the filter response and hence the PD charge can be determined directly from the maximum filter response [26, 31].

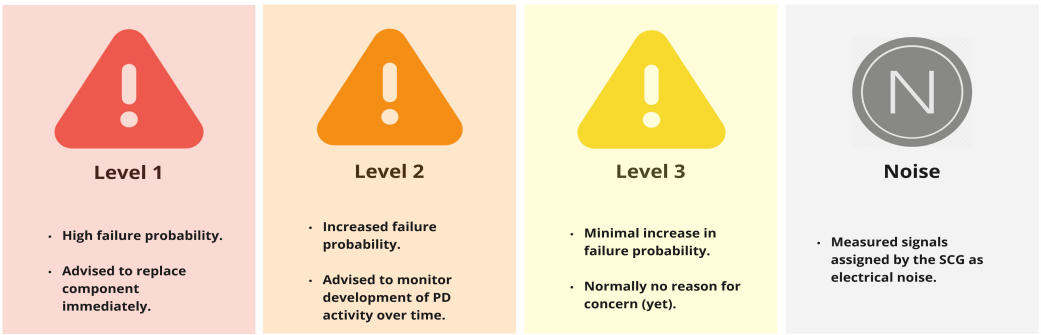

Figure 2.5: SCG Warning Levels

- Since multiple SCG systems are installed in the field locations, the PD measurements detected by the SIU are processed locally by the CU and communicated to the Control Centre Server via the internet. The data from both the SIUs is collected and combined to eliminate external pulses and to calculate the PD origin. Next, various statistical parameters are calculated to understand the PD behaviour. The trend in the PD data is evaluated using knowledge rules [1] and expert opinion and warnings are assigned to indicate the severity of the PD data (Figure 2.5). The PD data is made available every hour on the web platform along with the warnings assigned. The network operator can also view the trend of the PD activity in a 3D plot on the web platform provided by DNV GL (Figure 2.6).

The warnings for the SCG monitored circuit are assigned by a DNV GL expert who continuously observes the PD data sent to the control server every hour. The warning is a qualitative assessment from the DNV GL expert based on the development of the activity from the last

[1] These knowledge rules are proprietary to DNV GL and are not present in the public domain. The assessment methods employed were known through discussion with the DNV GL experts during the course of this thesis.

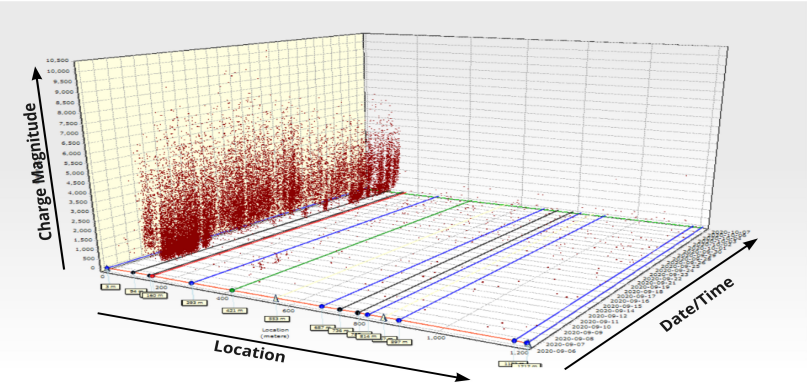

Figure 2.6: SCG Web Platform (Source: DNV GL)

time. This makes it a tedious task for the warning assignment as well as placing tremendous dependence on the availability of the DNV GL expert from the operator side. This makes it imperative to find a way to automate the process of identifying the development of this activity.

2.3 Summary

This chapter provides background information necessary to understand the system of interest. Section 2.1, provides an overview of Medium Voltage (MV) grid network with its structure and configuration characteristics.

In Section 2.1.1, the common causes of failures observed in the MV network components. The degradation mechanisms that may affect the component lifetime and can lead to their failure is summarized. One of the most prominent symptom and causes of degradation is the partial discharges occurring in the insulation.

The behaviour of partial discharges and its various types is discussed in Section 2.1.2.

The methods of monitoring are discussed in Section 2.1.3 along with the advantages of online monitoring of partial discharges.

Section 2.2 introduces the Smart Cable Guard (SCG) system used for online monitoring of partial discharge activity. The methodology employed by SCG for monitoring of PD activity is elaborately discussed in Section 2.2.1.

Chapter 3

Project Data

This chapter introduces the data provided by the SCG system to the network operator. The circuit characteristics are also briefly described which provide a high-level information of the circuit. A preliminary descriptive analysis is performed on the available data to understand its characteristics. An introduction to the Automated Warning System developed at Alliander is provided with an overview of the development steps of the architecture. The development steps relevant to the scope of this thesis are addressed.

3.1 Features

There are approximately 1500 active SCG systems mounted on the Alliander circuit network. Each circuit data consists of the characteristic properties of the circuit along with the PD data from the SCG systems provided by DNV GL. In Table 3.1, an overview of the Alliander Circuit data is provided along with a brief description and type of data.

Table 3.1: Alliander Circuit data

	Datatype	**Description**
Circuit ID	integer	Unique identification for circuit with SCG system.
Cable Congifuration	data table	Information of the locations of RMU(s), Terminations, and Joints(type).
Circuit Length	float	Length of the circuit in metres (m).
Start/End Location	string	Geographical Start and End Location of the circuit.

The circuit data from Alliander describes the high-level characteristics of the circuit with a SCG system installed. The Circuit ID serves as an unique identifier for the circuit. The Cable Configuration provides information regarding the locations of RMU(s), terminations, and joints in the circuit. It also holds information of the type of joint installed. The SCG system is installed at the circuit ends and hence the length of circuit is of importance.

In Table 3.2, the SCG data received from DNV GL for each circuit is described. The PD data provides the information of discharges in the circuit. The data table consists of the location of each observed discharge (in metres), the timestamp of the occurrence of discharge and the magnitude of the discharge occurred (in picocouloumbs or pC). Every hour, one end of the circuit SCG sensor/injector unit injects a pulse which is detected by at the other end of the circuit to measure the propagation time of the pulse along the circuit. The SCG sensors also

inject a high frequency pulse every hour to detect the pulse amplitude and calculate the charge associated with it. This is done to measure the PD detection sensitivity of the SCG sensors. The Warnings data table gives the warning levels (Figure 2.5) assigned to discharge locations of the circuit and time of occurrence of the discharge events and is updated every hour on the SCG web platform.

Table 3.2: SCG Data from DNV GL

	Datatype	**Description**
PD data	data table	Date/Time, Location, Charge Magnitude of the discharge activity
Propagation time	time(ns)	Propagation time of the circuit measured by the SCG systems every hour.
Sensitivity	charge(pC)	PD charge detection sensitivity measured every hour.
Warnings	data table	Date/Time, Location and Warning Level (1, 2, 3 or Noise)

3.2 Exploratory Data Analysis

This section explores the high level characteristics of the circuit data. From the actively monitored circuits, the boxplot of the length of the circuits (Figure 3.1) shows the median circuit length is around 3000m. Based on the circuit configuration, the average percentage of cable length (in metres) that consist of the PILC (paper insulated lead covered) insulation type in the circuit are plotted. It is evident from the boxplot that there is higher percentage of PILC insulated cables in the circuits.

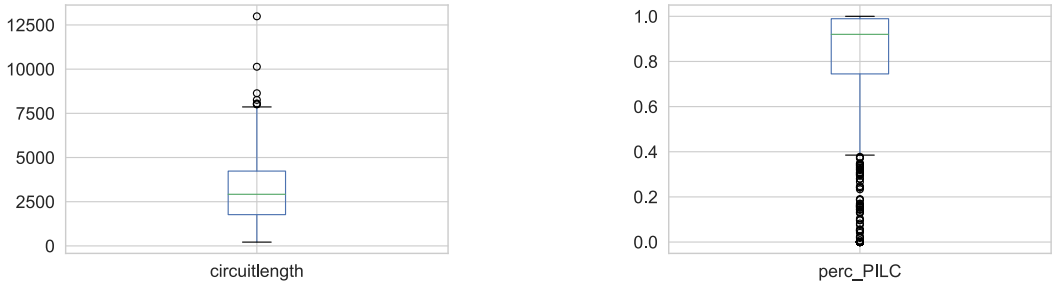

Figure 3.1: Average circuit length of the monitored circuits and the average percentage of XLPE and PILC type cable in the circuits

In the Figure 3.2, the scatter plots show the number of warnings observed for the number of joints in the circuits, the number of RMUs in the circuit, and for the length of the circuit. The warnings considered here are the DNV GL Level 1, 2, 3, and Noise warnings. There is very little correlation with the number of warnings with respect to the number of joints and number of RMUs in the circuit. The number of warnings observed for the increasing circuit length shows a negative correlation although it is very low.

The scatter plots for the number of warnings with respect to circuit length is plotted for warnings levels 1 and 2 collectively and also for warning level 1 alone (Figure 3.3). On further observing the correlation between the number of level 1, level 2 warnings and the length of

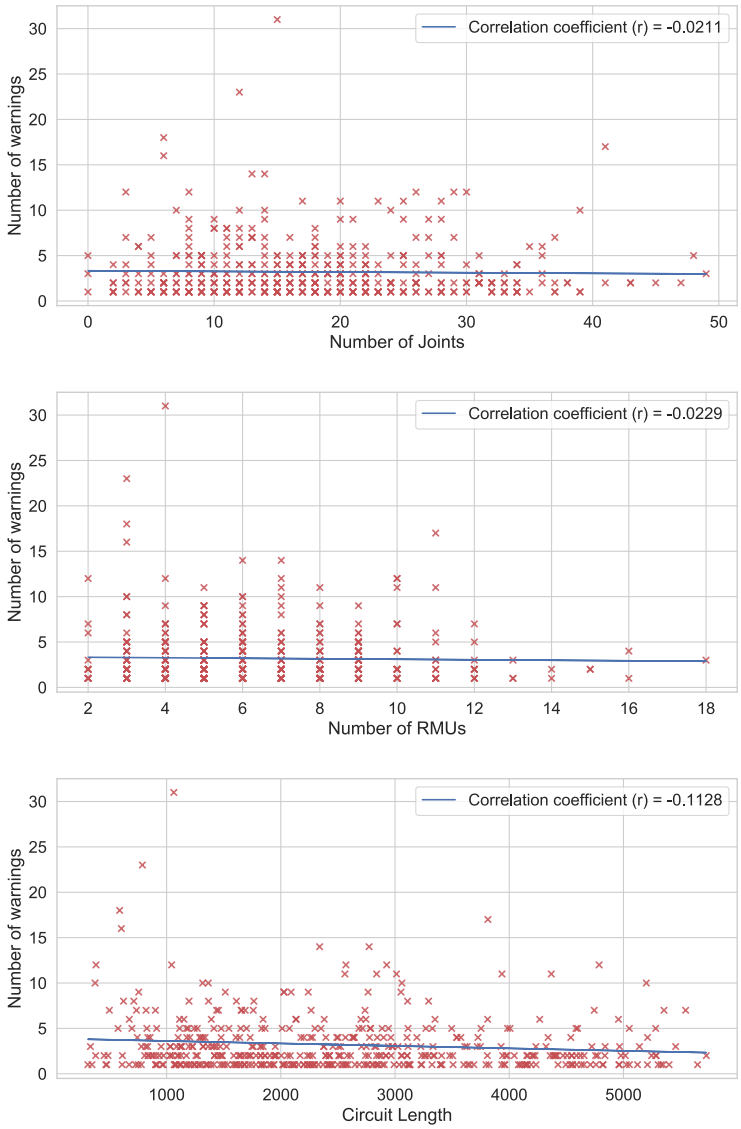

Figure 3.2: Number of warnings observed with respect to number of joints and RMUs in the circuit and length of the circuit

the circuit, the correlation becomes weak to draw upon any conclusions on the effect of circuit length on the number of warnings observed. It is interesting to note here that the proportion of level 1 warnings is far less than the total number of warnings observed (level 1, 2, 3, N) (Figure 3.2). It is also evident that the level 1 warnings are observed for circuit lengths less than 4000m which is also the average circuit length of the actively monitored circuits.

The SCG data from the actively monitored circuit provides the location of the PD defect, the timestamp of occurrence of the defect and also the magnitude of charge measured for the PD. The data for one circuit is visualized in three different views namely the front, top and side view (Figure 3.4, 3.5, 3.6). These views are a combination of the three attributes of the SCG data and the observation of each attribute with respect to each other. The front view gives

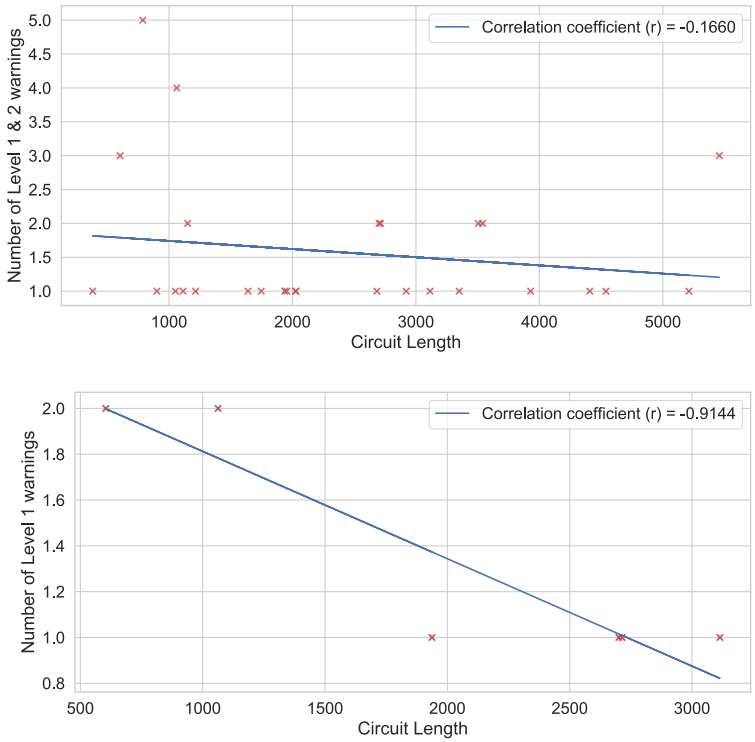

Figure 3.3: Number of warnings (Level 1 & 2, Level 1) observed for the length of the circuit

the charge magnitudes observed at each locations of the circuit. The top view helps observe the concentrations of discharges occurring at locations and see the trend in activity spread in time. The side view provides a way to visualize the variation in charge observed in time. For the observed circuit, the DNV GL experts assigned 3 warnings on location 349m (level 3) from February 2018 - May 2018 and on location 604m (level 2 and level 3) from January 2018 - February 2018.

From the SCG data, the number of discharges occurring per day for the entire circuit is visualized in the Figure 3.7. From the months February to March and from March to May, peaks can be observed in the number of discharges with the number reaching upto 35,000 discharges cumulatively in February for the circuit. This is also observed from the Top View (Figure 3.5) with dense activity for locations around 500m and around locations 300m-400m.

The average charge (in nC) per day is plotted to observe the magnitude of discharges measured by the circuit per day. The highest spike in October 2018 is attributed to the high discharge magnitude observed at the location ∼2200m. This location is the end of the circuit where the SCG sensor/injector units are clamped and also can be a place where there might be RMUs present (which is the case here). The RMUs have multiple MV feeders (cables) passing through and disturbances in the RMU terminations can be mistakenly picked up by the SCG sensors as high magnitude discharges occurring at the location [26]. The count of discharges observed at each locations in the circuit is plotted in Figure 3.9. It can be seen from the figure that the count of discharges is significantly high for locations around 300 to 600m which also holds true with the warnings assigned by DNV GL experts for the previously mentioned locations.

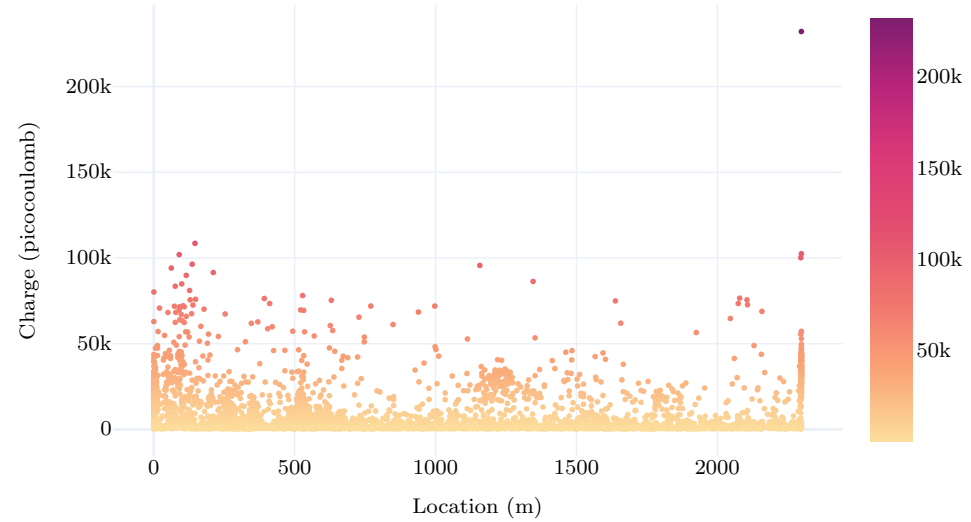

Figure 3.4: Front Visualization of the SCG PD data. *(Color bar on the right indicates the charge magnitude in picocoulomb)*

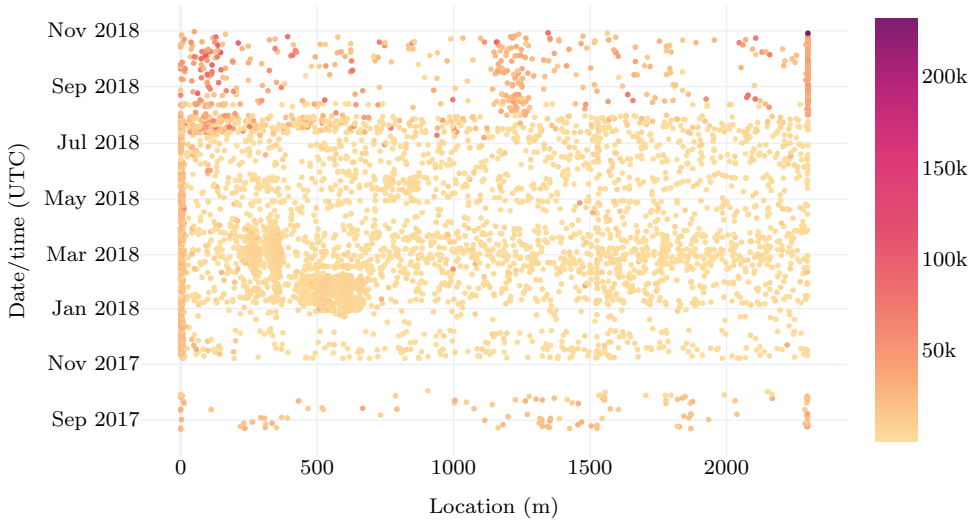

Figure 3.5: Top Visualization of the SCG PD data. *(Color bar on the right indicates the charge magnitude in picocoulomb)*

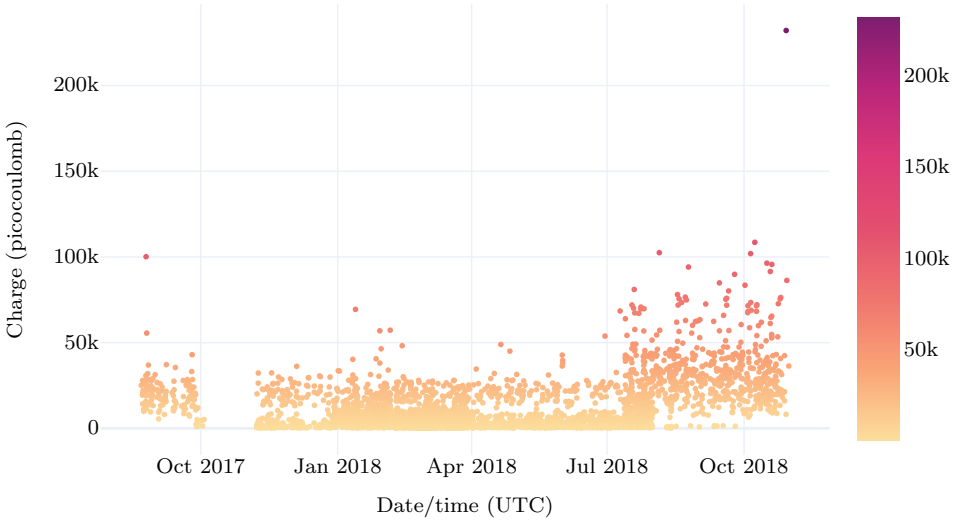

Figure 3.6: Side Visualization of the SCG PD data. *(Color bar on the right indicates the charge magnitude in picocoulomb)*

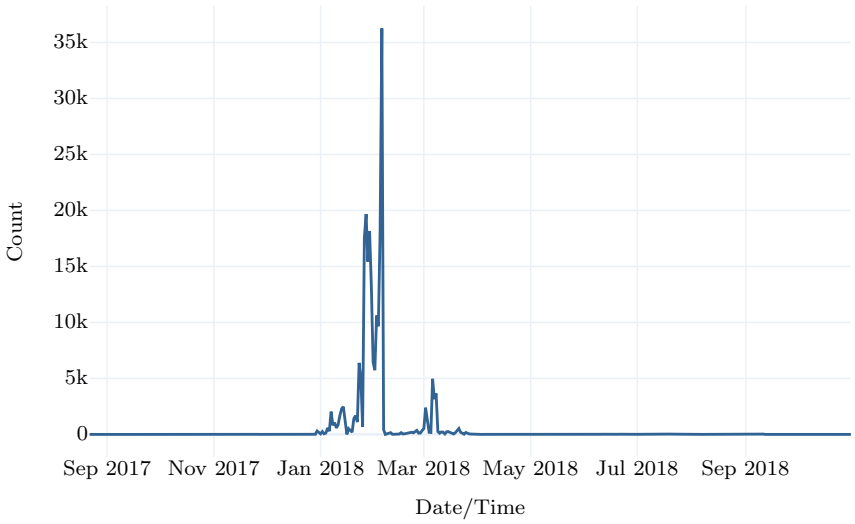

Figure 3.7: Number of discharges per day

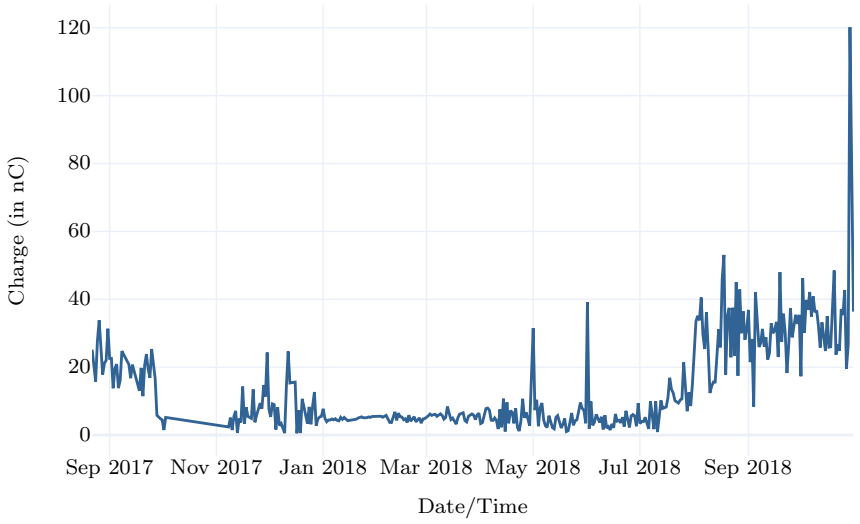

Figure 3.8: Average Charge per discharge event (in nC) per day

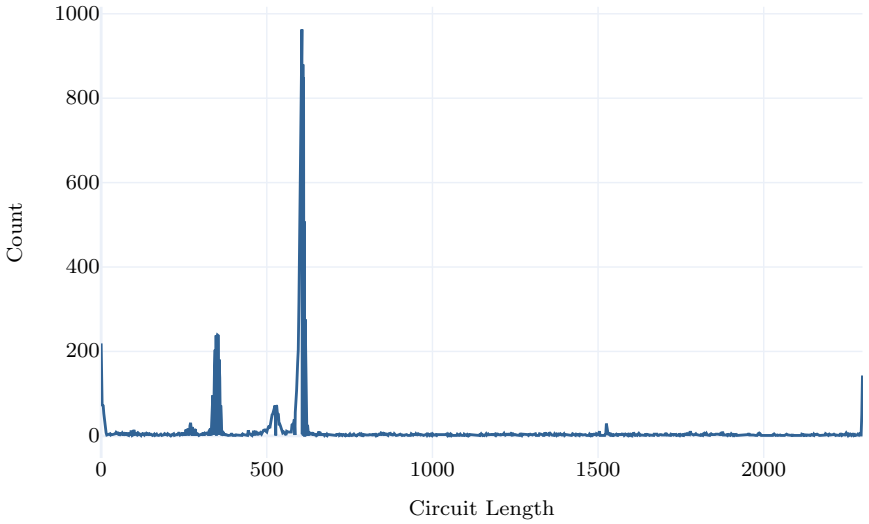

Figure 3.9: Number of discharges per location in the circuit.

3.3 Automated Warning System at Alliander

The need for preventive maintenance in the medium voltage cable network has accelerated the development towards automating the discovery of high density discharge events in the circuit and to develop knowledge rules for evaluating these events. The SCG Analytics team at Alliander plans to develop an Automated Warning System (Figure 3.10) which will serve as a decision support tool for the experts of the network operator.

The purpose of the decision support tool would be to collect the SCG data provided by DNV GL for the monitored circuits and employ noise reduction techniques and identify high density discharge events in the circuits. The identified clusters of high-density discharges would then be evaluated by calculating statistical parameters alongwith the underlying circuit configuration information. These parameters would then serve as features to the clusters and would help in deciding whether the identified cluster resembles PD activity (Level 1, 2, 3) or noise.

3.3.1 Architecture

The architecture [32] for the automated warning system is divided into four main development stages:

Data Preparation
In the data preparation stage, raw data collected from the active SCG circuits is collected and cleaned by removing inconsistent or missing data such as measurement that have missing location/timestamp/charge magnitude values. The data then undergoes a preprocessing step, wherein the PD data is checked for gaps in the measurement. If the gaps are encountered to be longer than 12 hours then the PD data missing for the gap duration is reported. The processed PD data is then sent to the Clustering stage.

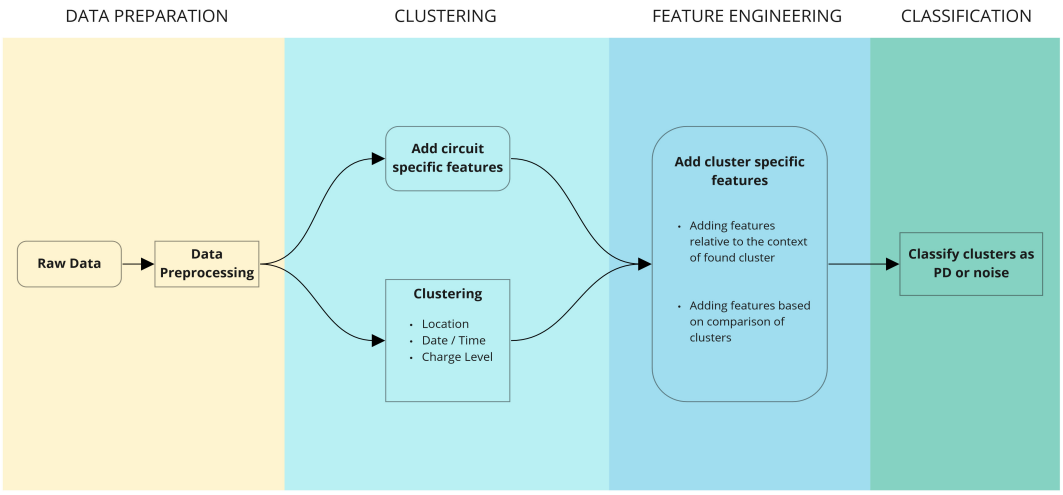

Figure 3.10: Architecture of the Automated Warning System at Alliander [32]

Clustering
In the clustering stage, the preprocessed data undergoes two processes. First, the circuit specific features such as the configuration of the cable circuit which includes the locations of the cables, joints, terminations, and RMUs present in the circuit as well as the their insulation type are added. This also gives the information of the total length of the circuit. Next, the SCG data containing the location, timestamp and charge magnitude of the discharges for the circuit

are passed into the clustering process. Here, clusters of high density discharges are identified among the measurements. The clustering stage serves as a noise filtering step for the PD measurement data. The clusters identified are checked for DNV GL warnings present in them and are evaluated using performance indicators.

Feature Engineering
The clusters identified in the clustering stage are evaluated by extracting features in the feature engineering stage. Features of the clusters provide information of the properties of the cluster. The features specific to the clusters are calculated by measuring the width of the cluster (in location), calculating the median location as it gives the information of location where maximum discharges have occurred, the charge distribution of discharges in the clusters, density of the cluster which gives the average discharges occurring in the cluster per day etc. This helps in distinguishing clusters from the list of clusters identified for the circuit. The circuit configuration information added in the clustering stage is used to calculate features of the clusters in relation to its configuration such as calculating the nearest joint and RMU location by taking the difference of the median location from the locations of the joints and RMUs in the circuit. This information provides insight into whether a cluster identified by the algorithm was identified in close proximity to a joint or RMU.

Classification
Once the features are calculated for the identified clusters and assessed for warnings present in them, the final step of the system is to distinguish PD or 'relevant' clusters from noise. To classify clusters with PD activity from noisy clusters, the features calculated in the feature engineering step will be evaluated by calculating correlation among them. Based on the feature correlation, the features will be ranked according to the order of importance. The identified cluster set is divided into training, testing and validation sets and the classifier will be trained on the training and validations set and the performance of the classifier will be evaluated on the test set.

3.3.2 Thesis Contribution

The scope of this thesis focuses on the *Clustering* stage and the *Feature Engineering* stage of the Automated Warning System. A clustering method is implemented to detect high-density discharge events in the SCG PD data. The clusters identified by the clustering method are validated against DNV GL warnings using performance indicators. Features are developed to extract information from the identified clusters and to give a possible way to describe them quantitatively which would further supplement the classification model in distinguishing the clusters with PD activity from noise clusters.

3.4 Summary

In Section 3.1, the features of the SCG and Alliander Circuit data were briefly described. The circuit data is specific to the monitored circuit and provides the configuration information of the circuit. The SCG data for the monitored circuits comprises mainly of the PD data and the warnings data provided by DNV GL server for the circuit. The propagation time and sensitivity feature of the SCG serves as an additional information of the circuit characteristic as well as calibration information of the SCG sensors.

Section 3.2 performs descriptive analysis of the overall information of the monitored circuits and the SCG data received for a circuit.

In Section 3.3, a detailed description of the Automated Warning System developed by Alliander is provided. Various development stages are explained and the contribution of this thesis is identified in two of the development stages.

The following chapter provides the description of the clustering technique employed on the

SCG data and the performance indicators to evaluate the clustering method are described.

Chapter 4

Data Clustering

This chapter provides an overview of clustering methods and an in depth description of the density based clustering methods. The methodology implemented for the data is discussed in detail. Performance indicators are formulated to evaluate the clustering method.

4.1 Clustering

Data mining involves the discovery of interesting patterns in large data sets which in turn can provide a way to understand the data and thus help in making decisions or predictions based on the data. One such task in data mining is clustering. Clustering task involves partitioning of a large number of data points into a smaller number of groups (or clusters) such that the data points in the same cluster are more similar to each other than to those in other clusters [33]. Greater similarity or homogeneity between the data points of the group and greater the difference between group, the clustering would yield better and more distinct clusters.

There are different charcteristics of clusterings used to form clusters which are summarized below [34]:-

Partitional versus Hierarchical
Partitional clustering divides a set of data points into non-overlapping subsets (clusters) such that each data point is in exactly one subset. In hierarchical clustering, clusters are allowed to have subclusters such that these group of nested clusters are organized as a tree. The nodes (clusters) of the tree are a union or collection of its children nodes (subclusters) and the root of the tree is the cluster containing all data points.

Exclusive versus Overlapping versus Fuzzy
When each data point is assigned a single cluster then the clustering is known to be exclusive. In cases where some data point(s) may belong to more than one cluster then such clustering is known as non-exclusive or overlapping. In fuzzy clustering, every data point is associated with every cluster with a membership weight ranging between 0 (meaning no membership) or 1 (meaning absolute membership).

Complete versus Partial
Complete clustering assigns every data point to a cluster. This not the case in partial clustering where some of the data points may not be assigned to any cluster. Such data points may represent noise or outliers and hence may not belong to any well-defined clusters.

Several models of clustering are present based on the above types of clustering [33, 34].

- *Connectivity-based clustering* model connects data points to form clusters based on their

distance and its result is often represented using a dendrogram and hence these models follow a hierarchical clustering. A dendrogram is a tree that iteratively splits the data into smaller subsets until each subset contains only one data point. The dendrogram can either be constructed from the leaves to the root by merging clusters at each step, also known as the agglomerative approach, or from the root down to the leaves by dividing clusters at each step, also known as the divisive approach (eg. SLINK [35], CLINK [36]). Although the model provides an ease of interpretation, its results are difficult to use as they provide a hierarchy that needs to processed further to find appropriate clusters. Also, the model does not have a notion of noise and hence is not robust to outliers.

- *Centroid-based clustering* model partitions the data into a set of k-clusters for a given parameter k. The algorithm is initiated with an initial partition of the data and then incorporates an iterative strategy to optimize an objective function. The clusters can be represented by the centre of gravity of the cluster (k-means algorithms) or by one of the data points of the cluster close to its centre (k-medoid algorithms). These partitioning algorithms are executed in a two step procedure - first, the k representatives are determined by minimizing the objective function and then each data point is assigned to a cluster with its representative closest to the considered data point. One of the drawbacks of this model is that it requires the number of clusters k to identified to be known a priori. Also, since the initialization criteria is often done randomly by sampling from the database, the presence of outliers might have a detrimental effect on clustering process and lead to the formation of singleton or empty clusters.

- *Distribution-based clustering* model assumes the data to be generated as result of statistical process and clusters can be identified using known probability distributions. The clustering process involves deciding on a statistical model for the data and estimating the parameters of that model from the data. One of the major disadvantages of this model is that for many real world data sets there might not be well defined mathematical models to describe the data and hence might not be able to produce clusters effectively.

- *Density-based clustering* model identifies areas of high density separated from areas of low density. One of the advantages of the model is that they do not require the number of clusters as input parameters and also do not make any assumptions of the underlying density or the variance within the clusters in the data. Unlike most algorithms which partition the data into predefined groups where the sum of squared pairwise dissimilarities between cluster data points or the sum of squared dissimilarities of all cluster data points with respect to some cluster representative (e.g., mean, medoid) using some dissimilarity measure, density-based clusters are not necessarily groups of points with a low pairwise within-cluster dissimilarity and thus need not have convex shape but can be arbitrarily shaped clusters. This property also provides a notion of noise to the clustering method and makes it robust against outliers.

In the context of the project data, the data received from the SCG system for the circuit provides the location, timestamp, and magnitude of the discharge events and the warnings assigned to the circuit. This information does not provide the number of clusters that may be present for the circuit and for the duration of observation. This makes it difficult to employ any centroid-based clustering model as it requires a prior information of the clusters (k) that are needed to be found in the data. Based on the representation of the data in Figure 3.5, it is seen that occurrence of discharge events has varying density and not every discharge event is of interest as discharge activity which is sustained over time has more severe effects on the operation and lifetime of the component. In hierarchical or connectivity-based clustering, each point is assigned to a cluster or an association link is provided which makes it less effective in separating noise and dense discharge activity as there would be a large number of clusters

found and also the partitioning criteria would be required as an input from the user. This is also true with centroid-based clustering where each point is assigned a cluster based on the point's distance closest to a cluster. For the distribution-based clustering model, which assumes the distribution of the data, the PD data from the SCG system do not follow any know distributions making it difficult to model the clustering with respect to a known distribution or a mixture of distributions.

With the density-based clustering model, the characteristics specified previously provide an advantage in the aspect of non-assumption of the distribution of the data as well do not require the number of clusters to specified apriori. This makes it suitable for using a density-based approach towards identifying dense discharge activity in the SCG data from the surrounding 'noise' or uninteresting activity. In the following sections, two density-based clustering techniques are explained in detail namely, DBSCAN or Density Based Spatial Clustering of Applications with Noise and ST-DBSCAN which is a variant of the DBSCAN method. For this thesis, the ST-DBSCAN method is employed on the PD data recieved from the SCG system to evaluate the performance of the clustering method in effectively identifying clusters of interest.

4.2 DBSCAN

Density Based Spatial Clustering of Applications with Noise (DBSCAN) [37] is one of the most widely used and cited density-based clustering model. The paper [37] presents a formal model for identifying density-based clusters and also a database-oriented algorithm to find clusters that adhere to the density model.

The DBSCAN model requires two parameters, a spatial threshold or ϵ-radius (with an arbitrary distance measure) which defines the minimum distance between two points and the *minPts* which is based on the threshold for the minimum number of neighbors of objects. If the distance between two points is lower or equal to the ϵ threshold then the points are considered as neighbors. Objects that have more than the *minPts* neighbors within the ϵ-radius are considered as *core points*. This forms the main intuition of the DBSCAN model, to find areas that satisfy the minimum density criteria and thus can be separated from areas of low density. Figures 4.1, 4.2, 4.3 illustrate the various concepts of the DBSCAN clustering methodology (Here for the sake of illustration the *minPts* is considered to be 4).

Below are the formal definitions adapted from the paper [37] that expand on the DBSCAN clustering model:

Definitions

1. *The ϵ − neighborhood of a point p, denoted by $N_\epsilon(p)$, is defined by $N_\epsilon(p) = \{q \in D | dist(p,q) \leq \epsilon\}$, where D is a set of points and dist(p,q) is a distance function e.g. Euclidean distance, between p and q.*

2. *A point p is a core point if $|N_\epsilon(p)| \geq minPts$.*

3. *A point p is directly density-reachable from a point q with respect to ϵ and minPts if $p \in N_\epsilon(q)$ and q is a core point.*

4. *A point p is a border point if p is directly density-reachable from a core point q and $|N_\epsilon(p)| < minPts$.*(Figure 4.3)

5. *A point p is density-reachable from a point q with respect to ϵ and minPts if there is a chain of points p_1,...,p_n, with $p_1 = q$ and $p_n = p$ such than p_{i+1} is directly density-reachable from p_i*

6. *A point p is density-connected to a point q with respect to ϵ and minPts if there is a point o such that both, p and q are density-reachable from o.*

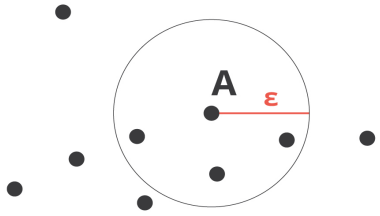

Figure 4.1: Core Point

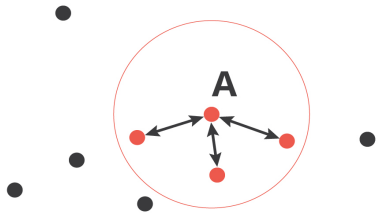

Figure 4.2: Directly density-reachable

7. *Let D be a set of points. A cluster C with respect to ϵ and minPts is a non-empty subset of D satisfying the following conditions:*

   (a) $\forall p, q$: *if* $p \in C$ *and* q *is density-reachable from* p *with respect to* ϵ *and minPts, then* $q \in C$.

   (b) $\forall p, q \in C$: p *is density-connected to* q *with respect to* ϵ *and minPts.*

8. *A point p is considered to be noise if it is neither a core point nor a border point. This implies that noise does not belong to any cluster.*

The paper also provides an algorithm to compute the clusters based on the above defined model. A pseudo code of the clustering algorithm is shown in 1. The database *DB* is linearly scanned for objects that are yet to be processed. When a core point is discovered, its neighbors are iteratively expanded and added to a cluster with a label assignment whereas the non-core points are assigned a noise label. If the neighbors of the core points is again a core point, then their neighborhoods are transitively included in the cluster (*density reachable*). The objects that have been assigned a cluster will be dropped from processing.

The algorithm of DBSCAN serves as an abstraction for the clustering model. Different variants of the algorithm are available such as scikit-learn 0.23 [38], first identify all the neighborhoods and then perform the cluster expansion on the core points that are identified. This does not improve the overall runtime complexity but is more efficient to execute in the Python environment.

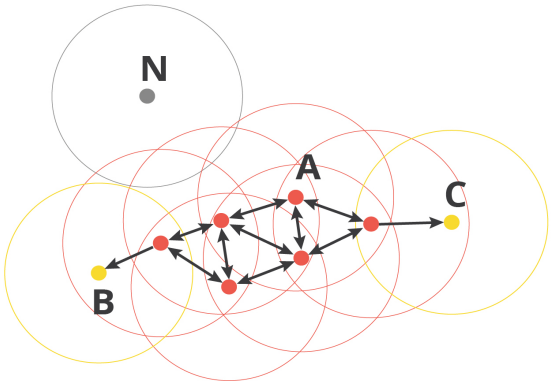

Figure 4.3: Illustration of DBSCAN cluster model (A: core point, B & C: border points, N: Noise point)

Algorithm 1: Pseudocode of DBSCAN Algorithm [39]

Input: *DB:* Database
Input: ϵ : Spatial Threshold or radius
Input: *minPts:* Density threshold
Input: *dist:* Distance function
Data: *label:* Point labels. Initially *undefined*

```
foreach point p in Database DB do                    // Iterate over every point
    if label(p) ≠ undefined then continue             // Skip processed points
    Neighbors N ← RangeQuery(DB, dist, p, ε)          // Find initial neighbors
    if |N| < minPts then                  // Non-core points labelled as noise
        label(p) ← Noise
        continue
    c ← next cluster label                            // Start a new cluster
    label(p) ← c
    Seed set S ← N / {p}                              // Expand neighborhood
    foreach q in S do
        if label(q) = Noise then label(q) ← c
        if label(q) ≠ undefined then continue
        Neighbors N ← RangeQuery(DB, dist, p, ε)
        label(q) ← c
        if |N| < minPts then continue                 // Core-point check
        S ← S ∪ N
    end
end
```

The DBSCAN method provides a possibility to cluster points which are close based on the spatial attribute of the data using the ϵ threshold. The PD data from the SCG system consists of the description of discharge event with its location and time of occurrence. The association of a discharge event with both time and location compels to find a more suitable method to incorporate the temporal attributes of the data and to impose a threshold on the extent of evolution of activity in time and thereby identifying clusters close in the spatial and temporal domain.

4.3 ST-DBSCAN

ST-DBSCAN is a variant of the DBSCAN model and stands for Spatio-Temporal DBSCAN [40]. The ST-DBSCAN model provides a way to use the temporal attribute of the data by defining a temporal threshold. The temporal dimension provides a description of the extent of evolution of the object whereas the spatial dimension provides a description about whether the objects considered are associated with a fixed or moving location [41].

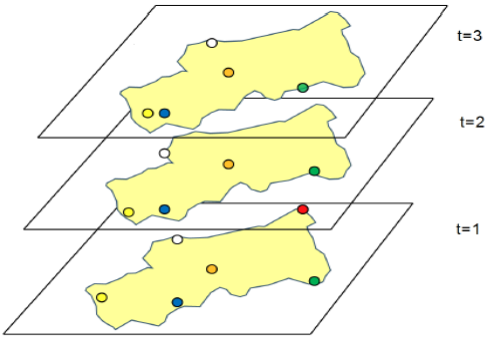

Figure 4.4: Spatio-temporal Representation [42]

In the context of this thesis, the data from the SCG systems (as seen in Figure 3.5) shows an evolution of the discharge activity (or events) in time for locations along the length of the circuit. This makes ST-DBSCAN a beneficial model to capture clusters of discharge activity that are close in time and in space (location). The ST-DBSCAN model uses three parameters namely, the spatial threshold or ϵ_1, the temporal threshold or ϵ_2, and the *minPts* or the minimum number of neighbors. The ϵ_1 measures the closeness of objects in the spatial dimension and the ϵ_2 measures the closeness of objects in the temporal dimension.

The working principle of ST-DBSCAN is such that if the time difference between successive points is within the temporal threshold or ϵ_2 and the distance between those points is is less than or equal to the spatial threshold or ϵ_1, then the points are spatial and temporal neighbors. If a region is dense then the it should satisfy the *minPts* criteria for the core point condition. The values of the parameters do not of have any defined rules and the selection is dependent on the intuition of threshold values based on the application specific to the use case.

For the ϵ_1 parameter, the value is intuitively set based on the information of the SCG accuracy of 1% of the length of the circuit. For the ϵ_2 threshold, there is no one specific value that is effective as the temporal characteristics of the data vary across circuits. To find an optimum threshold value for ϵ_2, the heuristic provided in [37] is employed. The usage of the heuristic is demonstrated in the next chapter during the evaluation of the circuit. The value identified for one circuit is used forward for evaluation of other circuits which share similar circuit characteristics.

4.3.1 Methodology

This section elaborates on the methodology of clustering adopted for the project data. It is divided into three main process steps (Figure 4.5).

To perform the clustering process, the circuit under evaluation is loaded with its associated circuit and PD data. The circuit data provides the configuration information such as the lengths of different types of cable insulation and the locations of RMU(s), joints, and terminations in the circuit. The PD data is the SCG measured PDs in the cable circuit along with the warnings assigned to the circuit.

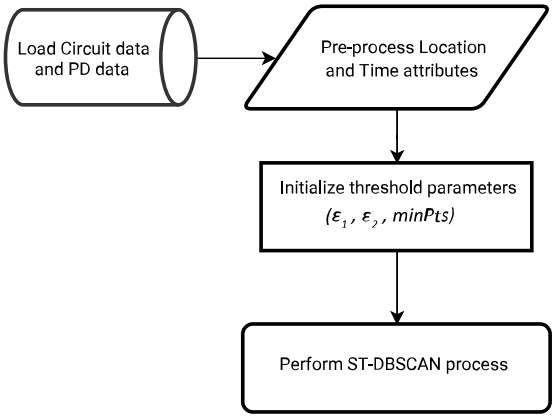

Figure 4.5: Clustering Process

	Date/time (UTC)	Location in meters (m)	Charge (picocoulomb)
0	2017-08-22 00:25:00	0.000000	25141.5
1	2017-08-22 18:24:00	1869.181579	19401.5
2	2017-08-22 21:20:00	1424.902004	14303.0
3	2017-08-22 22:35:00	235.358102	28000.0
4	2017-08-22 23:32:00	1524.610363	20693.5
5	2017-08-23 06:26:00	0.000000	22192.5
6	2017-08-23 15:25:00	1162.877714	11320.0
7	2017-08-23 17:13:00	351.298053	9921.0
8	2017-08-23 22:09:00	309.559671	21382.0
9	2017-08-24 00:16:00	1350.700435	20416.0

Figure 4.6: SCG data

The SCG data has three attributes - the timestamp of the discharge event, the location where it was observed, and the magnitude of discharge. For the clustering process, the location and timestamp of the discharge events is considered for identifying clusters in the spatial-temporal domain. The clustering method uses a distance measure to calculate the similarity or closeness of points both in location and time. The timestamps of the data are converted into hourly measurements based on the start date of the data. The location attribute consists of numerical data of floating type. For the location attribute, the values are discretized using the equal bin-width technique into 10m wide bins. The equal bin-width technique divides the range of observed values into k equal sized bins. The parameter k is set according to the length of the circuit. The discretization allows for converting the continuous datatype into discrete location intervals. This is helpful in aggregating points within the 10m interval and assigning their counts as weights for the measurement. Additionally, for the timestamps, the hourly resolution is chosen due to the discharge activity, in most cases, is spread for hours leading up to days. This can also be observed in the Figure 3.5, where the activity at locations (300-600) is dense and spread over for days/months. The discretization of the time and location attributes helps in aggregating the points which are similar in the dataset and assign their counts as weights. This is done because of the large amounts of data present in the PD data table. Since, ST-DBSCAN computes a distance matrix for all the points in the dataset this makes it computationally expensive in the cases of large amounts of observation points and hence discretization allows grouping of similar points and perform the distance matrix calculation easily.

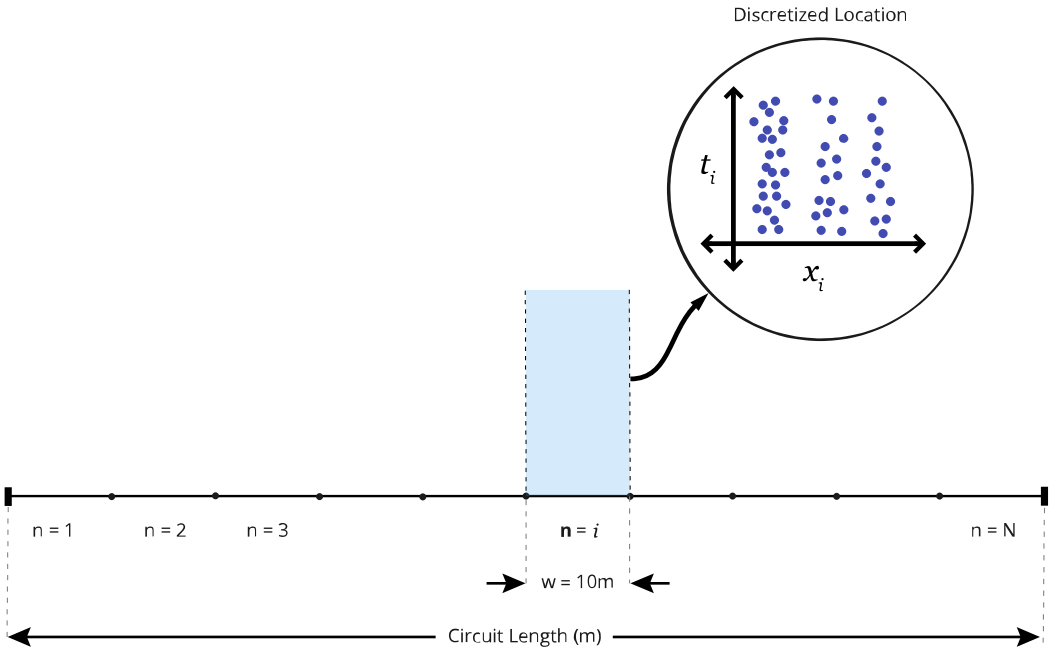

Figure 4.7: Discretization of the location attribute

Next, the parameters for the ST-DBSCAN clustering process are initialized. ST-DBSCAN requires three parameters - the spatial threshold (ϵ_1), the temporal threshold (ϵ_2), and the *minPts*. The threshold parameters can be set intuitively based on the data or can be set using an heuristic provided [37]. For the spatial threshold (ϵ_1), the parameter value is set intuitively at 1 bin-width or 10m due to location discretization. This is motivated by two reasons- (i) the location accuracy of the SCG sensors is close to 1% of the length of the circuit (domain knowledge) and median circuit length for the actively monitored circuits is close to 3000m so a 10m radius provides higher resolution in finding tight clusters, and (ii) circuit lengths which are less than 1000m are less than 10% of the total circuits, for which the bin-width can be set at 5m. For the temporal threshold (ϵ_2), it is difficult to set the parameter value intuitively for all the circuits, as the discharge activity may be spread for days or months for some circuits whereas it may be just spread for a few hours or days in others. Hence, a single value for all circuits may not be beneficial in identifying clusters. The value for the temporal threshold can be estimated using the heuristic provided [37] by setting the *minPts* parameter. The *minPts* parameter serves the purpose of smoothening the density estimate of the data [39]. Although papers [37, 40, 43] propose several ways to set the *minPts* parameter but with datasets with a lot of noise or that are very large it is favourable to set the *minPts* to higher value for better results [39].

Once the parameters for the spatial threshold (ϵ_1), temporal threshold (ϵ_2), and *minPts* are initialized for the circuit under evaluation, the next step that remains is the clustering process. A pseudo code for the ST-DBSCAN algorithm developed for this project is shown in Algorithm

2.

Algorithm 2: Pseudocode of ST-DBSCAN Algorithm [44]

Input: *DB:* Database
Input: ϵ_1 : Spatial Threshold
Input: ϵ_2 : Temporal Threshold
Input: *minPts:* Density threshold
Input: *dist:* Distance function
Data: *label:* Point labels. Initially *undefined*

```
1 dist(N_t) ← RetrieveNeighbors(DB, ε_2, dist)        // Temporal Neighbors
2 dist(N_st) ← RetrieveNeighbors(DB, ε_1, dist(N_t))   // Spatial Neighbors
3 DBSCAN(DB, ε_1, minPts, dist(N_st))                  // Apply DBSCAN model
4 Assign labels
```

The algorithm inputs are the Database DB (or the PD data for the circuit), the clustering parameters (ϵ_1, ϵ_2, *minPts*), and the distance measure or *dist* for the calculating the distance between the points in location-time which is the Euclidean distance used here. The RetrieveNeighbors function is called twice. First, to support the temporal characteristics of the data, data is filtered by retrieving the temporal neighbors and their corresponding spatial values by calculating the distance matrix with respect to the temporal threshold. For points whose distances satisfy the temporal threshold are retained with their corresponding spatial values and for points that do not satisfy the threshold the spatial distance for those points is set twice the spatial threshold and will evntually be labelled as noise points. The spatial distances retrieved from line 1 are then sent to the RetrieveNeighbors function to compute the distance matrix of spatial neighbors. The entire dataset is then passed to Dbscan along with the pre-computed distance matrix and the parameter values ϵ_1, *minPts*. The Dbscan function here represents the function provided by the sklearn library [38] in Python environment which gives the functionality to input a pre-computed distance matrix and not perform the distance computation in the DBSCAN method. The distance matrix is simply evaluated for identifying core points and their neighborhoods (density-reachability) and to carry out the label assignment.

4.3.2 Performance Indicators

The clusters identified for the circuit from the clustering process are compared against the DNV GL warnings assigned to that circuit. Performance indicators are formulated to evaluate the clustering. The efficiency of the algorithm is quantitatively expressed using performance indicators.

The DNV GL warnings assessment is considered as the ground truth for the evaluation of clusters. The performance of the clustering is assessed for Level 1, 2, and 3 warnings whereas the Noise warnings are dropped out of evaluation. To evaluate the warnings with respect to the clusters found the warnings are converted into a bounding box of length equivalent to the time duration of the warning and the width is equivalent to the warning assigned to the location $\pm$ 0.5m. For the clusters, the bounding box is created with width equivalent to the range of the location of the cluster and the length is equivalent to the duration for which the cluster was detected. To detect if a warning is present inside a cluster, the overlap between the bounding box of the warnings and the bounding box of the clusters are checked.

The following performance indicators are formulated:-

1. For a cluster identified, if there is a DNV GL warning detected in it then the identified cluster is assigned as a True Positive (TP).

2. For a cluster identified, if there is no DNV GL warning detected in it then the identified cluster is assigned as a False Positive (FP)

3. If there is no cluster detected for a DNV GL warning then the identified cluster is assigned as a False Negative (FN).

The performance indicators are used to calculate metrics such as precision and recall for the clustering algorithm. Precision is defined as the fraction of correctly identified clusters among all identified clusters (with or without warnings) whereas the Recall is defined as the fraction of correctly identified clusters among all detected clusters (i.e., clusters with warnings).

$$Precision = \frac{\text{TP}}{\text{TP + FP}} \quad \text{and} \quad Recall = \frac{\text{TP}}{\text{TP + FN}} \quad \text{and} \quad F1 = 2\left(\frac{Precision \times Recall}{Precision + Recall}\right)$$

The precision and recall scores are used to compute the F1 score which provides a harmonic mean of the two scores.

4.4 Summary

In Section 4.1, the characteristics of different type of clusterings along with the various models that employ the clustering methods are discussed. Here, the advantage of using a density-based clustering method is motivated based on the PD data received from the SCG system and the method employed for this thesis is identified.

In Section 4.2, the DBSCAN clustering model is explained with the formal definitions used by the model to identify clusters.

In Section 4.3, the ST-DBSCAN method which is a variant of the DBSCAN model is presented and the description of the parameters in context to the project are discussed. The methodology of the clustering method is described in detail and the performance indicators and metrics formulated to evaluate the performance of the clustering method are discussed.

The next chapter discusses different cases that are used to perform the clustering and their results.

Chapter 5

Results and Discussion

In this chapter, five circuits considered for the evaluation of the clustering method are discussed. For each circuit, the description of the circuit configuration is provided along with the DNV GL warnings assigned to the circuit. Next, we tune the parameters in case 1 and perform the clustering on the subsequent circuits. The clustering is evaluated using the performance metrics mentioned in the previous chapter. The clusters identified from the clustering method are analysed and features of the clusters are presented.

5.1 Case I

This circuit consists of mixture of PILC and XLPE type cables connected together with 23 joints (dashed blue lines, see Figure 5.1) of several type of insulation and and also consists of 4 Ring Main Units (RMUs) (solid green lines, see Figure 5.1). The length of the entire circuit is 2299 metres. The PD data received from the SCG system is graphically illustrated in Figure 5.1. From the figure, 4 clusters of varying density are visible at locations around 250m to 400m and locations around 500m to 650m.

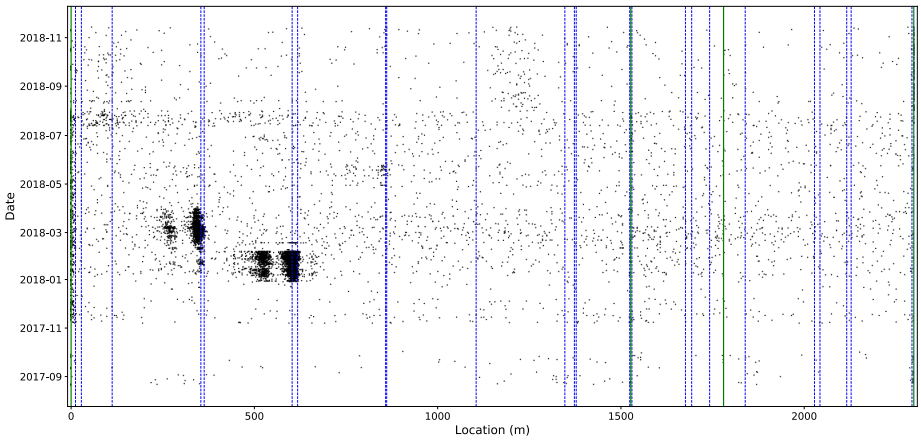

Figure 5.1: Location vs Date/Time PD data for the circuit

The warnings assigned by DNV GL experts for the circuit are represented in the following table:

Table 5.1: Warnings associated with the circuit

Location (in metres)	Start Date/Time	End Date/Time	SCG warning level (1, 2, 3, N)
604	2017-12-30 19:48:43	2018-02-05 00:03:25	3
604	2018-02-05 00:03:25	2018-02-23 16:08:12	2
349	2018-02-26 22:40:58	2018-05-27 08:45:43	3

In the table, we see that two types of warnings were assigned to location 604m (level 2 and 3) and a level 3 warning was assigned at location 349m of the circuit. The warning are projected on the PD frame and illustrated on a magnified view of the locations in Figure 5.2. The yellow bands represent the level 3 warnings and the level 2 warning is represented with an orange band.

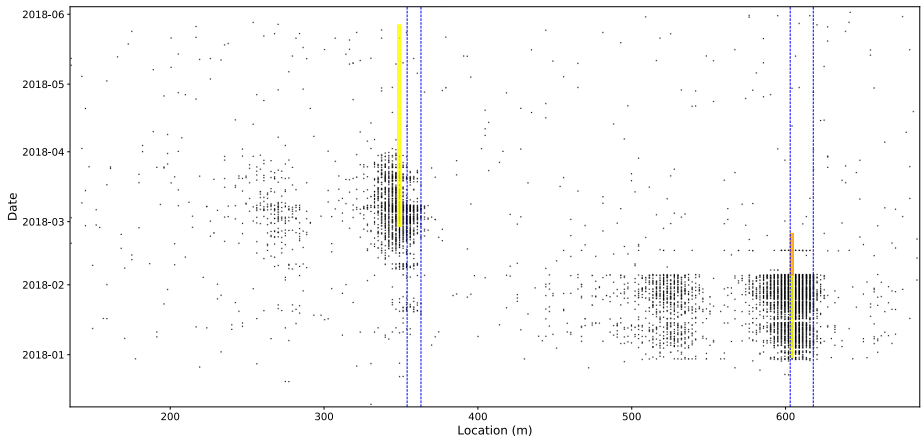

Figure 5.2: Warnings assigned at locations 349m and 604m. *(Level 2: orange band, Level 3: yellow band)*

Parameter Tuning

The parameter values of the clustering algorithm are initialized, with ϵ_1 threshold set at 1 bin-width or 10 metres. To set the value for ϵ_2, the heuristic presented in [37] is used to estimate the parameter value. The heuristic suggests calculating the k-th neighbor distance for each point in the dataset. The parameter k is set using the *minPts*. The distances calculated are sorted in descending order of their k-distance values which provides some hints in the density distribution of the dataset. Then for an arbitrary point p if the ϵ_2 value is set at k-dist(p) then all points with equal or smaller k-dist values will be core points. To find an optimal value for the ϵ_2, the threshold point is ideally the first point in the first 'valley' or 'knee-point' (from the left side) in the sorted k-distance graph [37].

For the ϵ_2 threshold, the sorted k-th distances are calculated for $k = (100, 200, 300, 400, 500)$ which are the sets of *minPts* values that are evaluated. The distances are plotted in Figure 5.3. The horizontal dashed lines are the ϵ_2 bands of values from 100 to 500. It is evident from the graph that the number of 'valleys' in the distance plots increases with higher value of k.

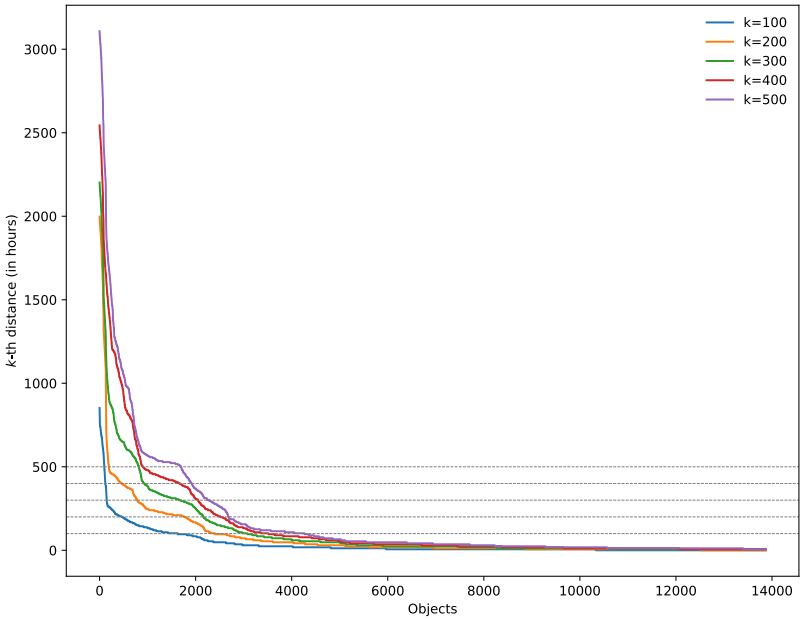

Figure 5.3: Sorted k-th neighbors distance plot for $k = (100, 200, 300, 400, 500)$.

This makes it difficult to determine the correct threshold point or 'valley' which would help in yielding thin clusters. With k value of 100, the first valley is detected between bands 200 to 300. The discrete steps visible are due to the low numerical precision of the temporal attribute since the temporal resolution is set at hourly precision.

Clustering results

The clustering is performed with the following parameter values:

- **Spatial threshold or ϵ_1:** 10m (1 bin-width)
- **Temporal threshold or ϵ_2:** 250 hours (estimated from Figure 5.3)
- ***minPts*:** 100

The result of the clustering is illustrated in the following plot (Figure 5.4).

From the Figure 5.4, it is evident that the estimation of the ϵ_2 parameter value is able to identify the four visible clusters. The clusters identified are checked for overlap of warnings from DNV GL by creating bounding box of the warning and clusters. From Figure 5.5, clusters 1 and 3 do not have an overlap with the DNV GL warnings whereas for the DNV GL warnings (ref. Table 5.1) are detected forming an overlap with clusters 2 and 4. Therefore, clusters 1 and 3 are the false positives identified by the clustering algorithm and clusters 2 and 4 are true positives with DNV GL warnings present.

For the above, parameters the performance indicators are evaluated with the Precision, Recall and F1 scores.

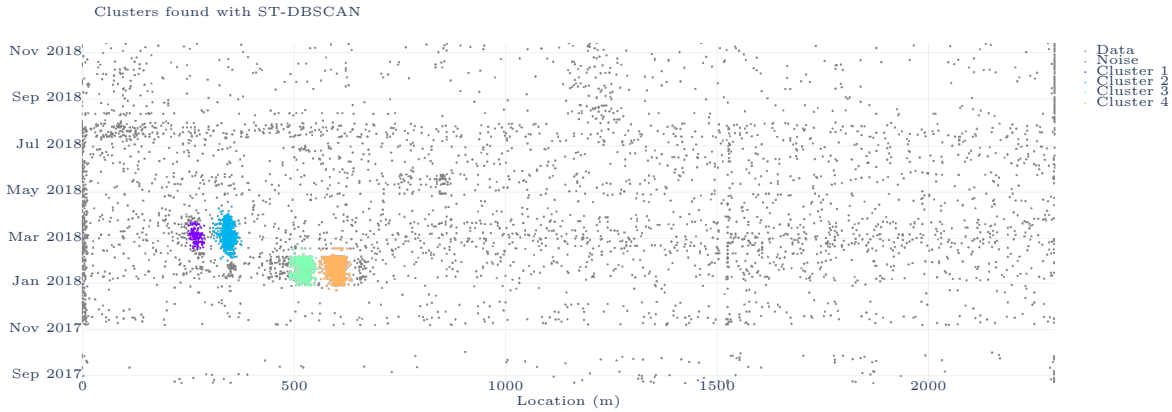

Figure 5.4: Clustering result

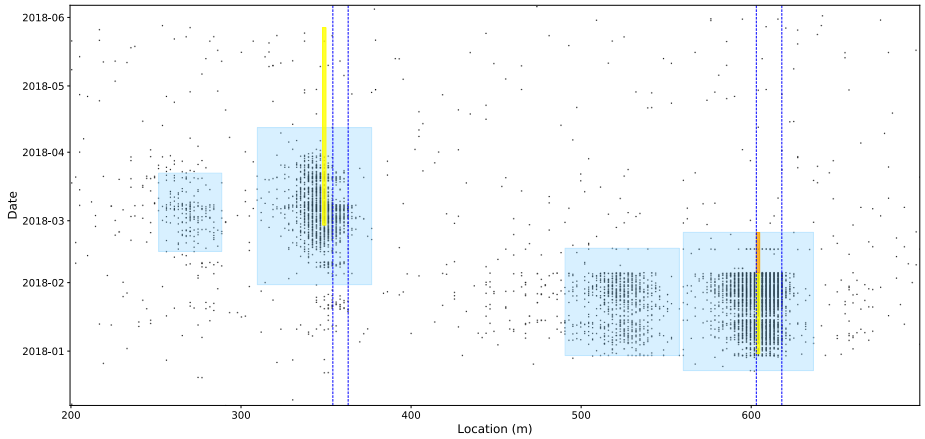

Figure 5.5: Overlap of clusters with the DNV GL warnings

Performance Indicators					
True Positives	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
3	2	0	0.6	1	0.75

Cluster Features

For the identified clusters, it is interesting to evaluate the clusters by calculating the features of the clusters. Features help in defining the characteristics of a cluster and help in providing insight into the behaviour of the discharge activity. The features identified are summarized in the Table 5.2.

To begin with, the density of the cluster is calculated by counting the number of discharges and dividing them by the duration (start and end time) of the cluster. From the Table 5.2, the density of clusters 1 and 3 is low compared to the clusters 2 and 4. This serves as a good

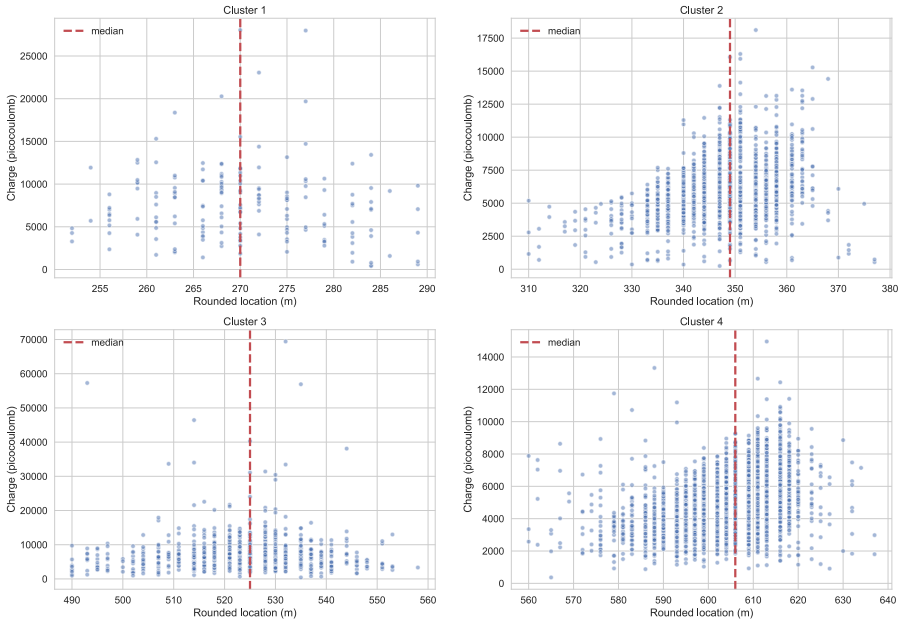

Figure 5.6: Location vs Charge scatterplot of cluster

distinction between the clusters with DNV GL warnings (true positives) versus the clusters without DNV GL warnings (false positives).

The location versus charge scatter plots are plotted to visualize the cluster in the front view (Figure 5.6). This helps in identifying the width of the cluster which is calculated by taking the difference of the maximum and minimum of the location values of the cluster. The location attribute is further evaluated by calculating the statistical moments such as the mean, standard deviation, skew, and kurtosis which help in providing the shape characteristics of the clusters. The median locations of the clusters 2 and 4 is same as the warnings assigned for locations (Table 5.1).

The charge attribute helps in identifying the distribution of charge in the cluster (Figure 5.7). This provides information into the magnitude of the extent of discharge activity that was observed in the identified cluster. The statistical moments are also calculated for the charge attribute. For clusters 1 and 3, the maximum magnitude of charge is higher compared to clusters 2 and 4. This is also evident from the mean and standard deviation of the charge in clusters. The skew characteristic of the charge shows closeness to a normal distribution for clusters 2 and 4 whereas it is positively skewed for clusters 1 and 3.

The time attribute of the clusters can be utilized by calculating the duration (height) of the cluster. This is done by taking the difference between the start and end time of the cluster activity. The temporal feature handling provides both the 0^{th}-order and 1^{st}-order handling of the temporal characteristics of the clusters. The 0^{th}-order feature helps in bucketizing the timestamp of the discharges into number of days, hours, minutes etc. and the temporal resolution chosen for the clustering process is hourly resolution of the discharges. The temporal characteristic of the data can be further taken advantage of by calculating the inter-arrival time of the discharges present inside the cluster (Figure 5.8). Inter-arrival time is defined as the time

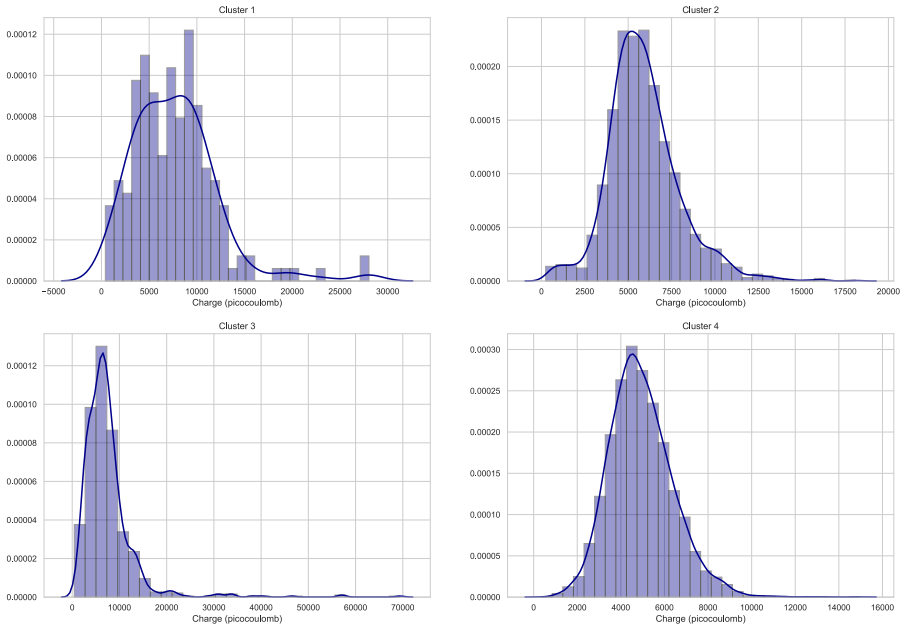

Figure 5.7: Charge distribution in the identified clusters

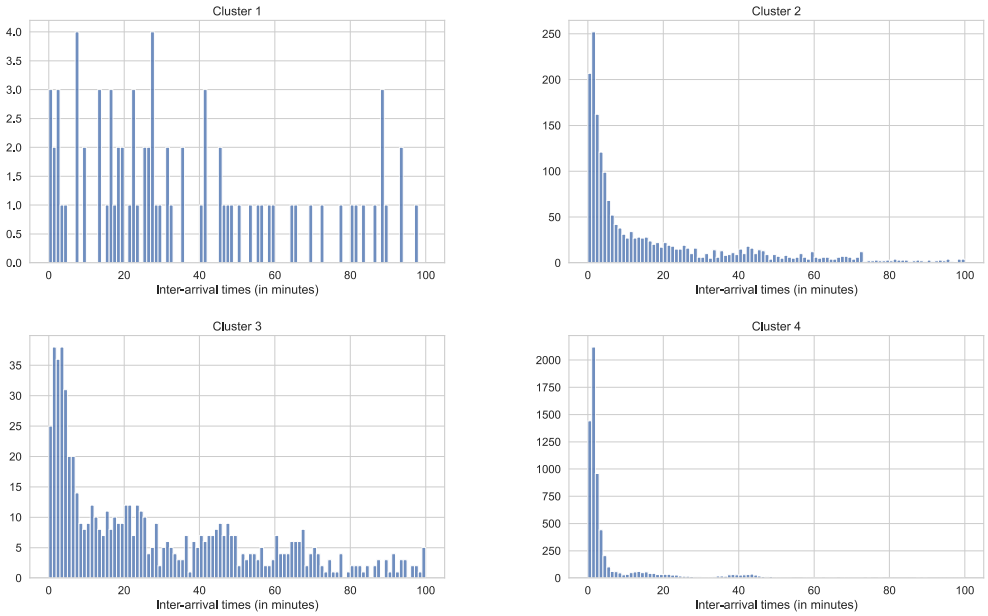

Figure 5.8: Histogram of inter-arrival times of discharge events within the cluster

interval between two consecutive discharge events within the cluster. The inter-arrival time of discharges within the clusters serve as the 1^{st}-order handling of the temporal feature. From

the Figure 5.8, the histogram of inter-arrival times measured in minutes is plotted for all the clusters. The histograms are binned for 100 minutes of the inter-arrival times with a bin width of 1 minute. Cluster 1 shows an interrupted activity of the discharge events within it. For cluster 3, there are multiple peaks observed in the occurrence of discharge events happening inside the cluster. Clusters 2 and 4 show a high number of discharge events happening in quick succession. Since the reporting of the PDs from the SCG is limited to 1 minute the inter-arrival times can only be measured with maximum resolution of 1 minute.

Apart from the features extracted from the attributes of the data, the circuit configuration data can be utilized to identify circuit specific features for the clusters. For the median location of the cluster, the distance from the nearest joint and RMU location are found out. This gives the information of how close the cluster identified was from a joint as the most likely defects or PDs are observed in joints. One of the assessments from the DNV GL experts was that during high loading cycles vibrations in the RMU may be picked up as PD signals by the SCG system and are characterized by a repititive pattern of discharges in time and hence such patterns in the PD data are regarded as noise. The Table 5.2 summarizes the distances measured for the median cluster locations. For clusters 2 and 4, it is evident that the discharge activity was observed in a joint whereas clusters 1 and 3 appear far away from the nearest joint location. Since the nearest RMU located is at the 0m of the circuit, the distance from the nearest RMU location does not provide any information in this case.

Discussion

Case I demonstrates the method involved in setting the parameter values for the temporal threshold using the heuristic to estimate the threshold value. The result of the clustering is able to capture the visible denser regions in the PD data from the otherwise sparse discharge events in the rest of the circuit. The process of extracting features that define the characteristics of the identified clusters is illustrated. The case was demonstrated at the start as the warning assigned for location 604m (level 3 and level 2) is a known and most commonly used example of a partial discharge happening in the circuit of Alliander's medium voltage circuit database. This motivated the implementation and testing of the clustering method to ensure the clustering method is able to identify clusters and capture the warnings assigned by DNV GL.

Table 5.2: Cluster Features

Cluster Density (Discharges per day)			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
12.11	564.45	78.73	6268.60

Cluster Width (in metres)			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
37	67	68	77

Location (in metres): Descriptive Statistics

Cluster	Min	Max	Mean	Median	St. Dev	Skew	Kurtosis
1	252	289	270.24	270	8.68	0.13	-0.52
2	310	377	347.90	349	8.53	-0.62	1.37
3	490	558	523.47	525	12.48	-0.36	0.15
4	560	637	605.03	606	8.01	-1.02	2.83

Charge (pC): Descriptive Statistics

Cluster	Min	Max	Mean	Median	St. Dev.	Skew	Kurtosis
1	407.50	28039.00	7645.05	7300.50	4470.63	1.47	4.49
2	244.50	18110.50	5897.13	5674.25	2077.25	0.82	2.41
3	436.00	69419.00	7480.45	6493.50	5786.62	4.71	35.21
4	366.00	14957.50	4936.50	4805.50	1455.06	0.58	1.14

Cluster Height (in days)			
Cluster 1	Cluster 2	Cluster 3	Cluster 4
35 days 11:14:00	71 days 01:28:00	48 days 13:29:00	62 days 13:49:00

Inter-arrival time [△, IAT] (in minutes): Descriptive Statistics

Cluster	Min	Max	Mean	Median	St. Dev.	Skew	Kurtosis
1	0	2353	286.93	121	412.25	2.54	7.77
2	0	4926	49.14	8.5	234.04	14.99	262.20
3	0	6432	86.43	27	324.09	13.91	238.28
4	0	10721	13.62	1	201.41	42.48	2011.27

Circuit specific features

Cluster	Median Location (m)	Distance from the nearest joint location	Distance from the nearest RMU location
1	270	84	270
2	349	5	349
3	525	78	525
4	606	3	606

5.2 Case II

This circuit shares similar characteristics as the previous circuit. It consists of 23 joints and 5 RMUs connecting a combination of PILC and XLPE type cables. The length of the circuit is 2240m which is also close to the case discussed above. The PD data from SCG system is graphically illustrated in Figure 5.9. It can be seen from the plot that the PD data represented in location versus the timestamp of the discharge activity consists of a noisy background with a few clusters visible. The actual number of clusters is difficult to determine solely from visual inspection unlike the previous case.

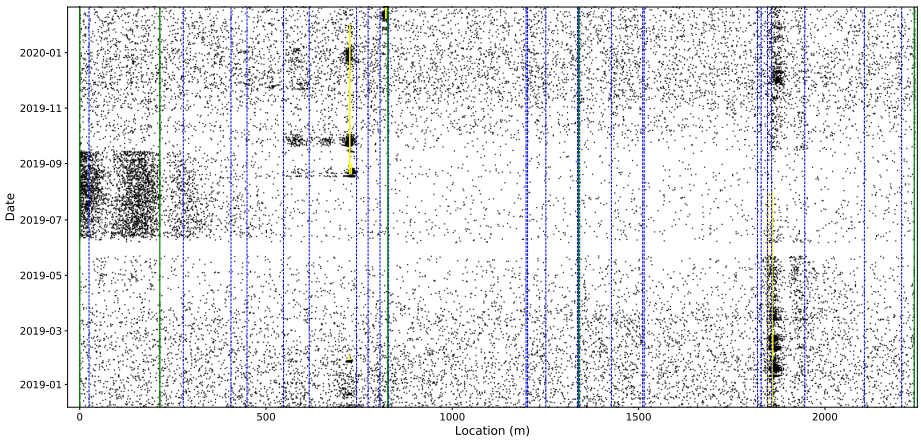

Figure 5.9: Location vs Date/Time scatter plot of PD data with warnings from DNV GL plotted (*yellow bands*)

The warnings associated with the circuit are represented in the Table 5.7. It is evident from the warnings table that the three warnings at locations 724m, 725m, and 729m and three warnings at locations 1846m, 1858m, 1860m are closer in locations. On visualising the warnings assigned to the circuit, one of the warnings assigned to location 724m showed an interesting ambiguity in the start of actual activity and the start time of the warning assigned by the experts at DNV GL. The warning assignment is a manual operation where an expert at DNV GL oversees the PD data associated with the circuit and based on the activity observed assigns a warning to signal the operator. Therefore, there can be a difference in the start time of the activity and the eventual assessment and warning assignment by the DNV GL expert. This can be seen in the warning band plotted on the PD data (Figure 5.19).

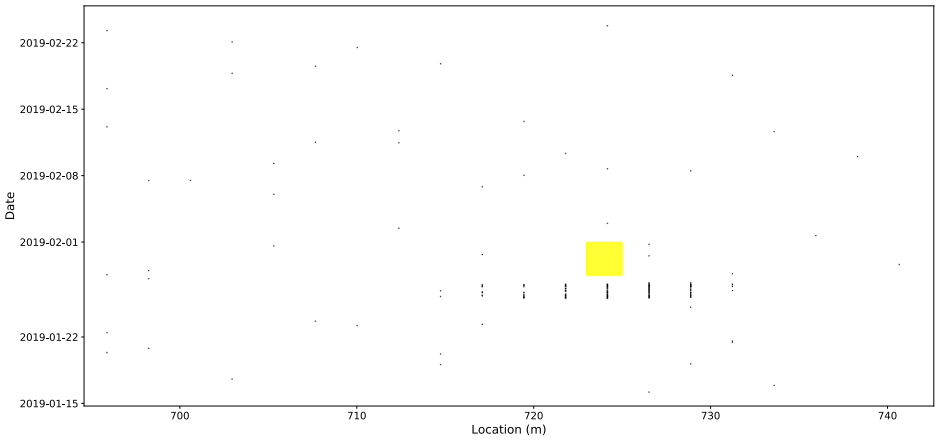

Figure 5.10: Warning assigned to location 724m

This mismatch in the overlap of the warning assignment and the start time of activity can have severe downside when evaluating the performance of the clustering as there might not be any overlap detected and the cluster would be assigned as a false positive and a false negative would be assigned for a cluster not found for the DNV GL warning. To overcome this issue, upon consultation with the DNV GL experts it was decided to introduce a buffer of 2 weeks that would be added to the timestamp of the warning for the start and end of the warning. The buffer for the end of the warning can be considered optional.

Table 5.3: Warnings associated with the circuit

Location (in metres)	Start Date/Time	End Date/Time	SCG warning level (1, 2, 3, N)
724	28-01-2019 14:21:41	31-01-2019 23:59:00	3
725	20-08-2019 14:52:07	31-01-2020 23:53:00	3
729	20-08-2019 14:52:07	30-08-2019 22:46:00	3
822	07-02-2020 16:51:39	29-02-2020 23:25:00	3
1846	10-11-2017 16:10:15	31-07-2019 23:53:00	3
1858	10-11-2017 16:10:15	31-07-2019 23:53:00	3
1860	28-01-2019 14:21:52	28-02-2019 23:51:00	3

Clustering results

Since the characteristic of the circuit match with the previous case, the clustering is performed with the same parameter values.

The parameter values are set as follows:- ϵ_1=10 metres, ϵ_2=250 hours, *minPts*=100

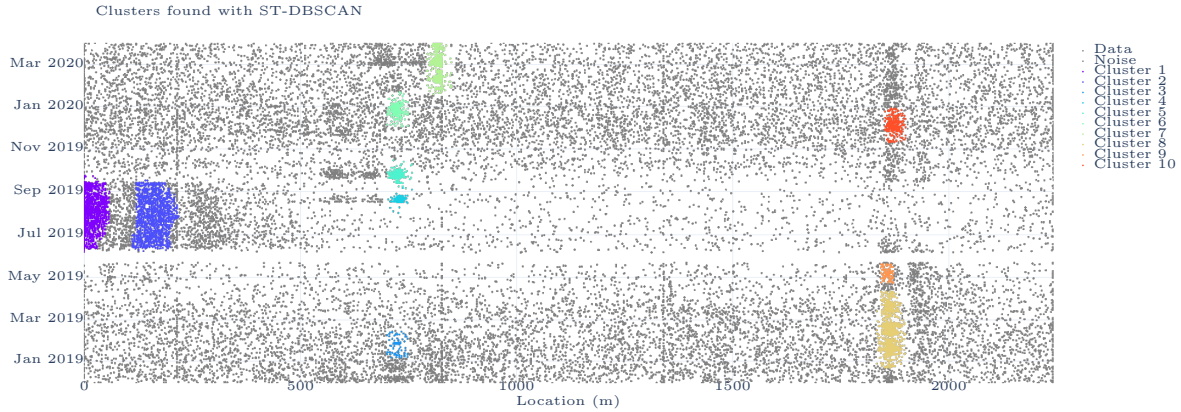

Figure 5.11: Clustering result with parameter set 1. *Clusters labelled from left to right (bottom cluster first).*

Table 5.4: Performance Indicators

Parameter set (ϵ_1=10m, ϵ_2=250, *minPts*=100)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
10	12	3	0	0.80	1	0.89
Parameter set (ϵ_1=10m, ϵ_2=24, *minPts*=100)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
12	21	0	0	1	1	1

The clustering results show that using the parameter values from the previous case, the algorithm yeilds 10 clusters of which 3 clusters do not have an overlap with any DNV GL warning. Out of the three clusters, the clusters at the start of the circuit show a repetitive pattern in the discharges and are closer to the RMU placed in the circuit and are hence regarded as noise points that are clustered. Although there are no false negatives, upon closer inspection of the clusters, on magnifying the cluster no. 8 it is visible that the cluster formed does not form a thin and tight cluster but in fact has several clusters merged together. This is attributed to the large value of the temporal threshold which coupled with a lower minimum number of points leads to a merging of multiple clusters together. The clusters identified have several 'outliers' or noise points which are included in the cluster due to large temporal threshold.

To identify thin clusters, the ϵ_2 parameter value is estimated in the same way as the previous case. With the *minPts* set at 100, the k-th neighbor distance is plotted (Figure 5.12). From the graph, the 'valley' or the 'knee-point' of the distance graph is visually detected to lie closer

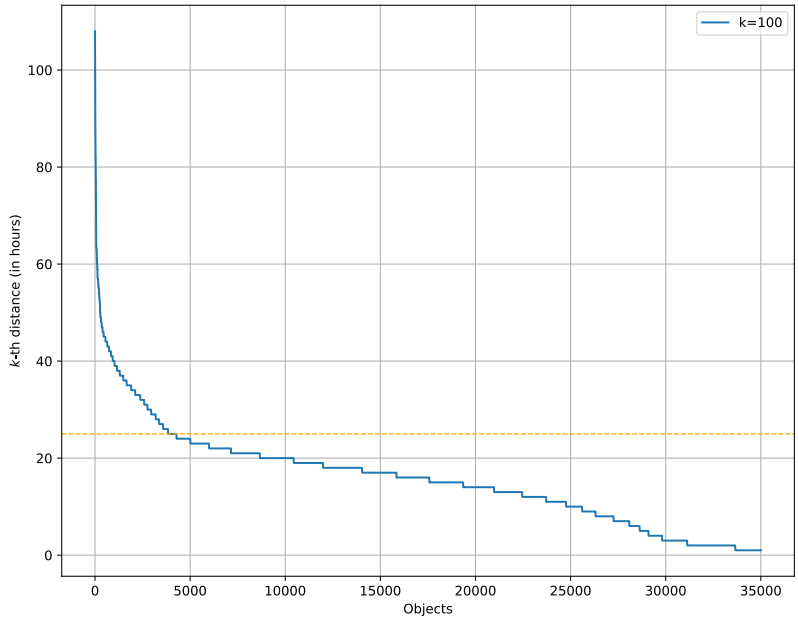

Figure 5.12: ϵ_2 parameter tuning

to 20 hours. An orange dashed line is plotted across the graph to set the threshold value at 24 hours.

The updated parameters of the clustering algorithm are set as follows:- ϵ_1=10 metres, ϵ_2=24 hours, *minPts*=100.

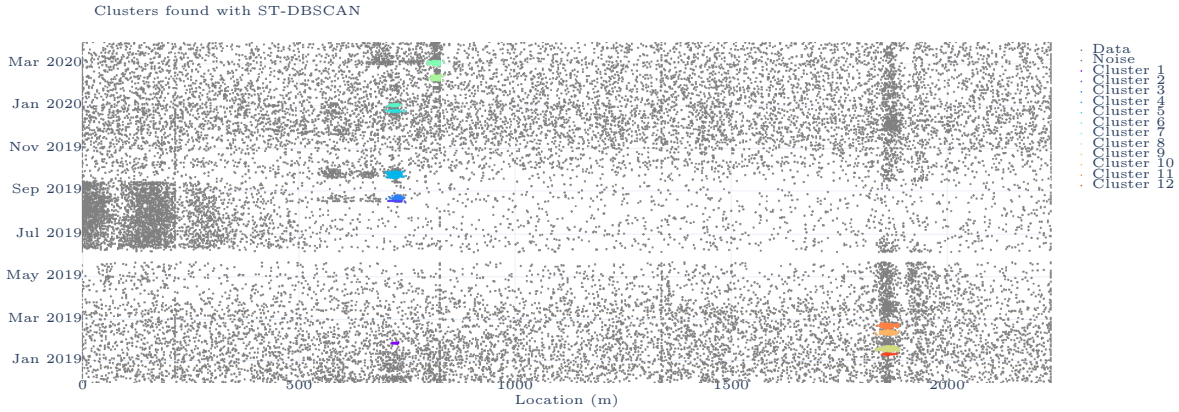

Figure 5.13: Clustering result with parameter set 2. *Clusters labelled from left to right (bottom cluster first).*

On running the clustering algorithm for the updated parameters, 12 clusters are found with no false positives or false negatives yielding the overall F1-score as 1 (Table 5.4). Upon visually

inspecting the clusters, it is evident that a lower temporal threshold is beneficial in achieving thin and tight clusters. The clusters along the terminations of the cable are also not detected as the discharge activity is not so dense within the 24 hour threshold.

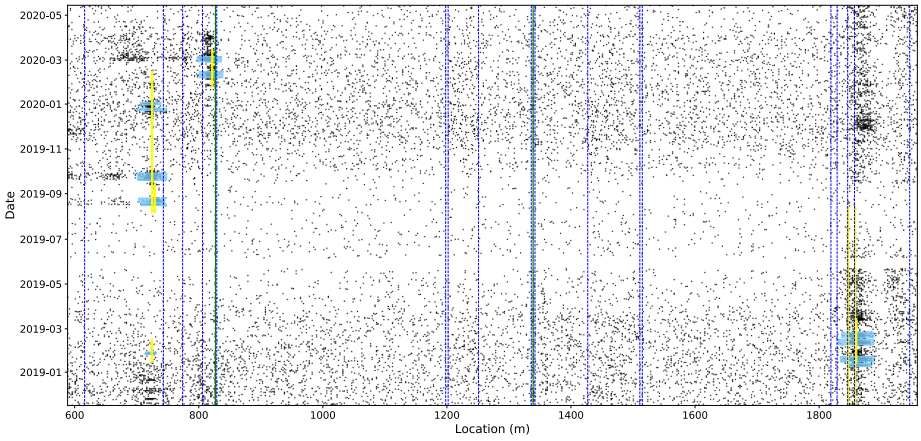

Figure 5.14: Overlap of clusters with the DNV GL warnings

However, a lower temporal threshold can also be detrimental in identifying clusters which may have been missed by the DNV GL experts. As in the above two parameter sets, although the first set of parameter values performed poorly in terms of identifying tight clusters, there were additional three cluster (or false positives) which were identified. These clusters can later be evaluated by extracting their features and may provide insight into the activity that was captured within the cluster and thereby providing more clusters to process in the classification stage. This is not possible in the second set of parameter values due to low temporal threshold and low number of *minPts*.

Discussion

The PD data presented in Case II showed a more strong surrounding 'noise' or infrequent discharges around the dense clusters visible near joints and RMUs. An inconsistency with the warning assignment was encountered which was corrected by adding a buffer of 2 weeks to start of the timestamp of the warning received from DNV GL. The clustering method re-used the parameters estimated in the previous case and was able to identify high-density regions in the PD data. For the first parameter set, the clustering method identified 10 clusters out of which 3 clusters did not have any overlap with any DNV GL warning. Out of the 3 clusters, two clusters were found close to an underlying RMU in the circuit. On discussion with the DNV GL experts to understand why these regions of discharges were not signalled as warnings for the circuit, the assessment that was reached during the discussion was that due to high amount of vibrations in the RMU, the SCG systems may pick up on these vibrations as discharges in the circuit. These events captured near the RMUs in the PD data is likely to be noise due to repetitive pattern of discharge events. Upon tuning the temporal threshold for the circuit the estimated threshold - depicted visually - suggested a lower temporal threshold. On clustering with the updated temporal threshold it was observed that the previously found 3 clusters were not clustered and instead of one big cluster (cluster 8) there were 4 thin clusters identified.

5.3 Case III

In this case, the circuit under evaluation is double the length of the previous two circuits with a circuit length of 4818m. It has 22 joints and 8 RMUs connecting the cable which is entirely of the PILC insulation type. From the PD data plotted the data has visibly uniform distribution of discharge events with a minor activity present near the termination location at 0m of the circuit whereas the activity near location 2177m is not easily visible in the plot. The activity along with the warning assigned to it is plotted on to the PD data and can be seen in Figure 5.15.

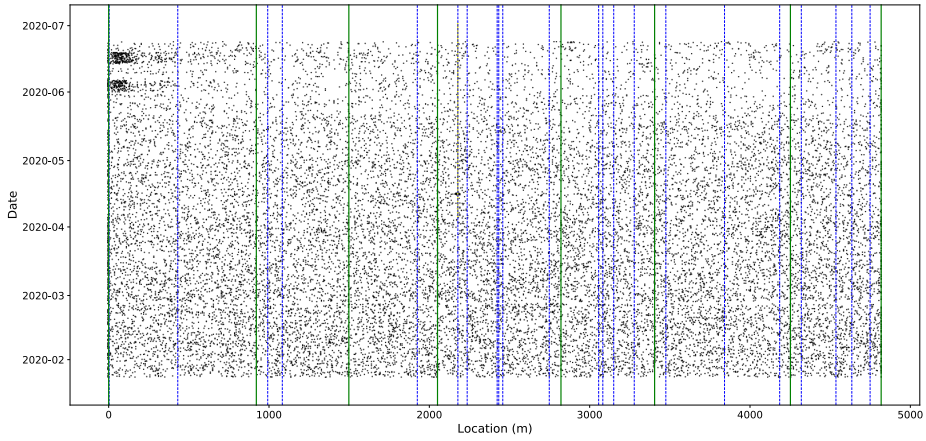

Figure 5.15: Location vs Date/Time PD data for the circuit

There was one warning assigned to this circuit at location 2177m with a warning level 3. The warning information is represented in the table below:

Table 5.5: Warnings associated with the circuit

Location (in metres)	Start Date/Time	End Date/Time	SCG warning level (1, 2, 3, N)
2177	19-04-2020 10:15:05	18-06-2020 03:58:38	3

To perform the clustering, the parameters estimated from the previous two cases are reused to evaluate whether they are transferable to other circuits with varying circuit length. On using the first parameter set of $\epsilon_1 = 10$m, $\epsilon_2 = 250$ hours, and $minPts = 100$, the clustering identifies two clusters at locations near the start of the circuit (50-90m) and at location around 2177m. The clustering performance yields one true positive and one false positive (for locations 50-90m). Upon a closer visual inspection of the clusters it is found that a lot of noise points are also clustered for the cluster found near location 2177m whereas the activity present near the termination of the circuit shows a sparse activity resembling noise which may be picked up by the SCG systems because of the close proximity to RMU.

Using the parameter set of $\epsilon_1 = 10$m, $\epsilon_2 = 24$ hours, and $minPts = 100$, the clustering yields a better result with a tight cluster identified at location 2177m. The activity at locations 50 - 90m is not clustered as previously seen visually that it is not dense when a 24 hour temporal threshold is set for the clustering method.

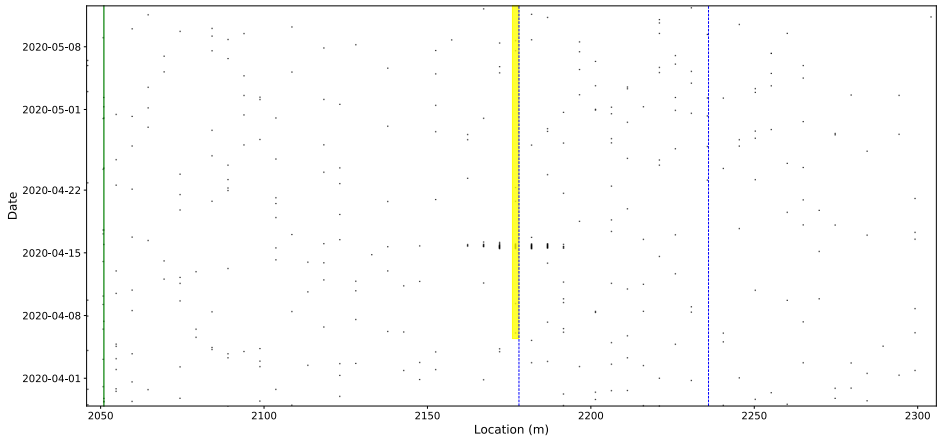

Figure 5.16: Warning assigned for location 2177m on PD data (magnified view)

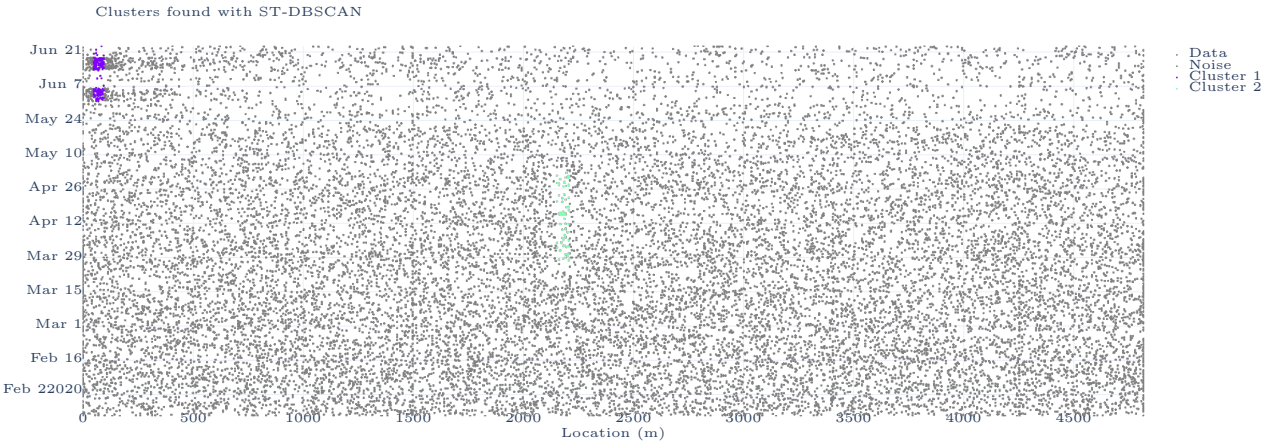

Figure 5.17: Clustering result with parameter set 1. *Clusters labelled from left to right.*

Discussion

In Case III the circuit has different circuit characteristics such as longer length of the circuit and also the type of the cable in the circuit is of PILC type solely. Instead of tuning the temporal threshold, the clustering is performed using the parameter sets from the previous two cases. The clustering from the parameter set 1 values identified 2 clusters of which one was identified correctly and the other 'false positive' was identified near the RMU at the termination of the circuit. Based on the inference from the previous case the parameter set 1 performs well with identifying the actual warning present and also an additional cluster which can be evaluated as noise by identifying cluster features which are in context to the circuit (distance from the RMU, inter-arrival time). On visual inspection of the clustering it is noted that the cluster identified at location 2177m had quite a number of discharge events with longer inter-arrival times with 67 number of discharges with inter-arrival time greater than 20 minutes. Clustering using the parameter set 2 identified a thin cluster at location 2177m and the number of discharges with

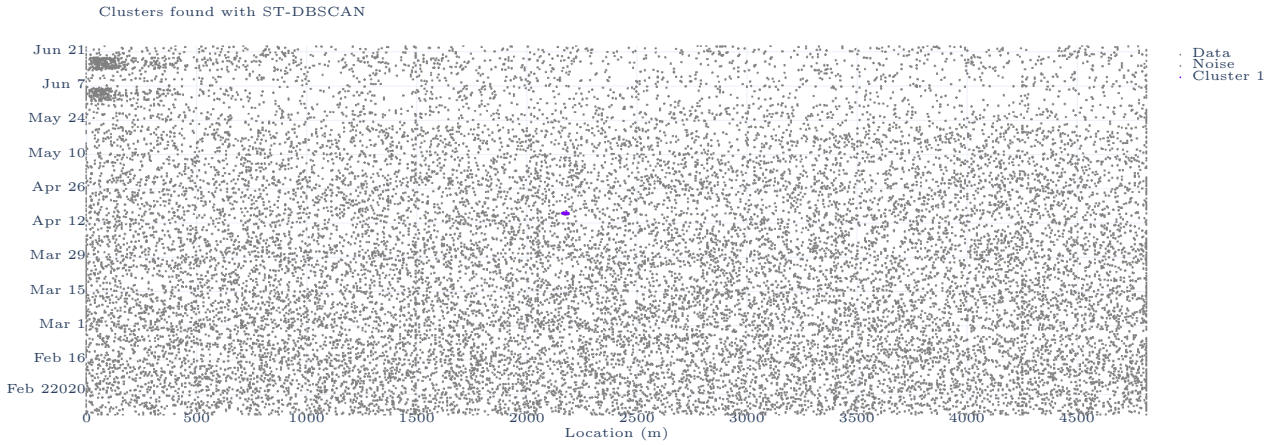

Figure 5.18: Clustering result with parameter set 2.

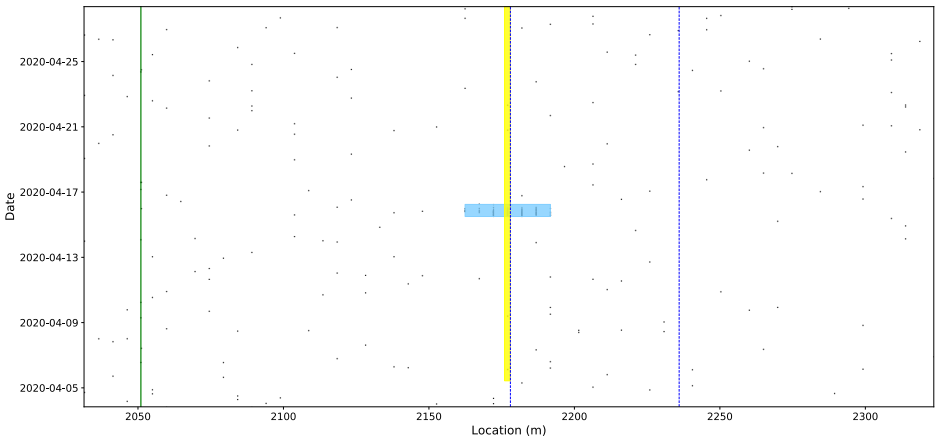

Figure 5.19: Overlap of cluster found at 2177m with the DNV GL warning.

inter-arrival times exceeding 20 minutes was 7. Of course, this is a very specific time threshold of the inter-arrival times and either of the parameter sets were able to identify the DNV GL warnings assigned.

Table 5.6: Performance Indicators

Parameter set (ϵ_1=10m, ϵ_2=250, *minPts*=100)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
2	1	1	0	0.5	1	0.67
Parameter set (ϵ_1=10m, ϵ_2=24, *minPts*=100)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
1	1	0	0	1	1	1

5.4 Case IV

This circuit consists of a PILC type cable circuit connected via 10 joints and 5 RMUs with a total circuit length of 1937m. The PD data for the circuit is represented in Figure 5.20. In the Figure 5.20, it is visible that no data is present during the mid of January, 2020 to mid February, 2020. This is due to new SCG systems installed during that period. There are two warnings assigned to this circuit - a level 3 and a level 1 warning. Level 3 is shown by yellow band whereas level 1 warning is shown using a red band (Figure 5.20).

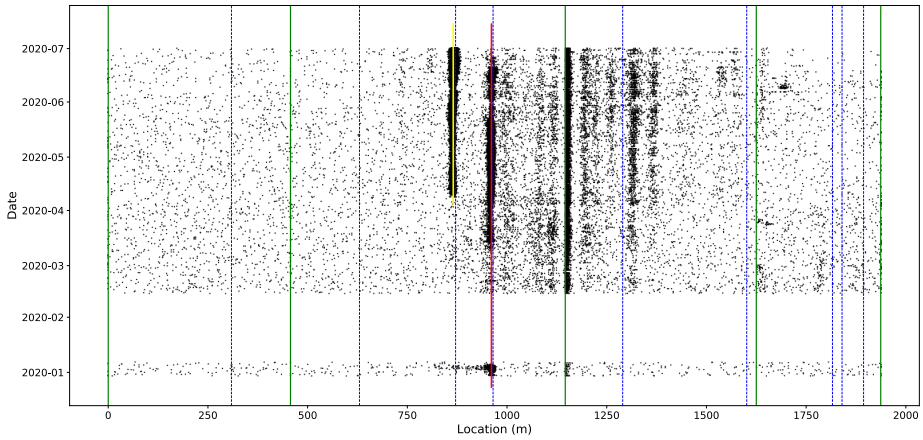

Figure 5.20: Location vs Date/Time plotted using PD data (also showing warnings assigned for the circuit).

The warning information is presented in the following table:-

Table 5.7: Warnings associated with the circuit

Location (in metres)	Start Date/Time	End Date/Time	SCG warning level (1, 2, 3, N)
865	17-04-2020 22:59:55	30-06-2020 23:59:00	3
961	06-01-2020 12:39:05	30-06-2020 23:59:00	1

The warnings assigned to the locations 865m and 961m are close to joint locations in the circuit. There is also a dense activity present around the warning locations. There is a dense activity of discharges present exactly on the location of RMU at 1146m which was not assigned any warning from DNV GL.

The parameter values used in the previous cases are used for the clustering method. The results of the clustering are presented below:

Parameters: ϵ_1 = 10m, ϵ_2 = 250 hours, and *minPts* = 100

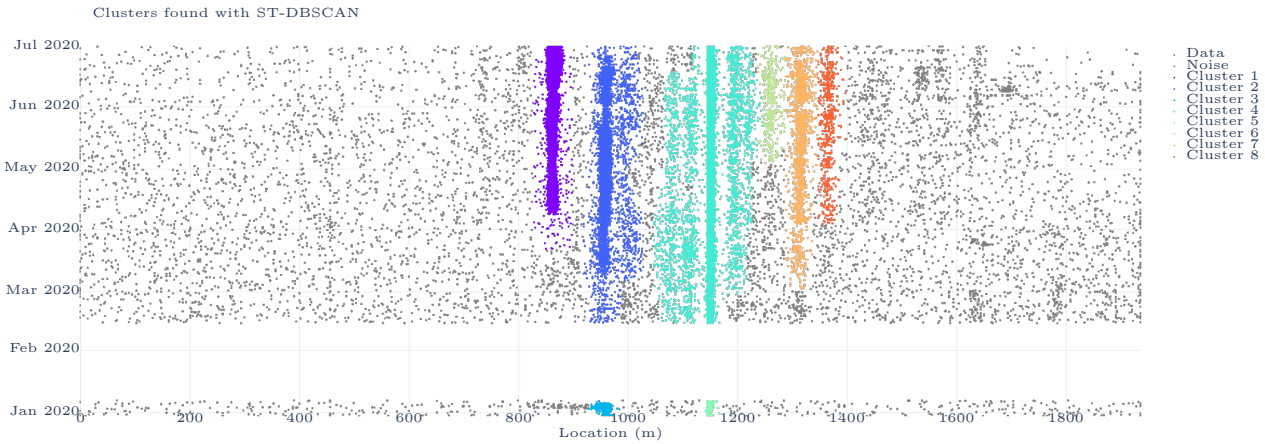

Figure 5.21: Clustering result with parameter set 1. *Clusters labelled from left to right (bottom cluster first).*

Parameters: ϵ_1 = 10m, ϵ_2 = 24 hours, and *minPts* = 100

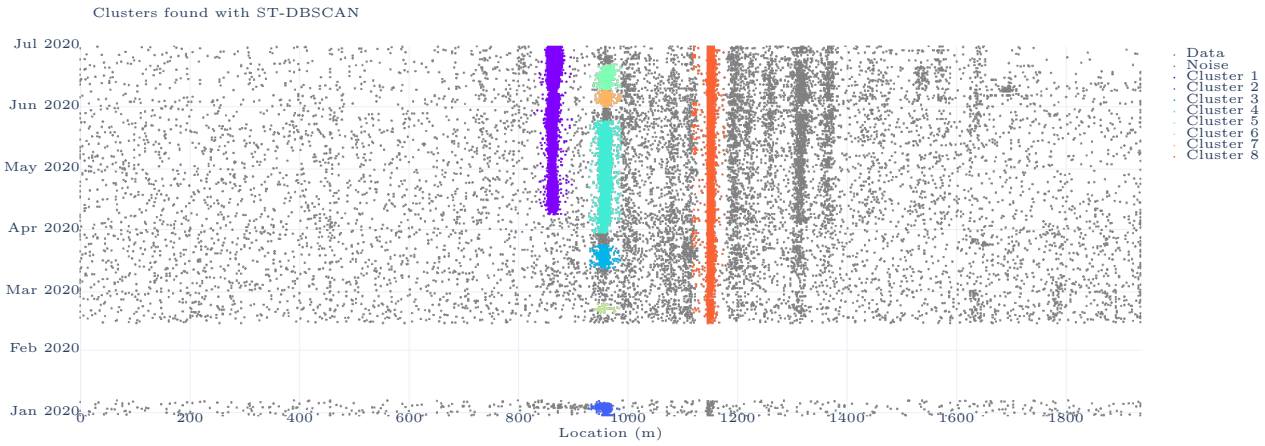

Figure 5.22: Clustering result with parameter set 2. *Clusters labelled from left to right (bottom cluster first).*

Parameters: ϵ_1 = 10m, ϵ_2 = 250 hours, and *minPts* = 500

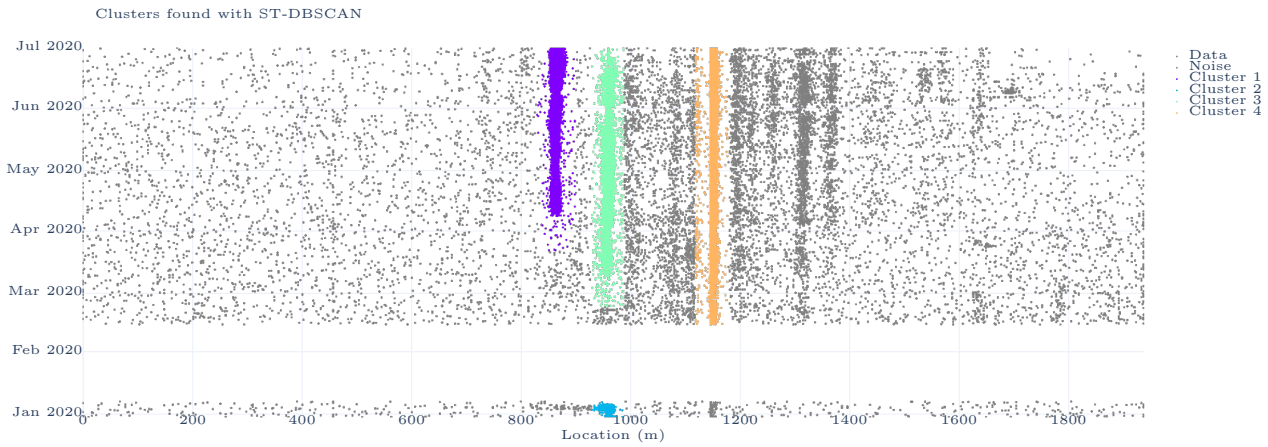

Figure 5.23: Clustering result with parameter set 3. *Clusters labelled from left to right (bottom cluster first).*

Table 5.8: Performance Indicators

Parameter set (ϵ_1=10m, ϵ_2=250, *minPts*=100)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
8	3	5	0	0.375	1	0.54
Parameter set (ϵ_1=10m, ϵ_2=24, *minPts*=100)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
8	7	1	0	0.87	1	0.93
Parameter set (ϵ_1=10m, ϵ_2=250, *minPts*=500)						
$N_{clusters}$	*True Positives*	*False Positives*	*False Negatives*	*Precision*	*Recall*	*F1-score*
4	3	1	0	0.75	1	0.85

From the above parameter set of values, the clustering results show that visibly all parameter sets are able to capture the warning assigned to the circuit. Parameter set 2 yields multiple clusters at location 950-980m which include the DNV GL warnings. This puts a dilemma of sorts of what can be considered as good or effective clustering. Based on the performance indicators, the precision and F1 score suggest that parameter set 1 performs poorly due to the high number of false positives whereas set 2 and 3 perform good. In set 2 the number of clusters is same as the set 1 but the since set 2 is identifies clusters on the same location 5 times the precision and F1 score are not affected. In set 3, the *minPts* was increased to remove the excess clusters identified in set 2. The choice of increasing the *minPts* was made because

varying the temporal threshold resulted in not much change in the multiple clusters found on a single location and hence the idea was to see the effect of increasing the *minPts* on the result of the clustering which would impose stricter constraints on clusters with low density and due to the longer temporal threshold the result of the clustering showed a one big cluster on the locations. This results in the loss of clusters around the dense clusters.

Discussion

This case demonstrates inadequacy of the method of evaluation of the clustering using the precision, recall and F1 scores. The results of the three parameter set clustering and their respective performance scores show that the solely relying on the performance indicators may not be such a good idea to accept or reject the results of the clustering. As it is evident from the 'visual' results that every clustering identified the warnings assigned from DNV GL. For set 1 the number of clusters was high leading to a higher number of false positives resulting in a low precision and F1-score. The set showed similar number of clusters with lower false positives and an improved precison and F1 score whereas for set 3 the overall clusters were low with a low number of false positives.

In this case, we encountered a level 1 warning which as seen from the Figure 5.20 is present for a long time prompting the question regarding the action taken for such a warning. Upon discussions it was made clear that the underlying resin joint for which the warning was signalled was replaced by a heat shrink joint on the 7th January, 2020. The reason for the persistence of the activity as well as the warning was later learned as the cause of the presence of cavities which are present in the joint during the connection of the cable. The material of the joint does not shrink well to the cables when using heaters. These cavities later disappear when adjusted to the cable temperatures and is a known phenomenon in heat shrink joints and hence no cause for alarm. If the PD persists for a long time then it is replaced because of improper installation of the joint.

5.5 Case V

In this case, the circuit consists of purely PILC type insulation cables and with 6 joints and 3 RMUs in the circuit. The length of the circuit is 973m. There were two level 3 warnings assigned to the circuit and are shown in Figure 5.24.

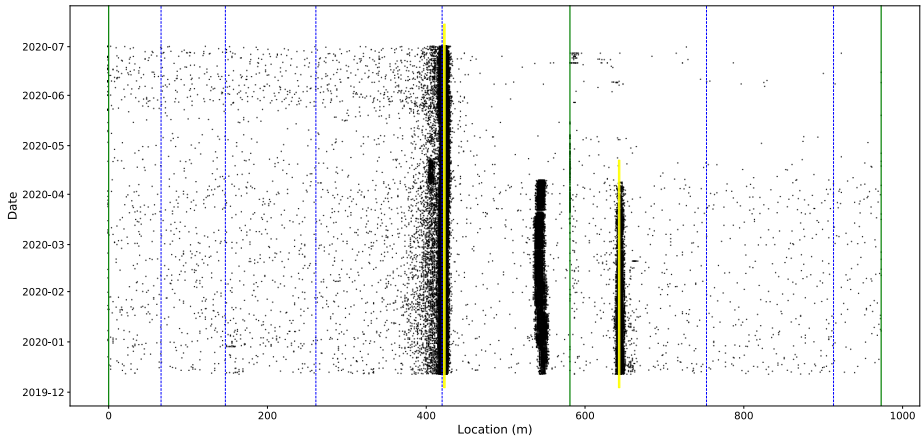

Figure 5.24: Location vs Date/Time PD data for the circuit along with the warnings assigned to it

The clustering was performed with the parameter values set as $\epsilon_1 = 10$m, $\epsilon_2 = 250$ hours, and $minPts = 500$. The results of the clustering is shown in the plot below with the warnings overlaid on the clusters found.

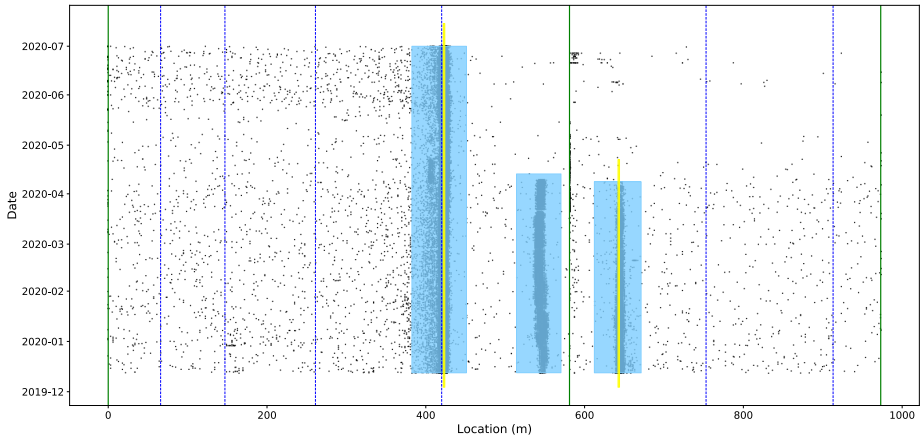

Figure 5.25: Overlap of clusters identified with the warnings assigned to the circuit.

The inter-arrival times of the clusters are plotted below (Figure 5.26). From the plots it is interesting to see that for cluster 1 the discharges were in quick succession. The underlying configuration showed that the cluster was found over a oil-filled joint. The activity shows discharges happening quickly in the oil-filled joint due to continuous migration of the insulation fluid (oil) from the joints into the cable insulation during high loading cycles or increased temperature [15]. Although there are high number of discharges, the oil level in the joint will return back to normal after cooling of the fluid and hence the level 3 warning. The inter-arrival times for cluster 2 and cluster 3 were plotted as well (Figure 5.26). The clusters were found on the cables and were also close to a RMU in the circuit - this based on the information received during discussions seems not something which can be assigned a warning. Although, for cluster 3 there is a level 3 warning assigned. The inter-arrival time of discharges observed in cluster 2 are higher in density than the inter-arrival time of discharges in cluster 3, but was not flagged as a warning by DNV GL experts. This prompts us to investigate into the behaviour of the discharges observed in the identified clusters 2 and 3.

The inter-arrival time distribution of the clusters showed to follow a trend in the time between discharges and seemed close to a lognormal distribution. The lognormal distribution was fit on the distribution data of the inter-arrival times of cluster 2 and 3 (Figure 5.27). For better visualization, the cumulative distribution of the lognormal distribution was fit on the empirical cdf of the inter-arrival time data of the cluster (Figure 5.28).

For the sample that follows a lognormal distribution, the logarithm of the sample data (here the inter-arrival time of discharge events) is known to follow a normal distribution [45]. To assess the goodness of fit of the lognormal distribution, the data is tested for normality or how close the data distribution follows a normal distribution. This was done because when the data distribution was tested using the Kolmogorov-Smirnov (K-S) test which is used to compare a sample data with a known probability distribution, the chi-square statistic and the p-statistic of the test were significantly low. This was because the K-S test needs the location, shape and scale parameter of the data to be specified and an estimated parameter values invalidates the test. Also, the p-statistic of the test is heavily dependent on the amount of sample data and if

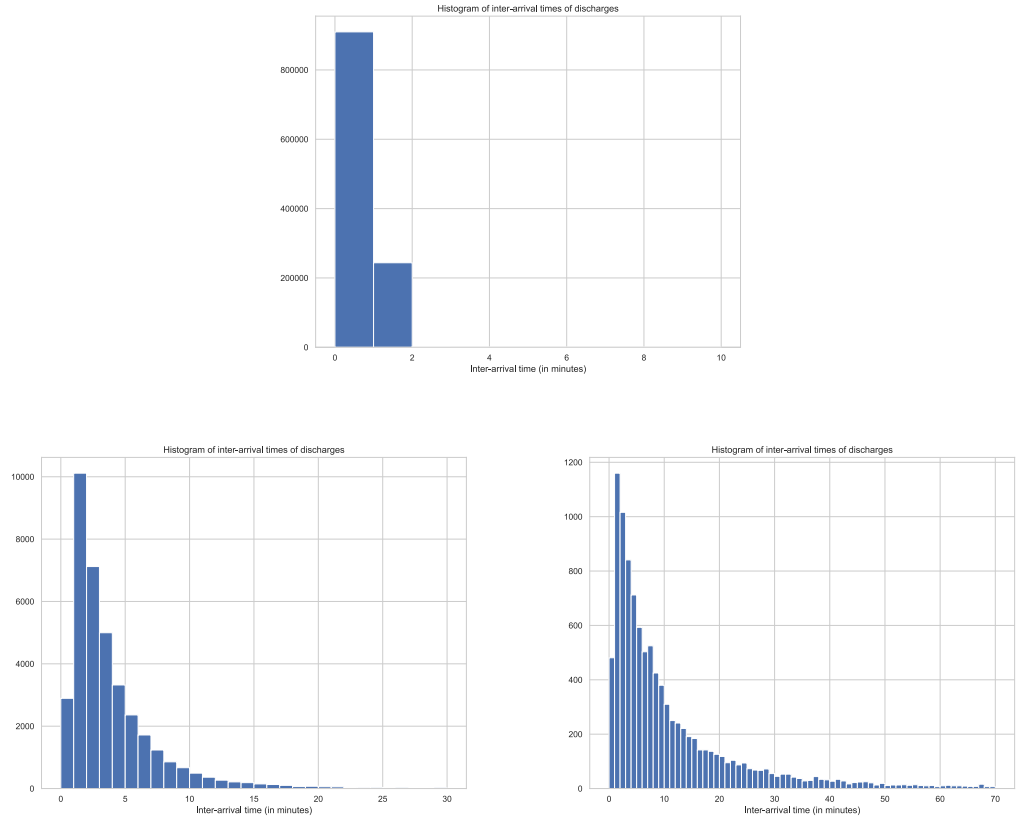

Figure 5.26: Histogram of inter-arrival time of discharges (*Top:* Cluster 1, *Bottom left:* Cluster 2, *Bottom right:* Cluster 3)

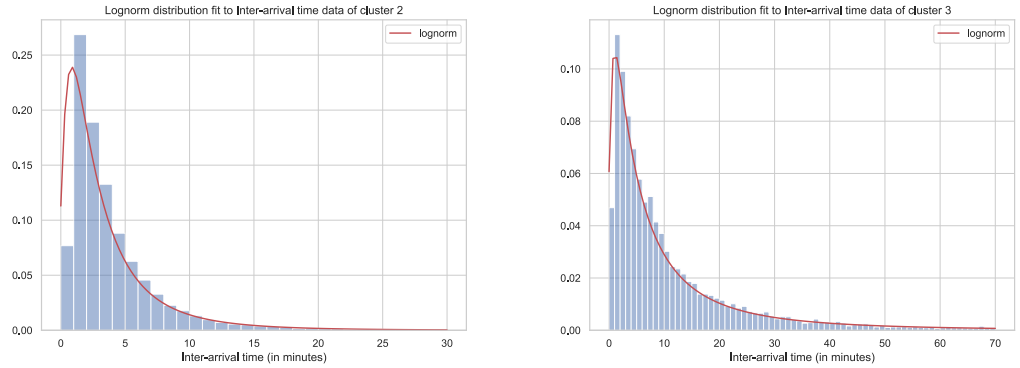

Figure 5.27: Lognormal distribution fit on the inter-arrival times distribution of cluster 2 and 3 respectively

the sample data is large then the number of deviations (which are checked by the p-statistic) can be large and can often lead to rejection of the hypothesis.

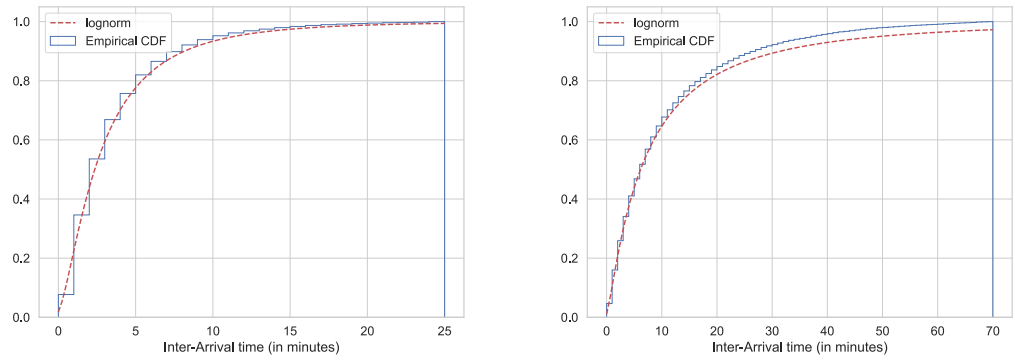

Figure 5.28: Lognormal cumulative distribution function (cdf) fitted for the empirical CDF of the inter-arrival time for Cluster 2 (*left*) and Cluster 3 (*right*).

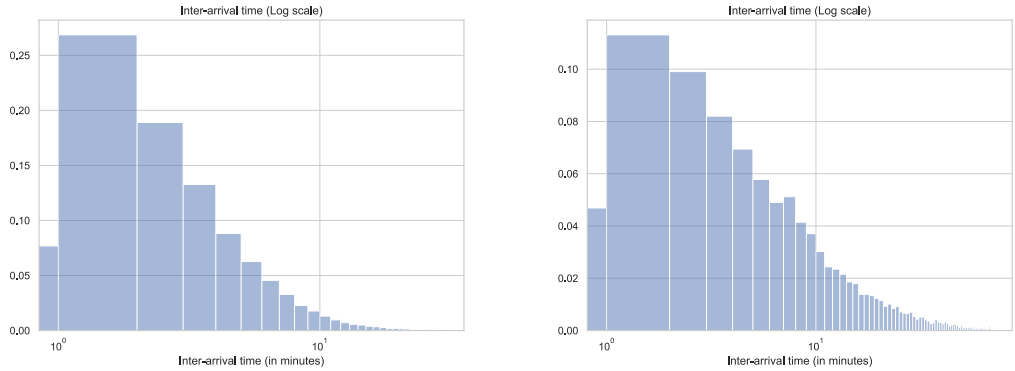

Figure 5.29: Inter-arrival time distribution of Cluster 2 (*left*) and Cluster 3 (*right*) represented on logarithmic scale

Since, the lognormal distribution is known to follow a normal distribution on a log scale the distribution of the data was plotted on a logarithmic scale on the x-axis. The data did not seem to follow a normal distribution and showed a skewed distribution (Figure 5.29). The data was then transformed to follow a normal distribution by using the 'Yeo-Johnson' power transformation [46]. A power transformation helps in converting data using a continuously varying function with respect to the power parameter λ. The Yeo-Johnson power transformation helps to convert non-normal data into normal data by raising the power of the distribution to a power of lamda (λ). The reason behind choosing the Yeo-Johnson power transform is that since the inter-arrival time contains values which are zero minutes, this makes the common Box-Cox transform [47] ineffective as it requires values to be positive and greater than zero. The parameter λ is estimated from the data using the *scipy* package [48] from the equation: [46] (where y_i is a data vector)

$$y_i^{(\lambda)} = \begin{cases} \left((y_i+1)^{\lambda}-1\right)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \log\left(y_i+1\right) & \text{if } \lambda = 0, y \geq 0 \\ -\left[(-y_i+1)^{(2-\lambda)}-1\right]/(2-\lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log\left(-y_i+1\right) & \text{if } \lambda = 2, y < 0 \end{cases} \tag{5.1}$$

After the data is transformed, a normal distribution is fit to the transformed data (Figure 5.30). The transformed data is plotted on a Q-Q plot to compare the probability distributions of the data and the normal distribution (Figure 5.31).

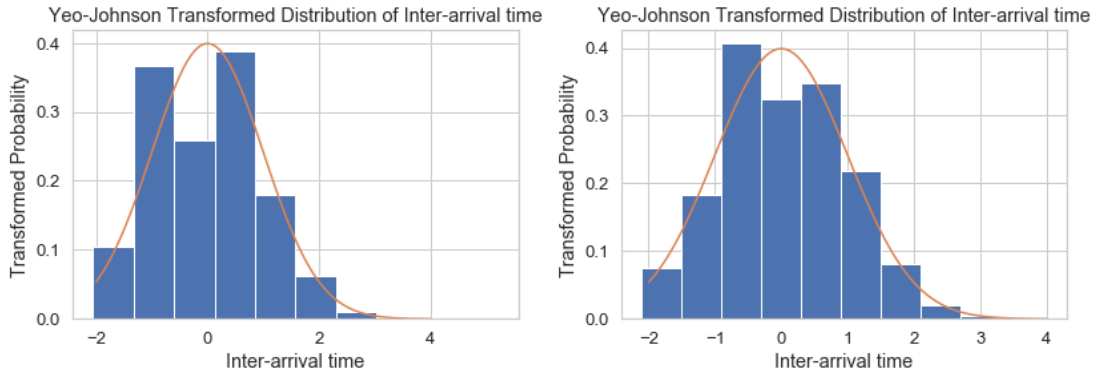

Figure 5.30: Yeo-Johnson Transformed data of the inter-arrival times for Cluster 2 (*left*) and Cluster 3 (*right*) fit with a normal probability distribution

The chi-square and p-statistic was computed for both the transformed data of cluster 2 and 3. For cluster 2, the chi-square statistic was 2.33 and the p-statistic is 0.31 which fails to reject the null hypothesis and states that the distribution of the sample data follows a normal distribution. Although, looking at the distribution fit (Figure 5.30) and the Q-Q plots (Figure 5.31) of cluster 2, the visual depiction contradicts the p-statistic and the data seems to deviate from the normal distribution and shows a positively skewed distribution. For cluster 3, the chi-square statistic was 64.87 but the p-statistic was 0 and rejects the null hypothesis but the Q-Q plots visually depict that the data is distributed much close to a normal distribution. Since the sample data is quite large the Q-Q plot provide a better approximation of the normality of the data than calculating the p-statistic.

The goodness of fit test was done to understand whether the inter-arrival times of certain clusters follows a trend. This case serves as an exploratory analysis of the inter-arrival times and should not be considered as only the lognormal distribution is a possible distribution. Several tests need to be formulated to understand the behaviour of such clusters identified and how many of them are found with an underlying joints or type of cables.

Discussion

This case does not demonstrate the results of clustering method, instead, it analyses the inter-arrival time feature of the identified clusters. It is important to note that this analysis was carried out for only this case and not for the previous cases. The motivation behind the analysis was compelled by a trend in the inter-arrival time distribution of the clusters identified. The lognormal distribution was chosen because the trend in the clusters visually seemed to follow a distribution similar to lognormal. To support this visual depiction of the trend, the goodness of fit test were carried out.

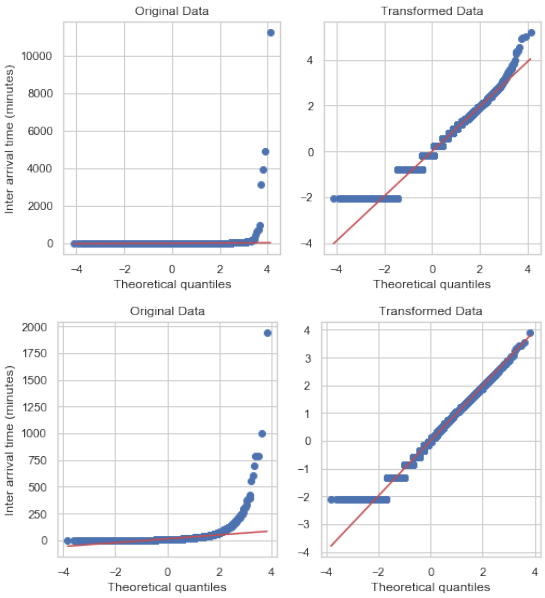

Figure 5.31: *Top:* Q-Q plot of the original and transformed data for cluster 2, *Bottom:* Q-Q plot of the original and transformed data for cluster 3

5.6 Conclusions

This chapter provided results from the cases for the evaluation of the ST-DBSCAN clustering method and also analysed features of the identified clusters.

Cases I and II demonstrate the parameter tuning method for estimating the value for the temporal threshold. In Case II we also see the difference in the results of the clustering between the parameter values reused from the previous case and the tuned parameters.

Case III demonstrates the result of the clustering using the previously identified parameter values and showcases the difference in the result for the two sets of parameter values. Here, we also see that although both sets of values identify clusters where warnings were present it shows that having a higher temporal threshold leads to inclusion of discharge events farther from the actual high-density activity.

Case IV demonstrates the dilemma introduced due to different results of clustering for different sets of parameter values and also showcases the inadequacy of the performance metrics formulated to evaluate the results of the clustering method.

Case V provides perspective into the analysis of the inter-arrival time characteristic of the identified cluster and demonstrates the tests conducted to support the hypothesis of the analysis.

In the next chapter we conclude the discussion and this thesis work and provide possible recommendations for future work.

Chapter 6

Conclusions

This thesis proposed a clustering methodology developed using the ST-DBSCAN density-based clustering technique for identifying high-density discharge events in the partial discharge data for the medium voltage cable circuits obtained from DNV GL's Smart Cable Guard (SCG) systems. This method serves as a contribution to the clustering phase involved in the development of the Automated Warning System at Alliander. The method allows a promising approach in utilizing the spatial and temporal characteristics of the PD data and helps in automatically separating high-density regions from low density noisy areas.

To evaluate the performance of the clustering, indicators are formulated to quantitatively measure the results of the clustering method. This is done by comparing the overlap of manually labelled DNV GL warnings, considered as the ground truth for assessment of the discharge events observed in the PD data for the circuit, with the clusters identified from the clustering method. These indicators are used to calculate the precision, recall and F1-score metrics for the method.

The cases presented in the previous chapter demonstrate the results of the proposed clustering methodology. The method and tuning approach for the temporal threshold are explained and tested for the cases presented in the previous chapter to observe the clustering result. There are real world considerations that should be taken into account, like, (i) variation of density of discharge events among various circuits, (ii) there is also a variation in the timing of occurrence of discharge events among various circuits. With each case we see that the temporal threshold estimated in the first case does not always identify thin clusters and for some cases results in inclusion of discharge events spread out further away in time of occurrence or identification of clusters that do not have an overlap with any DNV GL warnings. During the execution of the cases the mismatch between the assignment of the warning and the actual start of the activity was discovered. This was fixed upon discussion with the DNV GL experts to extend the time margin of the warning timestamp by a period of 2 weeks.

For the clusters identified from the clustering method various features were calculated to quantitatively describe them using statistical moments such as mean, standard deviation, skew, and kurtosis. The PD attributes are utilized to calculate the cluster width using the location attribute, duration of the cluster using the time attribute, and the discharge magnitude observed in the cluster. The cluster densities are calculated to evaluate how many discharges per day were observed within the cluster. Since the data from the SCG is representation of discharge events spread across the length of the circuit and in time of observation, the clusters are further described using the inter-arrival time of occurrence between two consecutive discharge events to identify how close the discharge events were observed in the clusters. This feature also led to an interesting observation demonstrated in case 5 which showed a trend followed by discharges in clusters identified.

A caveat which concerns the validity of the performance results is demonstrated in Case 4 where the dilemma of accepting a clustering to be effective is presented by showcasing the inadequacy of the performance indicators formulated and the visual results of the clustering. This can be challenging in many circuits and can require a visual inspection of the identified clusters to accept or reject the result of the clustering. A possible solution is to identify as many clusters during the clustering phase and then 'post-process' the identified clusters to see if two or more clusters that are found on the same location have a small time gap within them. These clusters can then be merged to form one cluster. Of course this would mean that there be an additional parameter required to be set for minimum time gap between clusters found on the same location. This can be tested for a list of circuits using a tighter threshold (such as the parameters estimated in case 2) and the performance of the results can be evaluated using the metrics.

A challenge for the implementation of the method was encountered. The clustering methodology implemented is not scalable for large datasets as the distance computation of the temporal and spatial neighbors fails when executed on a standard laptop processor. This poses as a challenge where even with the resolution of the data lowered to hour or 12 hour precision and using the location discretization, the distance computation becomes expensive when the objects of the dataset increase leading to an increase in higher number of temporal and spatial neighbors and thereby terminating the computation due to lack of memory resource.

Despite the caveats and challenge, the work carried out during this thesis helps in utilizing the SCG data obtained for the actively monitored circuits and provides a way to describe the discharge events that are identified through the clustering method. This work serves as a preliminary exploration of the possible advantages of automating the identification of the partial discharge events occurring in the circuit and their description which would help in future by aiding in the characterization of these events of interests into actual PD activity or noise.

A few suggestions on the possibilities for future work to broaden the applicability of the current work and to cater more to some of the assumptions made in the clustering methodology and cluster evaluation are presented below:-

- To overcome the issues with the application of the clustering method for large datasets a possible recommendation is to implement a batchwise clustering method which would split the datasets into smaller subsets for distance computation and clustering and set an overlap to merge the results of the clustering back to its original dataset [49].

- The parameter tuning employed using the heuristic in case 1 and 2 was performed by visual inspection of the 'knee-point' in the sorted distances plot. The process of identifying the 'knee-point' or the optimal threshold can be automated to avoid manual errors and can be efficient in setting of the values if the tuning is performed for a large group of representative circuits.

- In this thesis, the cases presented involved circuits with either a mixture cable sections of PILC and XLPE type or were entirely of PILC cables sections. This was also because PILC type cables are more common in the circuit as seen in Figure 3.1 and purely XLPE circuits form less than 20 % of the monitored circuits. Since the newly installed cables are more commonly of the XLPE type, the clustering can be tested for purely XLPE or high percentage of XLPE cable sections in the circuit. This would also provide insight into the partial discharge activity observed in these circuits and the extent of temporal threshold to be set for different types of circuit configurations.

- The inter-arrival time feature explored in case 5 can be extended to test for such trends across circuits and test for the distribution fit. A set of distributions needs to be identified and evaluated across the observed trends to perform a hypothesis test for the explanation of such phenomenon.

Currently, the development of the automated detection of the 'interesting' discharge events and their characterization is being done in collaboration with the DNV GL experts and Alliander. In future a possible addition to the development can be a collaborative effort with other network operators and their data from the SCG systems and create a knowledge sharing platform for evaluating the patterns observed in the medium voltage network. Additionally, different sources of data such as the cable loading capacity, weather data, soil temperature data etc. can be included as features for the evaluation of the discharge events detected in the circuits and investigate their correlation.

Bibliography

[1] Peter Fraser. Netherlands - regulatory reform in the electricity industry. http://www.oecd.org/regreform/sectors/2497385.pdf, 1998. Accessed: 2020-07-25. 1

[2] The Netherlands Authority for Consumers and Markets (ACM). Incentive regulation of the gas and electricity networks in the netherlands. https://www.acm.nl/en/publications/publication/17231/Incentive-regulation-of-the-gas-and-electricity-networks, 2017. Accessed: 2020-07-25. 1

[3] Directive 2009/72/ec of the european parliament and of the council. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009L0072, 2009. Accessed: 2020-07-25. 1

[4] Swasti R. Khuntia, José Luis Rueda, Sonja Bouwman, and Mart A. M. M. van der Meijden. A literature survey on asset management in electrical power [transmission and distribution] system. *International Transactions on Electrical Energy Systems*, 26(10):2123–2133, 2016. 1

[5] Richard E. Brown. *Business Essentials for Utility Engineers*. CRC Press, Boca Raton, 2010. 1

[6] J. J. Smit, B. Quak, and E. Gulski. Integral decision support for asset management of electrical infrastructures. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2622–2628, 2006. 2

[7] L. Bertling. *Reliability centred maintenance for electric power distribution systems*. PhD thesis, Dept. Elect. Power Engineering, KTH, Stockholm, Sweden, 2002. 2

[8] Saxena A. Mann, L. and G.M. Knapp. Statistical-based or condition-based preventive maintenance? *Journal of Quality in Maintenance Engineering*, 1(1):46–59, 1995. 2

[9] Swasti R. Khuntia, Jose L. Rueda, and Mart A. M. M. van der Meijden. Smart asset management for electric utilities: Big data and future. In Joseph Mathew, C.W. Lim, Lin Ma, Don Sands, Michael E. Cholette, and Pietro Borghesani, editors, *Asset Intelligence through Integration and Interoperability and Contemporary Vibration Engineering Technologies*, pages 311–322, Cham, 2019. Springer International Publishing. 3

[10] Netbeheernederland. Betrouwbaarheid van elektriciteitsnetten in nederland. https://www.netbeheernederland.nl/_upload/Files/Betrouwbaarheid_van_elektriciteitsnetten_2016_76.pdf, 2017. Accessed: 2020-06-07. 3, 4, 9, 10

[11] Alliander N.V. Annual report 2019 - working together on transition. https://2019.jaarverslag.alliander.com/FbContent.ashx/pub_1037/downloads/v200320083558/Alliander_Annual_Report_2019.pdf, 2020. Accessed: 2020-08-10. 5

[12] C. Zhou, M. Michel, D. M. Hepburn, and X. Song. On-line partial discharge monitoring in medium voltage underground cables. *IET Science, Measurement Technology*, 3(5):354–363, 2009. 5

[13] P. Wagenaars. *Integration of online partial discharge monitoring and defect location in medium-voltage cable networks*. PhD thesis, Department of Electrical Engineering, 2010. 9, 10, 13

[14] Fred Steennis, P. Wagenaars, and D. Harmsen. Smart cable guard – a tool for on-line monitoring a nd location of pd ' s and faults in mv cables – its application and business case. 2015. 9, 14

[15] F. Wester. *Condition Assessment of Power Cables using Partial Discharge Diagnosis at Damped AC Voltages*. PhD thesis, 2004. 9, 11, 56

[16] S. Mousavi Gargari. *Pattern recognition and knowledge extraction for on-line partial discharge monitoring with defect location*. PhD thesis, Department of Electrical Engineering, 2012. 10, 11, 12

[17] P.C.J.M. Wielen, van der. *On-line detection and location of partial discharges in medium-voltage power cables*. PhD thesis, Department of Electrical Engineering, 2005. 11, 13

[18] P.H.F. Morshuis. *Partial discharge mechanisms: Mechanisms leading to breakdown, analyzed by fast electrical and optical measurements*. PhD thesis, Department of Electrical Engineering, Mathematics and Computer Science, 1993. 11

[19] E. F. Steennis and F. H. Kreuger. Water treeing in polyethylene cables. *IEEE Transactions on Electrical Insulation*, 25(5):989–1028, 1990. 12

[20] IEC 60270 2000. *High-Voltage Test Techniques: Partial Discharge Measurement*. International Electrotechnical Commission (IEC), 2000. 12

[21] G. C. Montanari. On line partial discharge diagnosis of power cables. In *2009 IEEE Electrical Insulation Conference*, pages 210–215, 2009. 12

[22] N. Schaik, van, E.F. Steennis, W. Boone, D.M. Aatrijk, van, and E. Hetzel. Medium voltage cable diagnostics condition based maintenance on power cables. In *Proc. Nordic Insulation Symposium (NORD-IS), Stockholm, June 11-13, 2001*, pages 3–10, 2001. conference; Nordic Insulation Symposium; 2001-06-11; 2001-06-13 ; Conference date: 11-06-2001 Through 13-06-2001. 12

[23] N. van Schaik, E. F. Steennis, A. van Dam, B. J. Grotenhuis, M. J. van Riet, and C. J. Verhoeven. Condition based maintenance on mv cable circuits as part of asset management; philosophy, diagnostic methods, experiences, results and the future. In *16th International Conference and Exhibition on Electricity Distribution, 2001. Part 1: Contributions. CIRED. (IEE Conf. Publ No. 482)*, volume 1, pages 5 pp. vol.1–, 2001. 12

[24] E. F. Steenis, R. Ross, N. Van Schaik, W. Boone, and D. M. Van Aartrijk. Partial discharge diagnostics of long and branched medium-voltage cables. In *ICSD'01. Proceedings of the 20001 IEEE 7th International Conference on Solid Dielectrics (Cat. No.01CH37117)*, pages 27–30, 2001. 13

[25] P. C. J. M. van der Wielen and E. F. Steennis. Experiences with continuous condition monitoring of in-service mv cable connections. In *2009 IEEE/PES Power Systems Conference and Exposition*, pages 1–8, 2009. 13

[26] J. Veen. *On-line signal analysis of partial discharges in medium-voltage power cables*. PhD thesis, Department of Electrical Engineering, 2005. 13, 15, 19

[27] P. C. J. M. Van Der Wielen, J. Veen, P. A. A. F. Wouters, and E. F. Steennis. On-line partial discharge detection of mv cables with defect localisation (pdol) based on two time synchronised sensors. In *CIRED 2005 - 18th International Conference and Exhibition on Electricity Distribution*, pages 1–5, 2005. 13

[28] DNV GL. Smart cable guard (scg). `https://www.dnvgl.com/power-renewables/services/scg/index.html`, 2008. Accessed: 2020-10-20. 13

[29] E.F. Steennis, Denny Harmsen, Theo Rijn, Jan Mosterd, Leon Bokma, Piet Soepboer, Alfred Arts, Nico van Donk, and Branko Carli. The intriguing behaviour over time of pd's from defects in mv cables and accessories; lessons learned with scg, an on-line monitoring system. 6 2011. 13

[30] Paul Wagenaars Fred Steennis Sungin Cho, Ng Desmond. On-line fault detection and localization in the medium voltage network. 2016. 14

[31] P. C. J. M. van der Wielen and E. F. Steennis. On-line pd monitoring system for mv cable connections with weak spot location. In *2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century*, pages 1–8, 2008. 14, 15

[32] Ruud Wassink Sander Rieken. Architecture for the SCG Automated Warning System. 2020. 23

[33] Charu C. Aggarwal. *Data Mining*. Springer, New York, NY, 2015. 26

[34] Vipin Kumar Pang-Ning Tan, Michael Steinbach. *Introduction to Data Mining*. Addison Wesley, New York, NY, 2006. 26

[35] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 01 1973. 27

[36] D. Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 01 1977. 27

[37] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996. 28, 31, 33, 37

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 29, 34

[39] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, 42(3), July 2017. 30, 33

[40] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data Knowledge Engineering*, 60(1):208 – 221, 2007. Intelligent Data Mining. 31, 33

[41] Slava Kisilevich, Florian Mansmann, Mirco Nanni, and Salvatore Rinzivillo. *Spatio-temporal clustering*, pages 855–874. Springer US, Boston, MA, 2010. 31

[42] Hadi Fanaee-T. Spatio-temporal clustering methods classification. 01 2012. 31

[43] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Min. Knowl. Discov.*, 2(2):169–194, June 1998. 33

[44] Eren Cakmak. St-dbscan: Simple and effective method for spatial-temporal clustering. `https://github.com/eren-ck/st_dbscan`, 2020. 34

[45] NIST/SEMATECH . e-handbook of statistical methods. `https://doi.org/10.18434/M32189`, 2012. Accessed: 2020-11-12. 56

[46] In-Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000. 58

[47] G. Box and D. Cox. An analysis of transformations. *Journal of the royal statistical society series b-methodological*, 26:211–243, 1964. 58

[48] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. 58

[49] Iulian Peca, Georg Fuchs, Katerina Vrotsou, Natalia Andrienko, and Gennady Andrienko. Scalable Cluster Analysis of Spatial Events. In Kresimir Matkovic and Giuseppe Santucci, editors, *EuroVA 2012: International Workshop on Visual Analytics*. The Eurographics Association, 2012. 62