MÁSTER UNIVERSITARIO EN INGENIERÍA INDUSTRIAL

TRABAJO FIN DE MÁSTER

Detection of Hotspot Anomalies in LV Networks via Data Analysis

Autor: Carlos Prieto Rodríguez de Vera

Director: Dr. Miguel Ángel Sanz Bobi

Co-Director: Jesús Gutiérrez Serrano

Co-Directora: Itziar Lumbreras Basagoiti

Madrid
Agosto de 2025

Carlos Prieto Rodríguez de Vera, declara bajo su responsabilidad, que el Proyecto con título **Detection of Hotspot Anomalies in LV Networks via Data Analysis** presentado en la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas en el curso académico 2022/23 es de su autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: Carlos Prieto Fecha: 15 / 08 / 2025

Autoriza la entrega:

El Director del Proyecto

Dr. Miguel Ángel Sanz Bobi

Firmado por SANZ BOBI MIGUEL ANGEL - ***6599** el día 26/08/2025 con un certificado emitido por AC FNMT Usuarios

Fdo.: ..... Fecha: ..... / ..... / .....

Summary

Title: Detection of Hotspot Anomalies in LV Networks via Data Analysis

Author: **Prieto Rodríguez de Vera, Carlos.**

Director: Sanz Bobi, Miguel Ángel.

Industrial Supervisor: Gutiérrez Serrano, Jesús

Industrial Supervisor: Lumbreras Basagoiti, Itziar

Collaborating Entity: i-DE Redes Eléctricas Inteligentes, Iberdrola.

Collaborating Entity: ICAI – Universidad Pontificia Comillas.

Collaborating Entity: University of Strathclyde.

Summary of the Project

The increasing complexity of low-voltage (LV) distribution networks [1], driven by the integration of distributed energy resources (DER) [2], electric vehicles, and bidirectional power flows, necessitates a shift. These complex operations [3] alongside the need to increase supply quality and grid resilience require proactive maintenance strategies [4][5]. This thesis addresses the challenge of pre-emptively detecting hotspot anomalies in secondary substations and underground cable pits [6]. The work aims to support Distribution System Operators (DSOs) [7] in enhancing grid reliability, safety, and operational efficiency through data-driven predictive maintenance models [8]. Predictive strategies enhance system reliability by reducing unplanned outages and minimizing unnecessary maintenance operations, ultimately leading to lower OPEX and improved asset longevity [9] [10].

The topic of this work is hotspot anomalies [11]. These anomalies are localized thermal degradations that can lead to equipment failure, service interruptions, and safety hazards [9][12]. This thesis utilizes data from the Supervisión Avanzada de Baja Tensión (SABT) system and Advanced Metering Infrastructure (AMI) [13], including smart meter event logs and the operational management system (OMS) incident reports. These data sources enable the identification and classification of historical incidents, forming the basis for model training and validation.

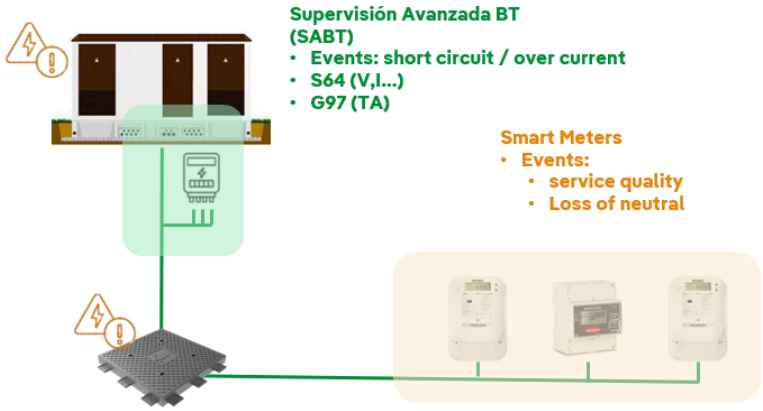
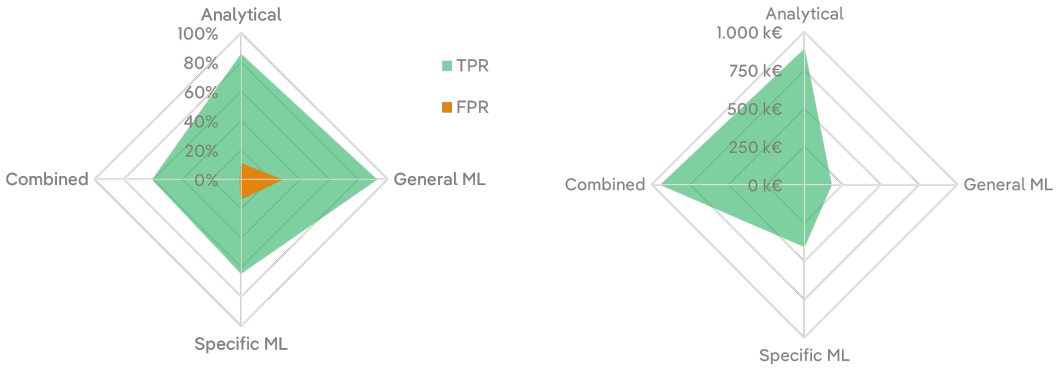


Figure 1: Data sources (SABT and Smart Meters) and visual representation of a secondary substation and an underground cable pit

The thesis proposes a hybrid approach combining analytical and machine learning (ML) models. For secondary substations, three models were developed:

- An analytical threshold model based on physical parameters such as temperature and current to create the statistical thresholds.
- A general-purpose multilayer perceptron (MLP) model trained on aggregated data for all secondary substations that reducing training but loses individualized complex pattern detection.
- A secondary substation-specific ML model tailored to individual asset behaviour.
- An and combination of the 3 above allows for fewer false positives and results in the economic optimum

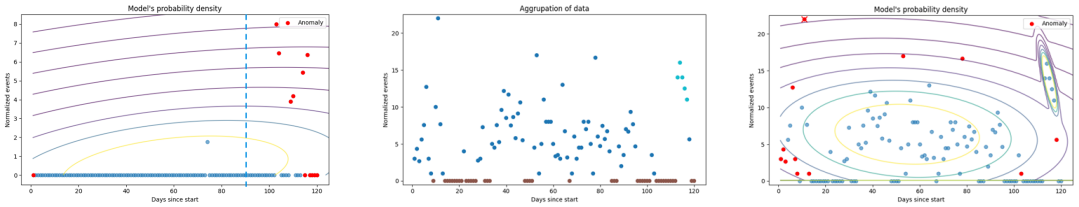


(a) Main KPIs summary

(b) Economic summary

Figure 2: KPI and economic summary for: analytical model, general ML model, specific ML model, combination of models

For underground cable pits, where direct measurements are unavailable, a Gaussian Mixture Model (GMM) was implemented using smart meter event logs. This probabilistic model identifies deviations from normal behaviour patterns to detect potential anomalies.



(a) Anomaly detection before and after the 90 day mark

(b) Data Gaussian aggregation for damaged line

(c) Gaussian probability spread for the same damaged line

Figure 3: GMM model examples, final model detects (post 90 day mark) either anomalous data or the need for a new Gaussian to explain the data and comparing the Gaussians mean number of events before and after the mark

The analytical model for secondary substations achieved the highest True Positive Rate (TPR) to False Positive Rate (FPR) ratio, demonstrating strong interpretability and reliability. The combined deployment of all three models yielded the best economic performance, with a projected annual benefit exceeding €900,000.

Table 1: Performance metrics summary of hotspot anomaly prediction GMM models for underground cable pits

Model Nº	**Model Description**	**TPR**	**FPR**
13	Rate of anomaly after $> 2 \times$ rate before (using p90 probability for 'before')	31%	4%
55	Number of normal instances after $>$ number of normal instances before	73%	88%
64	Mean of normal instances after $> 2 \times$ mean of normal instances before	39%	4%
68	Combination of models 55 and 64	36%	4%
100	Logical OR between models 13 and 68	46%	8%

Table 2: Economic results of the combined model

KPI	**Result**
True Positive Rate	60.71%
False Positive Rate	1.33%
TPR / FPR	45.5
Economic result	944,000 € / yr

In the case of underground cable pits, the GMM model achieved a TPR of 46% with an FPR below 1%, validating the feasibility of predictive maintenance in data-scarce environments. These results underscore the value of leveraging existing SABT and Smart Meter infrastructure for anomaly detection. Although economic break-even was not achieved for underground cable pits, the economic analysis performed did not include social and media costs of allowing hotspot events subsequent thermal incidents or unexpected grid element downtimes.

This thesis demonstrates that predictive maintenance for hotspot anomalies in LV networks is both technically (for both cases) and economically viable (for secondary substations). By integrating analytical and ML models with existing SABT and AMI data, the proposed framework enhances grid observability and supports predictive asset management. The work contributes novel methodologies for anomaly detection, extends predictive maintenance to previously under-monitored assets (such as underground cable pits), and aligns with the strategic goals of DSOs in the context of digitalization and regulatory evolution. Future work should focus on improving data granularity, integrating maintenance logs, and exploring more efficient ML techniques for real-time deployment.

Título: Detection of Hotspot Anomalies in LV Networks via Data Analysis

Autor: **Prieto Rodríguez de Vera, Carlos.**

Director: Sanz Bobi, Miguel Ángel.

Supervisor Industrial: Gutiérrez Serrano, Jesús

Supervisora Industrial: Lumbreras Basagoiti, Itziar

Entidad Colaboradora: i-DE Redes Eléctricas Inteligentes, Iberdrola.

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas.

Entidad Colaboradora: University of Strathclyde.

Resumen Del Proyecto

La creciente complejidad de las redes de distribución de baja tensión (LV) [1], impulsada por la integración de recursos energéticos distribuidos (DER) [2], vehículos eléctricos y flujos de potencia bidireccionales, exige un cambio de enfoque. Estas operaciones complejas [3], junto con la necesidad de mejorar la calidad del suministro y la resiliencia de la red, requieren estrategias de mantenimiento proactivo [4][5]. Esta tesis aborda el reto de detectar de forma anticipada anomalías térmicas (hotspot anomalies) en centros de transformación secundarios y arquetas de cableado subterráneo [6]. El trabajo tiene como objetivo apoyar a los Operadores del Sistema de Distribución (DSOs) [7] en la mejora de la fiabilidad, la seguridad y la eficiencia operativa de la red mediante modelos de mantenimiento predictivo basados en datos [8]. Las estrategias predictivas mejoran la fiabilidad del sistema al reducir las interrupciones no planificadas y minimizar las operaciones de mantenimiento innecesarias, lo que conduce a una disminución de los costes operativos y a una mayor longevidad de los activos [9] [10].

El tema principal de este trabajo son las anomalías térmicas (hotspot anomalies) [11]. Estas anomalías son degradaciones térmicas localizadas que pueden provocar fallos en los equipos, interrupciones del servicio y otros riesgos [9][12]. Esta tesis utiliza datos del sistema de Supervisión Avanzada de Baja Tensión (SABT) y de la Infraestructura de Medición Avanzada (AMI) [13], incluyendo los registros de eventos de los contadores inteligentes y los informes de incidencias del sistema de gestión operativa (OMS). Fuentes de datos que permiten la identificación y clasificación de incidentes históricos, constituyendo la base para el entrenamiento

y validación de los modelos.

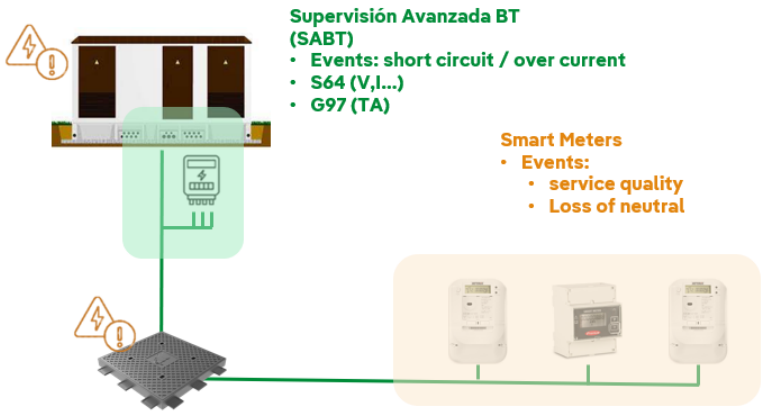
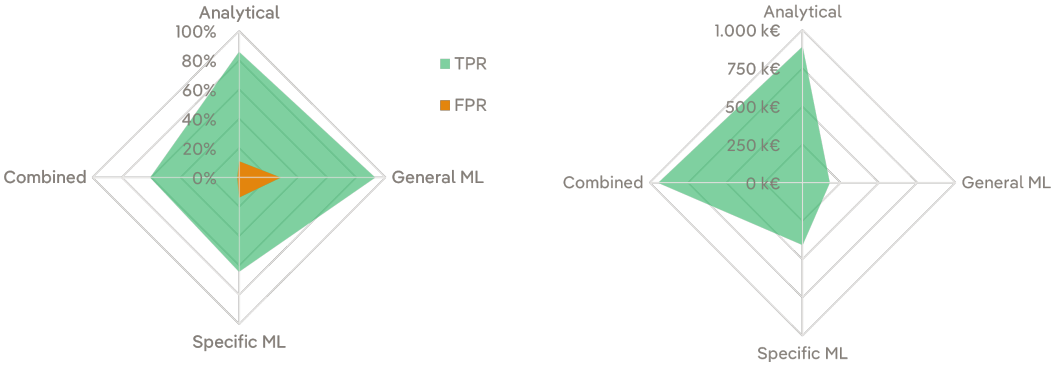


Figura 4: Fuentes de datos (SABT y contadores inteligentes) y representación visual de un centro de transformación y una arqueta de cableado subterráneo

La tesis propone un enfoque híbrido que combina modelos analíticos y de aprendizaje automático (ML). Para los centros de transformación secundarios, se desarrollaron tres modelos:

- Un modelo analítico basado en umbrales físicos como la temperatura y la corriente para establecer criterios estadísticos.
- Un modelo general de perceptrón multicapa (MLP) entrenado con datos agregados de todos los centros de transformación, que reduce el tiempo de entrenamiento pero pierde capacidad de detección de patrones complejos individualizados.
- Un modelo específico por centro de transformación, adaptado al comportamiento individual de cada activo.
- Una combinación lógica tipo “and” de los tres modelos anteriores permite reducir los falsos positivos y resulta en el óptimo económico.

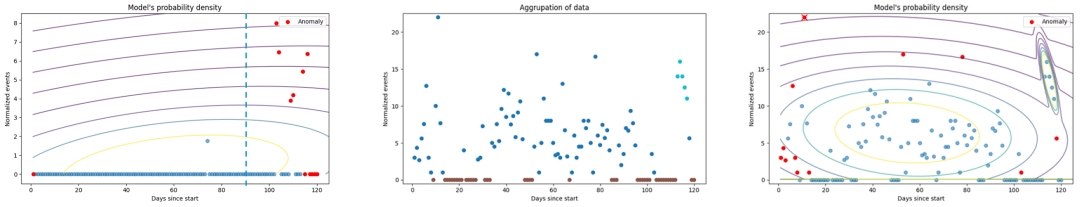


(a) Resumen de los principales KPIs

(b) Resumen económico

Figura 5: Resumen de KPIs y resultados económicos para: modelo analítico, modelo ML general, modelo ML específico y combinación de modelos

Para las arquetas de cableado subterráneo, donde no se dispone de mediciones directas, se implementó un modelo de mezcla gaussiana (GMM) utilizando los registros de eventos de los contadores inteligentes. Este modelo probabilístico identifica desviaciones respecto a los patrones normales de comportamiento para detectar posibles anomalías.



(a) Detección de anomalías antes y después del umbral de 90 días

(b) Agrupación gaussiana de datos para una línea dañada

(c) Distribución de probabilidad gaussiana para la misma línea dañada

Figura 6: Ejemplos del modelo GMM. El modelo final detecta (tras el umbral de 90 días) datos anómalos o la necesidad de una nueva gaussiana para explicar los datos, comparando además el número medio de eventos antes y después del umbral.

El modelo analítico para centros de transformación secundarios alcanzó la mejor relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR), demostrando una alta interpretabilidad y fiabilidad. La implementación combinada de los tres modelos desarrollados ofreció el mejor rendimiento económico, con un beneficio anual proyectado superior a los €900,000.

Cuadro 3: Resumen de métricas de rendimiento de los modelos GMM para detección de anomalías térmicas en arquetas de cableado subterráneo

Modelo Nº	**Descripción del modelo**	**TPR**	**FPR**
13	Tasa de anomalías después >2 × tasa antes (usando probabilidad p90 para el periodo 'antes')	31 %	4 %
55	Número de instancias normales después >número de instancias normales antes	73 %	88 %
64	Media de instancias normales después >2 × media de instancias normales antes	39 %	4 %
68	Combinación de los modelos 55 y 64	36 %	4 %
100	OR lógico entre los modelos 13 y 68	46 %	8 %

Cuadro 4: Resultados económicos del modelo combinado

KPI	**Resultado**
Tasa de verdaderos positivos (TPR)	60.71 %
Tasa de falsos positivos (FPR)	1.33 %
TPR / FPR	45.5
Resultado económico	944,000 € / año

En el caso de las arquetas, el modelo GMM alcanzó una TPR del 46 % con una FPR inferior al 1 %, validando la viabilidad del mantenimiento predictivo en entornos con escasez de datos. Estos resultados refuerzan el valor de aprovechar la infraestructura existente de SABT y contadores inteligentes para la detección de anomalías. Aunque no se alcanzó el punto de equilibrio económico para las arquetas, el análisis realizado no contempla los costes sociales ni reputacionales derivados de permitir el desenlace de puntos calientes en elementos de la red.

Esta tesis demuestra que el mantenimiento predictivo para anomalías de punto caliente en redes de baja tensión es viable tanto técnica (en ambos casos) como económicamente (en centros de transformación). Al integrar modelos analíticos y de ML con los datos existentes del SABT y AMI, el marco propuesto respalda la gestión predictiva de activos. El trabajo aporta metodologías novedosas para la detección de anomalías, extendiendo el mantenimiento predictivo a activos previamente menos monitorizados (como las arquetas) y se alinea con los objetivos estratégicos de los DSOs en el contexto de la digitalización y evolución regulatoria. Las futuras líneas de trabajo deberían centrarse en mejorar la granularidad de los datos, integrar registros de mantenimiento y explorar técnicas de ML más eficientes para su despliegue en tiempo real.

Abstract

Title: Detection of Hotspot Anomalies in LV Networks via Data Analysis

Author: **Prieto Rodríguez de Vera, Carlos.**

Director: Sanz Bobi, Miguel Ángel.

Industrial Supervisor: Gutiérrez Serrano, Jesús

Industrial Supervisor: Lumbreras Basagoiti, Itziar

Collaborating Entity: i-DE Redes Eléctricas Inteligentes, Iberdrola.

Collaborating Entity: ICAI – Universidad Pontificia Comillas.

Collaborating Entity: University of Strathclyde.

The increasing complexity of low-voltage (LV) distribution networks, driven by the proliferation of distributed energy resources (DER), electric vehicles, and bidirectional power flows, demands a paradigm shift in grid operation and maintenance. This thesis addresses the critical challenge of detecting hotspot anomalies—localized thermal degradations that pose significant reliability and safety risks—in secondary substations and underground cable pits. Leveraging the data-rich environment enabled by Advanced Metering Infrastructure (AMI) and the Supervisión Avanzada de Baja Tensión (SABT) system, the work proposes a hybrid predictive maintenance framework combining analytical threshold models and machine learning (ML) techniques.

For secondary substations, three predictive models were developed and validated: a physically interpretable analytical model, a general-purpose multilayer

perceptron (MLP) model, and a substation-specific MLP model. Each was evaluated in terms of predictive accuracy, economic feasibility, and computational efficiency. The analytical model demonstrated the highest TPR/FPR ratio at 7.71, while the combined deployment of all three models yielded the most favourable economic outcome, with a projected annual benefit exceeding 900,000€ and the general ML model had the highest TPR at 92.86%. For underground cable pits—where direct measurements are unavailable—a Gaussian Mixture Model (GMM) was implemented using smart meter event logs. Despite data limitations, the model achieved a 46% true positive rate with a false positive rate below 1%, demonstrating the feasibility of anomaly detection in these challenging environments.

The thesis also presents a comprehensive economic analysis, aligning model performance with operational costs and regulatory incentives. It concludes that predictive maintenance, when supported by robust data analytics, offers a scalable and cost-effective solution for enhancing grid reliability and safety. The work contributes novel methodologies for anomaly detection in LV networks, extends predictive maintenance to previously under-monitored assets, and supports the strategic evolution of Distribution System Operators (DSOs) toward proactive, data-driven asset management.

Título: Detection of Hotspot Anomalies in LV Networks via Data Analysis

Autor: **Prieto Rodríguez de Vera, Carlos.**

Director: Sanz Bobi, Miguel Ángel.

Supervisor Industrial: Gutiérrez Serrano, Jesús

Supervisora Industrial: Lumbreras Basagoiti, Itziar

Entidad Colaboradora: i-DE Redes Eléctricas Inteligentes, Iberdrola.

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas.

Entidad Colaboradora: University of Strathclyde.

La creciente complejidad de las redes de distribución de baja tensión (BT), impulsada por la proliferación de recursos energéticos distribuidos (DER), vehículos eléctricos y flujos de potencia bidireccionales, exige un cambio de paradigma en la operación y el mantenimiento de la red. Esta tesis aborda el desafío crítico de detectar anomalías térmicas —degradaciones localizadas que representan riesgos significativos para la fiabilidad y la seguridad— en centros de transformación secundarios y arquetas de cableado subterráneo. Aprovechando el entorno rico en datos habilitado por la Infraestructura de Medición Avanzada (AMI) y el sistema de Supervisión Avanzada de Baja Tensión (SABT), se propone un marco híbrido de mantenimiento predictivo que combina modelos analíticos basados en umbrales físicos con técnicas de aprendizaje automático (ML).

Para los centros de transformación secundarios, se desarrollaron y validaron tres modelos predictivos: un modelo analítico basado en criterios físicos, un modelo general de perceptrón multicapa (MLP) y un modelo específico por subestación. Cada uno fue evaluado en términos de precisión predictiva, viabilidad económica y eficiencia computacional. El modelo analítico demostró la mejor relación entre tasa de verdaderos positivos (TPR) y tasa de falsos positivos (FPR) con un valor de 7,72, mientras que la combinación de los tres modelos ofreció el resultado económico más favorable, con un beneficio anual proyectado superior a 900.000 € y el modelos de ML general obtuvo la mayor tasa TPR con un 92,86%. Para las arquetas de cableado subterráneo —donde no existen mediciones directas— se implementó un modelo de mezcla gaussiana (GMM) utilizando los registros de

eventos de los contadores inteligentes. A pesar de las limitaciones de datos, el modelo alcanzó una tasa de verdaderos positivos del 46% con una tasa de falsos positivos inferior al 1%, demostrando la viabilidad de la detección de anomalías en estos entornos complejos.

La tesis también presenta un análisis económico exhaustivo, alineando el rendimiento de los modelos con los costes operativos y los incentivos regulatorios. Se concluye que el mantenimiento predictivo, respaldado por análisis de datos robustos, ofrece una solución escalable y rentable para mejorar la fiabilidad y la seguridad de la red. El trabajo aporta metodologías novedosas para la detección de anomalías en redes BT, extiende el mantenimiento predictivo a activos previamente no monitorizados y respalda la evolución estratégica de los Operadores de Sistemas de Distribución (DSOs) hacia una gestión proactiva y basada en datos.

Acknowledgements

I would like to express my sincere gratitude to the individuals and institutions whose support and collaboration have been essential to the completion of this thesis:

- My academic director, Dr. Miguel Ángel Sanz Bobi, for his expert guidance, continuous support, and insightful feedback throughout the development of this work. His strategic vision and technical depth were instrumental in shaping the direction of the project.
- My industrial supervisors, Jesús Gutiérrez Serrano and Itziar Lumbreras Basagoiti, for their invaluable contributions from the field, their practical insights, and their commitment to bridging academic research with industrial application. Their involvement was key to ensuring the relevance and feasibility of the proposed solutions.
- The collaborating entity i-DE Redes Eléctricas Inteligentes (Iberdrola), for providing access to critical data sources, technical infrastructure, and operational expertise that made this research possible. Specifically the team of Transformative Projects, Innovation and Innovative Processes for their technical expertise.
- ICAI – Universidad Pontificia Comillas and the University of Strathclyde, for their academic support, resources, and the opportunity to work within a multidisciplinary and international framework.
- My family, for their unwavering encouragement throughout this journey. Their support has been a constant source of motivation.

I am deeply thankful for the collective effort, trust, and collaboration that have made this project a reality.

Contents

List of Figures

List of Tables

Acronyms

ICAI	Insitituto Católico de Artes e Industrias
SDG	Sustainable Development Goals
DER	Distributed Energy Resources
EVs	Electric Vehicles
V2G	Vehicle to Grid
MVP	Minumim Viable Project
CPU	Central Processing Unit
AMI	Advanced Metering Infrastructure
RTU	Remote Terminal Unit
SABT	Supervisión Avanzada de Baja Tensión
OMS	Outage Management System
GIS	Geographical Information System
CNMC	Comisión Nacional de los Mercados y la Competencia
PNIEC	Plan Nacional Integrado de Energía y Clima
DNO	Distribution Network operator
DSO	Distribution System Operator
TSO	Transmission System operator
ML	Machine Learning
SVM	Support Vector Machine
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
GMM	Gaussian Mixture Model
KDE	Kernel Distribution Estimation

Chapter 1

Introduction

This introductory chapter aims to explain the reasoning behind this thesis and its structure. Starting with the background and motivation, following with the objectives and methodology overview to end up discussing the organisation of the present document.

1.1 Background and motivation

1.1.1 Background

Historically, the electric sector has been divided into four distinct segments: generation, transmission, distribution and commercialization. With each segment operating with distinct and differentiated responsibilities and unidirectional power flows - from generation units connected to the transmission grid, down to consumption points, mainly connected via the distribution network. However, this conventional model is undergoing a transformation phase. Boundaries are no longer as clear as before and the hierarchical organisation of the grid is dissolving [1].

The main reason for this change is the introduction of emerging technology such as intermittent power generation from Distributed Energy Resources (DER) [2], capable of causing reverse power flows which render unidirectional protections obsolete and significantly increase operational complexity [3]. Additionally, the deployment of Battery Energy Storage Systems (BESS) - which are required for an efficient renewable energy integration, the proliferation of electric vehicle (EVs) and their potential grid interactions via schemes such as vehicle to grid (V2G) and demand response mechanisms, all contribute to a more dynamic, less predictable

and subsequently more complex to manage distribution grid. These developments challenge conventional voltage control strategies and complicate the forecasting of load and generation profiles [14] [15].

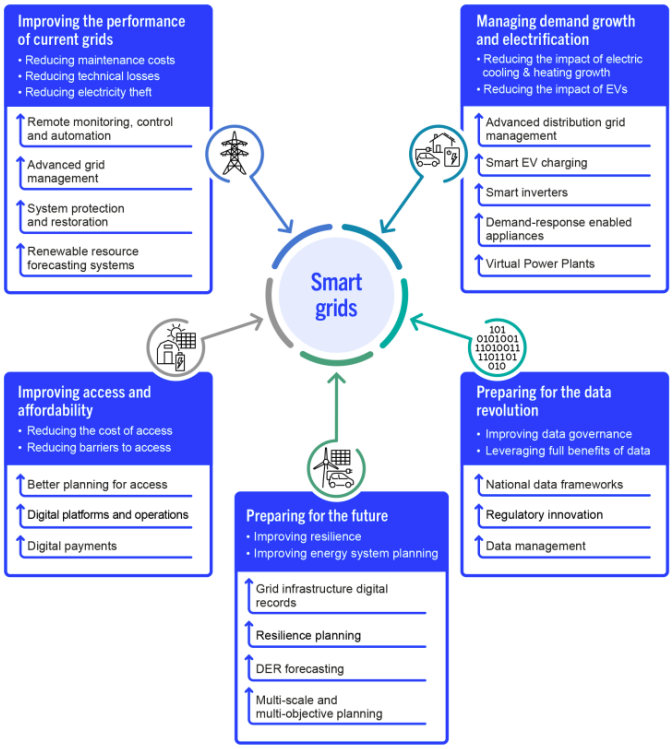


Figure 1.1: How can digital technologies and smart grids positively impact the grid [8]

As a result Distribution Network Operators (DNO) are transitioning from passive infrastructure managers to active Distribution System Operators (DSO) of these new smart grids. As a parallel trend regulators, governments and consumers are demanding a price reduction and an increase in reliability and service continuity. In Spain, this pressure is reflected in the regulatory incentives to reduce key reliability indices such as SAIFI (system average interruption frequency index) and SAIDI (system average interruption duration index) [4]. A way to cope with this added complexity and new demands is to optimise operations using large-scale data analytics to enhance grid reliability alongside a great investment in new more automated and digitalised power networks [16]. A viable course of action thanks to the currently deployed advanced metering infrastructure (AMI).

With a daily average exceeding 600 recorded incidents in large-scale distribution networks, it is crucial to pre-emptively detect anomalies and evolve into

a predictive maintenance to minimize interruptions, ensuring service continuity and minimising non-served energy. Amongst the most time-consuming and resource intensive incidents, that also poses a substantial safety threat, are instantaneous secondary substation thermal degradations, or also called hot spot events [11].

The deployment of SABT (Advanced Low Voltage Supervision, in Spanish "Supervisión Avanzada de Baja Tensión"), represents a key enabler in the transition towards grid digitalisation and predictive maintenance as a key data source for upstream elements. It also encompasses the remote systems and edge computing capabilities. For downstream data collection, at the consumption points, Advanced Metering Infrastructure (AMI) [13] is a key element. Both systems, collectively, enhance the observability of low-voltage networks.

Digitalisation allows for the continuous collection and processing of operational data, enabling detection of early signs of anomalies such as abnormal temperature rises, voltage irregularities, or load imbalances. These indicators, when properly analysed, can reveal the onset of conditions that may lead to critical thermal events or unplanned equipment downtime. By leveraging SABT data, operators can move from reactive to proactive maintenance strategies, improving both reliability and cost-efficiency in increasingly complex grid environments. It is for these advantages that the 'Comisión Nacional de los Mercados y la Competencia' (CNMC) is aligning regulation to push for digitalisation as a means for increased quality of supply [4] [5].

The evolution from conventional distribution systems to smart grids marks a fundamental shift in the design and operation of electrical infrastructure. Smart grids are defined by their capacity to integrate new technologies and enable real-time data communication and analysis. This transformation is driven as a digitalized solution to the complexity of modern grids [17].

The International Smart Grid Action Network (ISGAN) highlights that smart grids are not merely technological upgrades but foundational infrastructures for achieving climate and energy transition goals. Their capacity to integrate distributed energy resources, support bidirectional flows, and enable adaptive system management is essential for meeting decarbonisation targets and ensuring long-term grid sustainability [18].

In Spain, this transition is supported by regulatory and strategic frameworks such as the 'Plan Nacional Integrado de Energía y Clima' (PNIEC) [19] and the EU Clean Energy Package [20], which promote digitalisation, flexibility, and resilience in distribution networks. These policies encourage Distribution System Operators (DSOs) to adopt advanced monitoring, automation and data analytics to manage

increasingly dynamic and decentralised systems.

Predictive maintenance, underpinned by machine learning and statistical modelling, is a cornerstone of this new operational paradigm. It allows for the early identification of anomalies, like thermal stress or abnormal load patterns, before they turn into service-affecting incidents, improving reliability and reducing OPEX and CAPEX [10].

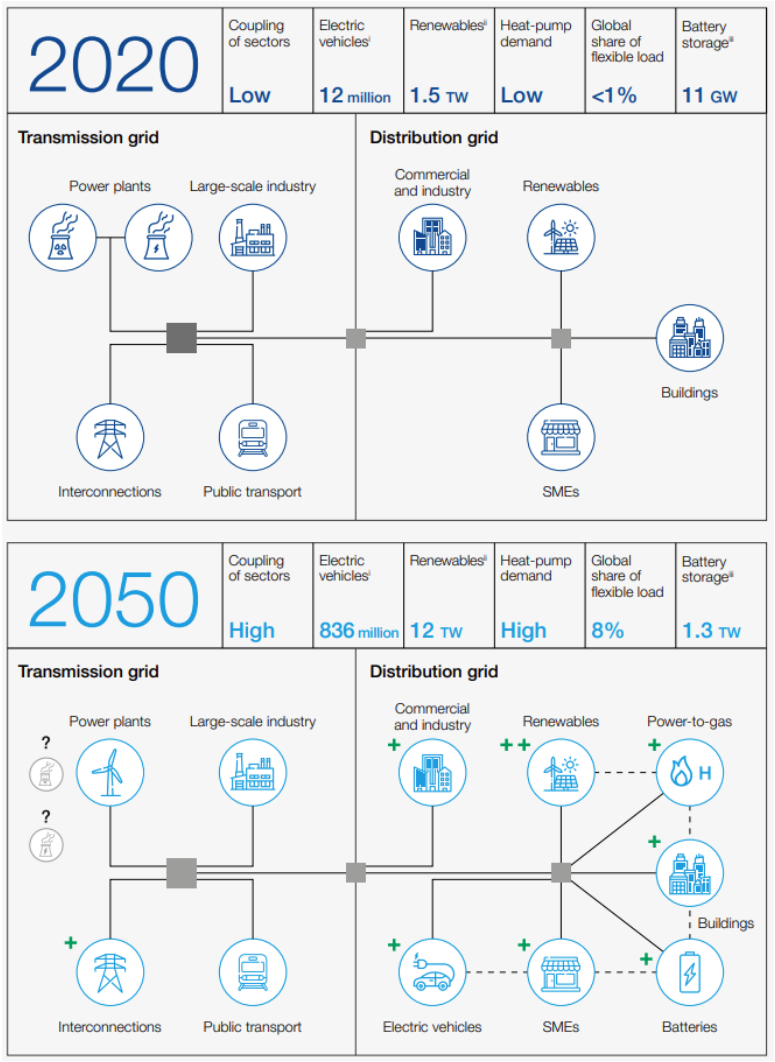


Figure 1.2: Transformation of global power systems as analysed by the World Economic Forum [21]

i-DE, as a leading Spanish Distribution System Operator (DSO), plays a leading role in the digital transformation of low-voltage networks through its STAR

project ('Sistema de Telegestión y Automatización de la Red') [22] and the subsequent PRADA project ('Proyecto de Renovación de Activos Digitales AMI'), deploying over 11 million smart meters and automating tens of thousands of secondary substations. This infrastructure enables continuous monitoring of LV assets and provides operational data essential for predictive maintenance algorithms. Within this context, the proposed methodology leverages SABT data to develop a hybrid anomaly detection system, combining analytical thresholds and machine learning models, to pre-emptively detect indicators of thermal stress and operational irregularities. Primarily, in secondary substations and, extensively, underground cable pits.

This approach aligns with Iberdrola's strategic emphasis on enhancing grid reliability and safety while reducing maintenance costs in increasingly complex environments. Furthermore, it complements the company's internal initiatives [7] which emphasises the importance of detecting alarm signals, such as temperature, humidity, and CO2 levels and integrating thermographic imaging [23] [24] to pre-emptively identify dielectric stress points and mitigate the risk of equipment degradation or service disruption [9].

1.1.2 Motivation

The aforementioned complexity of current LV networks with high DER penetration with a lower visibility, the increase of thermal degrading due to grid ageing and the vast amount of data currently being captured in the SABT systems are all both challenges and innovation opportunities.

Whilst there are several initiatives in favour of advances in digitalization and smart grids deployment, currently regulators remuneration schemes do not fully implement these incentives properly by favouring CAPEX investment rather than OPEX reductions and therefore investment lags on strategic vision [25] [26]. This is intended to be reverted in the following regulatory period [27] as it has been proven that grid reliability now requires a combination of grid reinforcement and more sophisticated flexibility and control schemes, as demonstrated by the 28A Spanish national blackout [28]. Therefore, obtaining new ways to effectively use the data collected via SABT for grid resilience, is a key topic for DSOs.

This project intends on proving the value and usefulness of such data to try and solve, pre-emptively, a current problem in distribution low voltage grids that severely reduces grid reliability and that, after research, has not been implemented prior to this project [29]. Such problem is hotspot anomaly events in secondary substations and more extensively in underground cable pits, where direct mea-

surements are unavailable. These events can lead to overheating or unexpected thermal related downtimes [23] [24] [9], compromising the grid integrity and the safety conditions around these elements. The proposed solution involves feeding operational data into a predictive maintenance algorithm that will be able to identify early signs of thermal problems or thermal degrading risk caused for example by a poor connection or unsafe operations. This approach is expected to increase the reliability and safety of the grid whilst also reduce its maintenance cost [30], key aspect that will be crucial for the future complex and expensive to operate distribution grids.

Additionally, as mentioned above, the novelty and relevance of this project is an exceptional and unique way to learn how the advanced metering infrastructure and smart meters can be used to obtain actionable insights through data analysis techniques capable of extracting relevant conclusions resulting in an increase in reliability. Contributing to both the scientific literature and practical grid management strategies.

Additionally, this project is aligned with several Sustainable Development Goals, as recorded in Appendix A.

1.2 Objectives of the work

The main goal is to generate an industrially viable predictive maintenance algorithm for hotspot events in both secondary substations and underground cable pits. In order to achieve this goal 4 main objectives have been devised for the project. Bellow they will be stated and further developed with their subsequent division into milestones and deliverables. Modularity is required to implement agile processes to keep improving the algorithm from a starting viable point. This modularity allows for continuous validation and scalability. Milestones are made so as to add key aspects at every development cycle of the predictive algorithm.

The four main objectives are:

OBJ1. Identification of historic secondary substation incidents by analysing the incident reports from the Outage Management System (OMS). Then apply graphical visualization of the line's state prior to the incident allowing for visual pattern recognition and anomalous state visualization using the LV – SABT data.

OBJ2. Develop a predictive maintenance algorithm for secondary substations. Design and implement a preventive anomaly detection system using both an-

alytical methods (threshold-based criteria) and machine learning techniques (RNN). The goal is to identify early signs of an overheating possibility using data from smart meters and secondary substation sensors.

OBJ3. Validate the prediction algorithm with real incident and field data. Compare the algorithm's predictions with historical incident reports and field measurements from secondary substations that experienced overheating. Emphasis will be placed on minimizing false negatives, even at the cost of increasing false positives, to ensure thermal anomalies are not missed.

OBJ4. Extend the methodology to underground cable pits. Apply the same detection framework to underground cable pits, where direct measurements are unavailable. Use customer smart meter data to infer abnormal behaviour and define detection criteria. Validate the new model with real incident data, where possible. This part introduces additional complexity due to the indirect nature of the data.

Next is a summary table with dates and a further breakdown of the main objectives:

Objective	**Key Technical Milestones**	**Target Date**
Misc	Obtain DataBase access	14/05/2025
	Get to know the DB and SQL language	19/05/2025
	Literature review	21/05/2025
OBJ1	Find historic CT incidents	26/05/2025
	Analyse the measurement graphs prior to incidents	31/05/2025
OBJ2	Obtain analytical thresholds	04/06/2025
	Find new viable features	22/06/2025
	Program and train ML model	03/07/2025
	Create predictive maintenance model (secondary substations)	03/07/2025
OBJ3	Application of thresholds to historic data	11/06/2025
	Result validation with incidents and field measurements	14/06/2025
	Test ML on historic data	07/07/2025
	Result validation with incidents (ML)	10/07/2025
OBJ4	Find historic cable pit incidents	15/07/2025
	Analyse the measurement graphs prior to incidents	21/07/2025
	Obtain analytical thresholds	25/07/2025
	Code predictive maintenance model (cable pits)	25/07/2025
	Result validation – cable pits	30/07/2025
	Field orders generation	01/08/2025

Table 1.1: Summary of objectives, deliverables and milestones for the project

1.3 Methodology overview

The research conducted was a data analysis with a deep connexion with the sources of the data and physical reasoning behind the behaviour of elements prior to the studied faults. As there is a strong industrial component the reasoning behind the tools produced during the investigation must and will be explained in the methodology section, chapter 3. Nevertheless, here is a brief introduction to the methods behind the project.

In order to achieve the objectives proposed, the methodology for the resolution of the problem is separated into 5 steps each with individualized tasks and inter-deliverables as mentioned above. The steps, or method followed, are:

1. Initial steps and data access
    1.1. Secure access to Iberdrola's database
        i. `CDSREADONLY`: OMS
        ii. `GENESIS`: asset inventory
        iii. `STG`: AMI measurements and event logs
    1.2. Familiarization with the database and the SQL environment
    1.3. Develop an automatized data extraction pipeline via python
2. State of the art and problem framing
    2.1. Study common thermal failure modes for secondary substations and the main causes
    2.2. Study common thermal failure modes for underground cable pits and the main causes
    2.3. Conduct literature review on predictive maintenance
    2.4. Conduct literature review on applications of machine learning in smart grids and digitalized low voltage environments
    2.5. Conduct literature review on anomaly detection techniques in power systems

3. Algorithm development and analysis
    3.1. Secondary substations analysis
        i. Identification of historical incidents
        ii. Data extraction and visual representation of statistical parameters of pre failure conditions
        iii. Definition of analytical thresholds based on expert knowledge and statistical distributions
    3.2. Secondary substations model development
        i. Model decision
        ii. Feature extraction and decision based on SABT and other available sources
        iii. Development of the model
        iv. Testing and training, model evaluation and improvement application
    3.3. Secondary substations model validation
        i. Extraction of test data
        ii. Validation against model
    3.4. Underground cable pit analysis
        i. Identification of historical incidents
        ii. Data extraction and visual representation of statistical parameters of pre failure conditions
    3.5. Underground cable pit model development
        i. Model decision
        ii. Model development
        iii. Testing and training, model evaluation and improvement application
    3.6. Underground cable pit model validation
        i. Extraction of test data
        ii. Validation against model

This methodology has been selected as the initial statistical analysis provides valuable insides on the possible relationships between the causes of hotspot events and the available data. The hybrid approach between analytical thresholds and ML ensures interpretability of the results and robustness. The modularity and method selected allows for result obtention and analysis of further steps in quick one week sprints. In that way the results can be analysed and improvements for the models can be implemented at a faster pace, always referring to the knowledge gathered on the past sections on the problem and its data distributions.

The resources required are a laptop with the capability to install python and SQLdeveloper. There are certain libraries and modules required alongside python such as Pandas, Oracledb, Matplotlib, Pyplot, Pyarrow, Datawrangler, Pytorch, Scikit-learn and TensorFlow. But most notably access to i-DE's databases is required for this project. The database accesses required, as stated above, are:

ICDS — acting as the OMS and is required as it records all the network's incidents, the grid elements affected, the durations and incident cause, origin or extra information in description form

GENESIS — is required as it contains all the tables related to inventory and its management

STG — is required as it contains all the SABT measurements from consumer smart meters and substation meters as well as event logs for the mentioned meters

1.4 Organisation of the document

Following the current introductory chapter, Chapter 2 describes the technical knowledge required for the interpretability of the gathered data and results by studying the causes for hotspot events in both grid elements analysed, as well as literature review comparing maintenance strategies and ML models. Chapter 3 details the methodology in more depth, explaining the models designed, the decisions and assumptions made. Chapter 4 contains the illustrative examples of the models alongside an evaluation of each model with respect to their validation. Chapter 5 provides the conclusions of the work.

Chapter 2

Theoretical framework and state of the art

The following chapter will review the common causes, impacts and issues related to secondary substations and underground cable pits. It is necessary to understand such causes as this will allow for an easier first filter on which data sources should be used to gather the most relevant information. It will also ensure the analytical study is performed with the required physical explicability. The impacts of secondary substation incidents will allow for a further economical analysis for break even minimum model requirements. Next, the state of the art is reviewed in all topics related to the project: maintenance strategies, low voltage anomaly detection and common machine learning approaches.

2.1 Common incidents in secondary substations: causes and impact

The incident this project mainly focuses on is hotspot anomalies. Such an anomaly in a low voltage electrical network refers to a localized area of overheating and excessive heat accumulation in a specific element of the grid. In secondary substations these appear mainly in the cable connections of the low voltage of the transformer, where currents are larger and many connections share the same low voltage panel or distribution switchboard.

The main causes of hotspot events in secondary substations are typically related to minimising the transfer area of the current therefore increasing the resistances and as such the temperature as seen in Figure 2.1. Some causes are: poor

contacts that have deteriorated with age, loose contacts that loosened due to a poor maintenance practice or strong short circuit currents that provoke magnetic forces on the wire, thermal and or conductive insulation failure due to extreme weather or ageing, presence of water or high humidity and there are other causes that whilst not greatly can have an accumulating effect on the problem like overloading, harmonic disturbances and phase unbalances [9][12].

Figure 2.1: Thermal image of a hotspot event in a secondary substation caused by a loose connection [source: internal maintenance reports from i-DE]

The consequences of such thermal events are varied from safety risks to maintenance crew, risk of equipment damage, damage to the integrity of the secondary substation, increased losses, system malfunctions and power outages.

Another relevant aspect of hotspot anomalies is their temporal evolution. These events often develop gradually, starting with minor increases in contact resistance or insulation degradation, which may go unnoticed in early stages. However, as the thermal stress accumulates, the degradation accelerates, potentially leading to abrupt failures. This progressive nature makes early detection through thermal imaging, continuous monitoring systems or data analytics crucial. Without timely intervention, a hotspot can evolve into a critical fault, causing unplanned outages, equipment failure, or other hazards as seen in Figure 2.2a. Therefore, integrating predictive maintenance strategies and anomaly detection algorithms becomes essential to mitigate the impact of such incidents on the reliability and safety of the low voltage distribution network.

(a) Damage caused by a hot spot event detected in a routine maintenance

(b) Thermal imaging of the hot spot event

Figure 2.2: Images depicting hotspot events and consequences in a secondary substation [source: internal maintenance reports from i-DE]

2.2 Incidents in underground cable pits

Hotspot events for underground cable pits are of a similar nature as hotspots in secondary substations. The igniting point tends to be a poor connection, nevertheless, due to the exposed nature of such grid elements there are several more ways in which the thermal or electrical insulation of the element can be initially damaged. Amongst some of the most common sources are NILED connections, where the original cable is perforated to make a new cable connection. Another common source is weather events, though if properly designed unless a different incident first took place environmental effects don't tend to be the sole cause of a fault but they are an aggravator.

Figure 2.3: Underground cable pit thermal event in Gastiez, Spain [*Source*: Newspaper Gastiez-hoy: "La Calle Diputación registra varias explosiones en una arqueta eléctrica"]

The consequences of hotspot events in underground cable pits tend to be loss of customer connection to the grid, short-circuit faults, temperature rises, equipment damage or degradation, power losses and thermal bursts due to arcing or gas accumulation. Gases might accumulate due to thermal insulation deterioration, poor cable joints, contaminated or sewage water draining [31]. Additionally, due to the distributed nature of this element, detection and correction is complex when not completely economically unviable (for both periodic and corrective maintenance schemes). Nevertheless detection, whilst complex due to the lack of direct measurements, is crucial as underground cable pits are located in urban areas and safety is a key aspect [6].

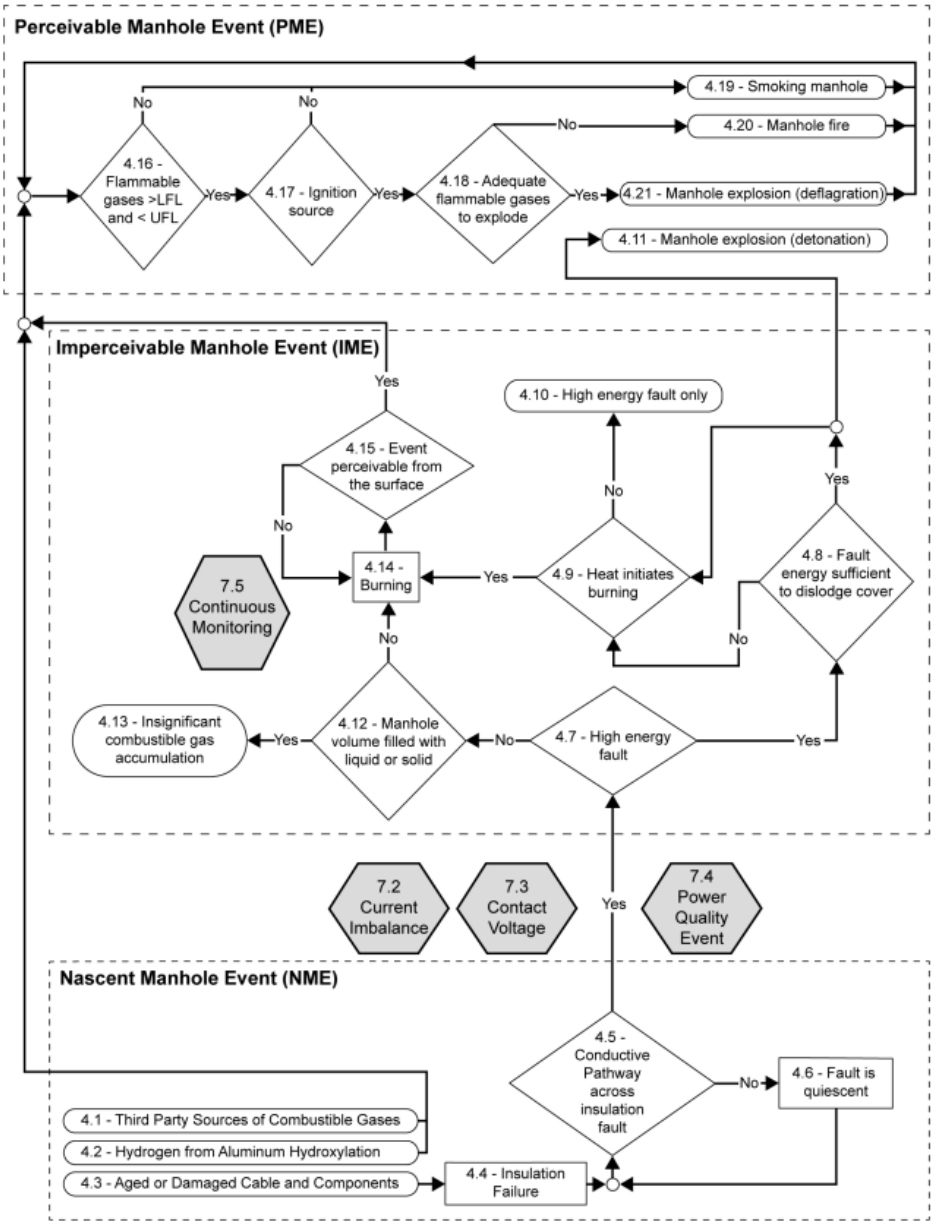


Figure 2.4: Common causes for underground electrical vault explosions as analysed by IEEE [6]

2.3 State of the art

Once understood the criticality of preventing these failure modes, a state of the art and competition analysis was performed to understand the different strategies for maintenance, how they could be applied to low voltage networks and how data analytics is used in different models to enhance these maintenance strategies. Later, in the Methodology chapter, Chapter 3, the basis of the actual model developed will be explained in further detail, taking this state of the art research as bases for the decisions made in model and structure.

2.3.1 Maintenance strategies

Maintenance can be, mainly, classified in three categories.

A maintenance strategy is a plan developed to minimise both downtime, maintenance costs (deemed both inefficiencies in the lean maintenance philosophy) and ensure operational continuity and efficiency [32]. As industrial systems become increasingly complex and data-driven, the range of available maintenance strategies have expanded significantly. Despite this, most approaches can be categorized into three main ones: **corrective**, **preventive**, and **predictive** maintenance.

Beyond these core strategies, several maintenance philosophies have emerged to enhance asset reliability and operational efficiency. These include **Reliability-Centered Maintenance (RCM)**, which prioritizes maintenance based on asset criticality; **Total Productive Maintenance (TPM)**, which involves all employees in maintaining equipment; **Risk-based maintenance** where the assets condition and its risk of failure is analysed; **Condition-based maintenance** by monitoring performance, control of corrective actions can be altered to best suit the asset [33] and **Lean Maintenance**, which applies lean manufacturing principles to eliminate waste and improve maintenance workflows. Each of these philosophies can fit within either corrective, preventive, or predictive frameworks, depending on the organization's goals, resources, and technological maturity.

In the following sections, each of these main strategies will be explored in greater detail, highlighting their principles, implementation methods and practical applications in modern DSO environments.

Corrective Maintenance

Corrective maintenance, also known as reactive maintenance, involves repairing equipment only once there is an element failure and corrective measures are required to achieve the original operational state. While simple and cost-effective in the short term, it often leads to unplanned downtime and higher long-term costs. Due to electricity being an essential good, regulators have enforced service continuity and quality standards which make corrective maintenance an undesired strategy for DSOs mainly due to the increased unplanned downtime and the immediate unavailability of spare parts for many electrical distribution grid elements.

Preventive Maintenance

Preventive maintenance is the next maintenance scheme, it aims to address potential failures before they happen. Achieving this where scheduled and periodic revisions or inspections take place to check the state of the element, typically revisions take place before and after peak usage of the element is expected or based on a yearly or multi-yearly scheme. Although it reduces unexpected breakdowns, it can lead to unnecessary interventions. Currently in distribution grids this maintenance strategy is the most deployed amongst DSOs [34] as it provides certain security against system failures although a compromise must be reached between constant inspections and reducing costs but allowing some room for failures and subsequent unplanned downtime to appear.

Predictive Maintenance

Predictive maintenance represents a significant advancement in the management of electrical distribution systems [12]. It relies on real-time data acquisition and advanced analytics to assess the current condition of system components and estimate their risk of failure. This approach enables maintenance activities to be scheduled based on actual usage and degradation patterns, rather than fixed intervals or post-failure interventions [35].

While reactive and preventive maintenance remain the most commonly implemented strategies due to their simplicity, the transition toward predictive maintenance is increasingly recognized as essential [9]. Predictive strategies enhance system reliability by reducing unplanned outages and minimizing unnecessary maintenance operations, ultimately leading to lower operational costs and improved asset longevity [10].

Despite its advantages, predictive maintenance in low voltage distribution networks still faces several challenges. Among them are the lack of commercially available turnkey solutions, the need for robust data infrastructure, and the integration of intelligent monitoring systems capable of detecting early-stage anomalies. These limitations highlight the importance of continued innovation and development in this area to fully realize the benefits of predictive maintenance [36].

2.3.2 Anomaly detection in LV networks

Anomaly detection is a growing fiend in the continuously digitalising and increasingly complex distribution systems. Some distribution companies, such as Enel's branch gridspertise, have started to use data for instantaneous anomaly detection [1] but the transition to predictive maintenance is yet to fully arrive. Other companies are using data intelligence to better understand failures and their causes and many startups have sprawl to cover the digitalization process and data intelligence knowledge gap.

Thermographic imaging is currently being leveraged [23] [24] to identify and prevent dielectric hotspots [9] and improve maintenance effectiveness. Also, as discussed above, anomaly detection and failure anticipation in LV networks is a growing research area, particularly with the increasing availability of data from grid digitalization. Nevertheless, currently there is no publicly available source or tool dedicated to predictive hotspot anomaly detection in low voltage networks via data analytics.

2.3.3 Review of Analytical and Machine Learning Approaches

To support this evolution towards predictive maintenance, certain tools can be leverage to various degrees of effectiveness. Here some of the most applied techniques in industry will be stated and referenced to the case study of anomaly detection in a DSO environment where possible.

A comprehensive review of anomaly detection techniques in energy systems [37] categorizes methods into:

- Physical models: Based on known system behaviour.
- Statistical thresholds: Triggered when measurements deviate from expected ranges.

- Machine learning: Including regression (e.g., neural networks), classification, clustering (e.g., k-means, GMM) and hybrid models.
- Image-based analysis: CNNs applied to thermographic images for hotspot detection [9] [23].

Notably, some studies highlight the lack of prior work, at least public, specifically targeting predictive maintenance for temperature-based anomalies in electrical distribution grids using ML [29], reinforcing the novelty and relevance of this project.

Analytical techniques

Analytical methods are a more traditional method of interpreting data. It relies on certain key characteristics like: being rule driven, having deep human interpretation and involvement, is applied to structured data (tabular, well organised data) and if the technical expertise is available its application is simplistic. It is basically a human based set of rules that come from experience or visual analytics.

This makes this method an attractive one for companies seeking deep explicability, although certain ML techniques are also explicable but the rules come from an automated process instead of a human. Other use cases of such models are well defined problems, when dealing with limited data available. regulatory requirements that need audit compliances (for the transparency and compressibility) and when there are resource constraints. Nevertheless, scalability is hindered and complex data patterns are typically not captured. All and all, this method provides several advantages and can be a good first level approximation to the data.

Machine Learning: supervised, unsupervised and hybrid approaches

Even though analytical threshold-based methods are just starting to be implemented for data-driven anomaly detection in low voltage grids, there is already a next step: machine learning (ML) techniques capable of identifying complex patterns and subtle deviations from normal behaviour [38].

Recent literature has explored a wide range of machine learning (ML) models for applications such as fault localization, cybersecurity anomaly detection [39], and non-technical loss identification. Although none of these studies directly address hotspot anomaly detection in secondary substations for failure anticipation,

the methodologies and insights derived from them are highly relevant to this research. The following categories summarize the most pertinent approaches:

- **Supervised and Ensemble Models:** Techniques such as Logistic Regression, Support Vector Machines (SVM), Decision Trees, and Naive Bayes have been widely applied in classification tasks [40, 41]. These models are interpretable and computationally efficient, but they require labelled datasets, which are often scarce in anomaly detection scenarios.
- **Unsupervised Models:** Algorithms like Isolation Forest, One-Class SVM, Local Outlier Factor, and Autoencoders are commonly used when labelled data is unavailable [41, 42, 37]. These models are effective in detecting novel or rare events but may suffer from high false positive rates and sensitivity to parameter tuning.
- **Deep Learning Approaches:** Multilayer Perceptrons (MLPs), Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have shown strong performance in real-time anomaly detection using smart meter data [43, 44]. LSTMs are especially suitable for time-series data due to their ability to model temporal dependencies. They can compare predicted or expected values against actual measurements to identify anomalies. While deep models require significant computational resources and data, their adaptability and scalability make them promising for dynamic grid environments.
- **Hybrid CNN Models:** CNN-based architectures have also been employed to identify regions of interest post-threshold, allowing for detailed simulation only in anomalous zones [45]. This selective modelling reduces computational overhead while maintaining diagnostic accuracy.
- **Contrastive Learning:** Recent studies propose contrastive learning to enhance feature representation and improve anomaly detection accuracy [46]. This approach is particularly useful in scenarios with limited labelled data, although it often requires careful design of positive and negative sample pairs.
- **Fault Cause Detection:** Many of the aforementioned models are evaluated based on their ability to detect root causes of faults such as overloads, phase imbalance, harmonic distortion, poor connections, insulation degradation, and thermal stress—factors commonly associated with hotspot anomalies [47, 48].
- **Clustering:** there are many clustering methods that are commonly used for anomaly detection, some will be mentioned here [49]:

– k-means clustering for fast and simple classifications but probabilistic starting positions for the centroids can yield different results also k-means uses mean value naively for cluster centroid detection, which not always is an efficient solution.

– Mean-shift clustering requires no initial estimate on the number of centroids but the selection of the observable window is non trivial and can greatly affect the results

– Gaussian Mixture Models (GMM) offer a more flexible probabilistic framework for modelling the distribution of normal operational data. They are particularly advantageous in scenarios with limited feature availability and can provide interpretable statistical insights offering variety in cluster covariance. However, they assume underlying Gaussian distributions and may struggle with very complex, non-linear patterns [50].

– Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is capable of identifying outliers as noise, a very important feature. But lacks performance if clusters are of varying densities.

– Spectral clustering uses graph properties to aggregate data points, very useful for non convex cluster structures, nevertheless, it is more computationally demanding than the above.

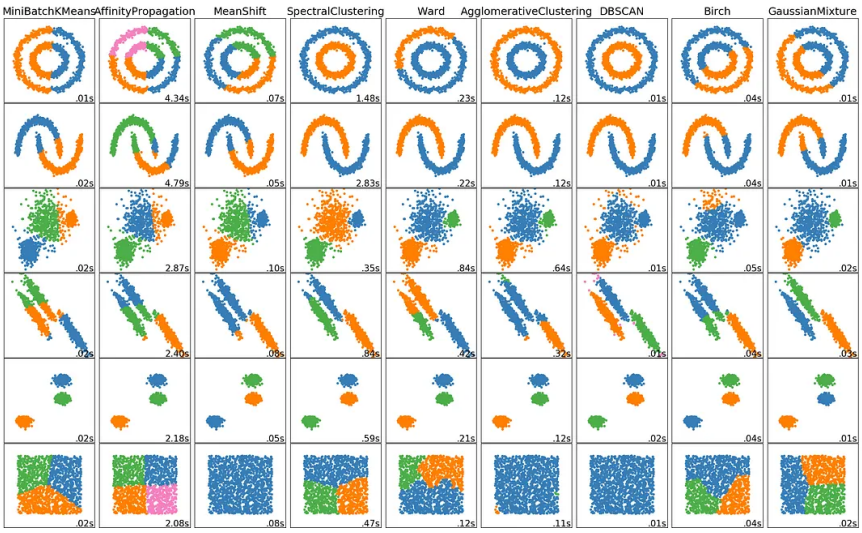


Figure 2.5: Cluster comparison of main techniques [49]

In the context of Distribution System Operators (DSOs), the choice of model depends on data availability, interpretability requirements, and computational con-

straints. RNNs, especially LSTMs, are favoured for their adaptability to time-series data and robustness in dynamic environments that have high temporal dependencies but require more computational power than multi layer perceptrons that are efficient in complex pattern recognition. GMMs, on the other hand, are valuable when feature selection is limited and a statistical understanding of system behaviour is desired.

RNN are deep neural networks used in machine learning models that train on time series data to make predictions on the next points. The distinguishing factor with other convolutional networks is that prior inputs can influence current inputs and outputs. Also, the weight parameter is shared for each layer as opposed to CNNs but both architectures are updated via back propagation and gradient descent. Although many different RNNs exist the standard RNN offers a computationally simple first step in the model development, required for a work based on agile philosophy. Also, vanishing gradients, the method by which the model loses grip of long term dependencies as they have a reduced impact on the next predictions gradients, is a lesser problem in anomaly detection of this type as immediate deep change is one of the indicators for hotspot events. Nevertheless, this vanishing gradient can be solved if required by the use of a long shot-term memory (LSTM) [51].

Multilayer Perceptrons (MLPs) are feed-forward neural networks composed of multiple layers of interconnected neurons, each applying non-linear transformations to the input data. Unlike recurrent architectures, MLPs do not retain memory of previous inputs, making them computationally efficient and well-suited for tasks where temporal dependencies are less critical or can be captured through feature engineering. In the context of hotspot anomaly detection in power distribution grids, MLPs offer a robust baseline due to their ability to model complex non-linear relationships between sensor readings, environmental variables, and operational parameters. Their simplicity facilitates rapid prototyping and deployment, aligning with agile development methodologies. Moreover, since hotspot events often manifest as abrupt deviations in spatial or operational patterns rather than long-term temporal trends, the lack of sequential memory in MLPs is not so much a limitation but rather an advantage for this study, reducing model complexity and training time [52].

On the other hand Gaussian Mixture Models are statistical or probabilistic model that select a limited number of multivariable Gaussian distributions to explain the data points. The distributions are so that the points could have been generated by the mixture of Gaussian distributions. With some key differences on the method, they are also classifier models like a more known k-means that group data into clusters, but in this case the data is classified into Gaussian distributions

with different means and covariances between the features. Its simplicity allows for fast a clustering technique.

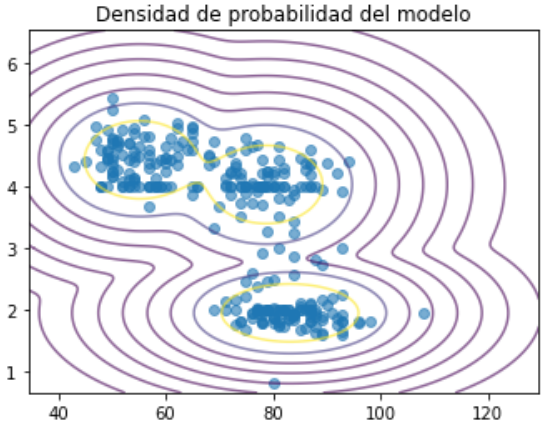


Figure 2.6: Bivariate GMM representation of normals as contour lines [53]

Chapter 3

Methodology

In this chapter the methodology, steps taken and justified decisions made for the completion of both hotspot predictive maintenance algorithms will be analysed. First reviewing the data sources and the preprocess of such data, continuing with the anomaly detection methodology for the secondary substations algorithm and finishing with the underground cable pit algorithm. For each algorithm the first data exploration and statistical analysis will be commented and then the model created is explained with its development, assumptions made, data structure, obstacles rencountered an model limitations.

3.1 Data Sources and Preprocessing

The data for all models will be extracted from i-De's internal data bases. As mentioned in the methodology overview, Chapter 1, there are three main data bases from where all the data is extracted: ICDS, GENESIS and STG. These databases are accessed through Oracle SQL Developer for exploratory means and via python for extraction purposes.

The ICDS database is mainly to store a recollection of all outages that occur in the distribution network. The database has three main tables, along with their auxiliary tables, that collect all the required data for this model. These are the incidence report table, the phase table and the 'ambito' or elements related to each incident table. The initial incident report contains the date, the approximate geografical location, the type of failure mode and a brief initial description. Incident logging follows a 5 step process called phases, recorded in the appropriate table. Phase one is the incident generation, then the mobile assignement, the fail-

ure location and isolation with non served energy, the provisional resolution for the disconnected clients and then phase 5 is the final definitive resolution of the incident. After this final step the system has returned to the initial state. The elements affectet table records the code of the elements that lost power because of the failure and for each element it provides a date time of lost of power and a date time for the reconnection of power, from wich the duration can be extracted.

The GENESIS, GIS - Geographical Information System, data base serves for inventory purposes. It is from this table from where the information for each element can be extracted. Some of the important information extracted are number of customers connected, power served, nominal currents and voltages of different elements ...

The STG records all the measurements as time synchronous **profiles**, with different time granularity depending on the type of measurement. It also records the **event logs**. Both profiles and events can originate from two different sources: either the **advanced low-voltage line supervisors** located in secondary substations, or the **smart meters** located downstream at consumption points.

For the purpose of this project, we focus on measurement profiles provided by the advanced line supervisors in secondary substations. These include voltage, current, power and temperature profiles which are collected in each line supervisor at 5 minute intervals, as an average of the measurements taken during those 5 minutes, and then sent all in one data packet at the end of the day.

In terms of the events, specially those this thesis focusses on, are generated by smart meters and arrive asynchronously. Once the event takes place the data is sent and it contains: the group, type, starting and ending date times and some additional information depending on the event.

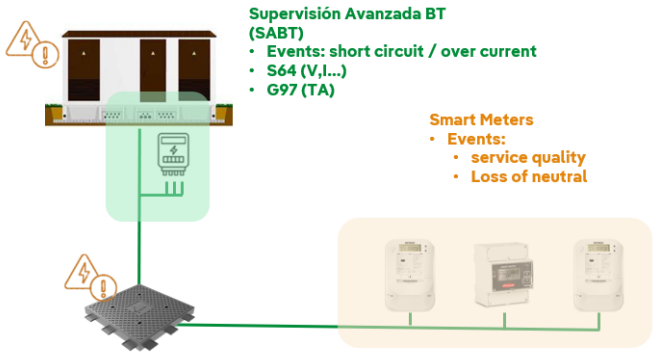


Figure 3.1: Data sources (SABT and Smart Meters) and visual representation of a secondary substation and an underground cable pit

Around September 2024 the last SABT meters were updated to allow for temperature readings, as it was deemed an important feature of the line to be taken care of analytically. This is why current SABT systems record and send back via STG 2 key temperature indicators. One is the ambient temperature recorded by the smart meter, which as the temperature sensor is placed adjacent to the line it has a certain relationship with the actual temperature of the line. The other is an estimated temperature of each of the phases and the neutral wire. The estimations is made by combining:

- **TPTMax**: Maximum allowable temperature of the cable, depends on the insulation material, in °C.
- **TPTlimit**: Ambient temperature at which the manufacturer specifies the nominal current of the conductor, in °C.
- **TPInom**: Nominal current of the conductor, in amperes (A).
- **TPThermalC**: Thermal constant of the conductor, depending on its cross-section and installation method (e.g., air, buried), in seconds.
- **TPCurrentMeasMax**: Maximum limit for current measurements in phases and neutral, used to avoid false emergency alarms due to measurement errors.
- **TPTamb**: Ambient temperature, in °C.
- **Fa**: Adjustment factor based on ambient and limit temperatures.
- **H(t)**: Thermal level of the cable at time t, dimensionless.
- **I(t)**: Measured current at time t, for each phase or neutral.

Formulas used in the calculation of cable temperature:
UNE.EN 60255 norm, European Standard [54]

$$Fa = \frac{TPTMax - TPTlimit}{TPTMax - TPTamb}$$

$$H(t) = \left(\frac{I(t)}{TPInom}\right)^2 \cdot \frac{15}{TPThermalC + 15} \cdot Fa + \frac{TPThermalC}{TPThermalC + 15} \cdot H(t-1)$$

$$Temp = TPTamb + H(t) \cdot (TPTMax - TPTamb)$$

For the predictive maintenance model related to underground cable pits, the data sources will come from the same database as the above information (STG - tstgreadings) but will come from a different set of tables, those related to meter events. These events are separated into the smart meters located at consumption points, also called Remote Terminal Unit (RTU), and those sourced from advanced line supervision system meters located in secondary substations, called LS cards events. The events are further categorised into synchronous events and asynchronous or spontaneous, it is these latter events that are of a greater interest for this study. The information retrieved is sent from the meters to the centralised information system using standardized protocols and reports such as S65, S67, S59, S63 ...

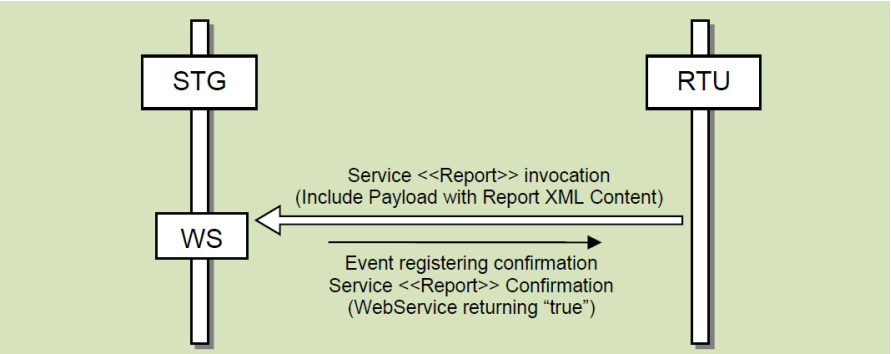


Figure 3.2: Protocol example: RTU asynchronous event report via the STG Web-Service [55]

For each event source there are several groups of events and a further division of the groups into event types. Here is a summary table on such division with a brief description of each event. Those that are of more interest to this project, as they are analysed to be the main flagging events for underground cable pit hotspot events, are marked. Missing types are reserved for future developments and are therefore not included in the summary table, also certain events which are similar to one another are gathered together for simplicity:

Table 3.1: RTU Event Groups and Types

Group	Type	Description	Relevant
1	1–3	Logger startup and power failure	
1	4–6	Phase voltage measurement failure (1, 2, 3)	
1	7	Neutral loss detected	✓
1	8	Low battery	
1	9	Critical internal error	
1	21–23	End of phase voltage failure (1, 2, 3)	
1	24–25	Official time change (Winter/Summer)	
1	26–29	Active energy impulse LED control (manual/remote)	
1	67	End of neutral loss	
1	68–71	Reactive energy impulse LED control (manual/remote)	
1	72–75	Active + reactive impulse LED control (manual/remote)	
1	30–34	Configuration changes (logger, ports, passwords, firmware)	
1	35	Battery reset performed	
1	36–43	Various configuration changes (auto I/V, integration period, transformation ratio, sync type, labels, output assignment)	
1	44–66	Contract 1–3 operations (closure, parameters, schedules)	
1	90–91	Timing changes for voltage variation and long interruption	
1	92–95	Reference voltage and threshold configuration changes	
1	96, 102	Contracted power changes (import/export)	
1	97	Firmware update (version change)	
1	99–101	Reset operations (keys, data, parameters)	
1	103	Max software update count reached	
1	108	Scroll mode change	
1	110	PLC modem reboot	
1	112–116	Auto-reconnection settings	
1	117–118	Smart 3-phase injection settings	
1	119–120	Overvoltage/neutral loss detection toggle	
1	121	Prime protocol change	
2	1–3	Manual and remote connection/disconnection (button/-command)	
2	4, 13	Disconnection due to contracted power limit (Contracts 1 and 3)	

Continued on next page

Table 3.1 – continued from previous page

Group	Type	Description	Relevant
2	5	Connection via power control (IGA)	
2	6, 14	Element lock/unlock due to PSC overlimit (import/export)	
2	7–8	Element enable/disable	
2	9–11	Residual power control: disconnection/connection	
2	12	Change in cut-off element control mode	
3	1–4	Phase voltage below lower limit (average and per phase)	✓
3	5–8	Phase voltage above upper limit (average and per phase)	✓
3	9–12	Long-duration outage (all phases and per phase)	✓
3	13–16	Phase voltage below lower limit (duplicate set)	✓
3	17–20	Phase voltage above upper limit (duplicate set)	✓
3	21–24	Long-duration outage (duplicate set)	✓
3	25	High impedance fault detected (BT Supervisor)	✓
3	26	End of high impedance fault (BT Supervisor)	✓
3	27–30	Phase voltage below lower limit (Distributor event)	✓
3	31–34	Phase voltage above upper limit (Distributor event)	✓
3	35–38	Phase voltage below lower limit (Distributor event, alternate set)	✓
3	39–42	Phase voltage above upper limit (Distributor event, alternate set)	✓
3	43	Reconnection due to overvoltage or neutral loss / automatic trip	✓
4	1–2	Manufacturer seal opened/closed	
4	3–4	Magnetic field detected/cleared	
4	5	Current detected without voltage	
4	6	Intrusion attempt (wrong password)	
4	7–8	Terminal cover opened/closed	
4	9–10	Voltage detected at output terminals during remote disconnection	
4	11–12	Infinite impedance detection at output terminals during remote disconnection	
4	13–14	Meter bypass started/ended (optional)	
5	1–3	Demand response order (critical power: residual, % reduction, absolute)	
5	4–12	Demand response order (non-critical residual power, % reduction, absolute)	

Continued on next page

Table 3.1 – continued from previous page

Group	Type	Description	Relevant
5	13	Change in contracted residual power value	
5	14–19	Activation/end (residual power, contracted power % reduction and absolute power reduction)	
5	20	Demand power near contracted limit (%)	
5	21–22	Auxiliary relay connected/disconnected	
5	23–24	IHD enabled/disabled	
5	25	Change in scheduled time	
6	1–2	PLC port communication started/ended	
6	3–4	Optical port communication started/ended	
6	5–6	Serial port communication started/ended	
7	1	Key reset	
7	2–6	Security key and policy changes (Master, Encryption, Authentication, LLS, Policy)	
7	7–13, 101–110	Errors in security configuration and secure communication (LLS, client, PLC)	
7	14–15	Secure client modification (optical port and reader)	
7	16–17	Public key update for firmware signing (Iberdrola and Manufacturer)	
7	18–19	Secure client reset (PLC and optical)	
7	111–117	Errors in public key (Iberdrola/Manufacturer) & Firmware update errors (signature and CRC verification)	

Table 3.2: LS Card event groups and types relevant to underground cable pit hotspot detection

Group	Type	Description	Relevant
1	1–2	Startup with/without data loss	
1	9–20	Manufacturer errors	
1	24–25	Seasonal time change	
1	26	High temperature detected	
1	27	End of high temperature	
1	30–34	Configuration changes (RS485, passwords)	
1	36–41	Parameter adjustments (transformation ratio, clock, labels)	
1	66–71	Period changes (profiles 1–5, real-time)	
1	97–98	Firmware update and synchronization	
1	100–101	Reset to default and data deletion	
1	102–117	Nominal voltage/current and limit changes	
3	26–31	Fuse activation/deactivation (phases R, S, T)	
3	40–45	Short-circuit activation/deactivation (phases R, S, T)	✓
3	46–51	Overload activation/deactivation (phases R, S, T)	✓
3	60–63	Thermal image protection active/inactive (phases, neutral)	
3	64–71	Thermal image alarm activation/deactivation (phases R, S, T, neutral)	
3	72–79	Thermal image emergency activation/deactivation (phases R, S, T, neutral)	
3	60–63	Thermal image protection status (active/inactive)	
6	5–6	DLMS association established/released	

As seen above, the amount of data sources is varied and the granularity is big. Therefore the amount of data managed is vast. Clear knowledge of the root causes, consequences and possible system manifestations is key to filter out the features and desired data sources to avoid hugely inefficient processing times.

The most relevant information from these event logs are in group 3, which for both RTU (consumer smart meter) and LS advances supervision systems, is related to power quality. This is because a hotspot event in an underground cable pit will alter the voltage profile of the line, with higher loses and higher voltage drops which will cause quality indicators to flag the line and send the events back to the main database. Whilst a monophasic grounding fault in an underground cable pit could be passed by the meters as a new load due to its small current, meaning that current measurements would not be useful for this case. Return to

earth fault detection systems are required for this current detection.

To visualize this amount of data and taking just the secondary substation model for simplification. The data is collected at five-minutes intervals, we typically are interested in current, voltage, temperature measures alongside other descriptive values like element reference number, location, exterior temperature (which is obtained by using the date time and the GPS coordinates of the element to access a publicly open source available climate API called Open Meteo to collect this data, cleaning the data to mold it into the 5 minute interval measurements) and some values like ventilation type and type of secondary substation. If around 16 values are collected every 5 minutes for the almost 550,000 lines and the 11 million smart meters in a one year study there would be around 9.3×10^{11} data points and that would equate to around 900 GB of data to be processed (extrapolating some lower scale database extractions performed). If each data point was a grain of rice, 8.7 Olympic swimming pools could be filled or if each byte was written in a page and printed the paper pile would stack to 90km, almost reaching the Kármán Line, the height where space starts, according to the FAI (Fédération Aéronautique Internationale). Though achievable for a company with access to cloud computing, too extensive for an individual model validation process.

This is one of the key limitations of the models presented below, even for small scale studies the amount of data and processing time is significant, so a trade of needs to take place between industrial scalability and manageability. Nevertheless, due to the critical nature of identifying all possible hotspot events, due to the safety risk related consequences, all models presented have been trained with the data of all detected hotspot anomalies in the past year and then a control population of undamaged lines that did not contain hotspot events was selected to validate the model and verify false positives.

3.2 Anomaly Detection in secondary substations

This section intends on explaining the method and decisions taken towards the development of both the analytical model of the data with physically explainable criteria and the machine learning models that result in the predictive maintenance algorithm for hotspot anomalies in secondary substations. But first the methods behind a statistical analysis of the data alongside some preliminary observations that have founded the models will be displayed, alongside the method used to perform an economical review that will serve as a limiting factor for the model precision required to break-even.

3.2.1 Historic incident, statistical and economical study

Historic incidents

The data available for this study was presented in Chapter 2. These were measures coming from the SABT system for each line of the secondary substation. Including date-time, current of the 3 phases plus the neutral wire, voltage of the 3 phases, thermal image of the 4 wires (obtained via the calculation explained in the above section), ambient temperature (as measured by the sensor in the smart meter which, as it is placed on the line has certain relationship with the lines temperature), external temperature (from a weather API), geographical zone code, type of ventilation, type of secondary substation and the respective line and secondary substation codes for identification.

To decide on the model, the features and the criteria studied an initial statistical distribution and variable correlation study was performed. The main objective for this study is to detect key differences between healthy lines and damaged lines that could then be exploited by the predictive algorithm models to distinguish between the two. For this key first step we require a dataset of damaged and healthy lines that keep the majority of things constant such as type of secondary substation, time of the year, loading ... To achieve this first the historic incidents related to secondary substation hotspot events had to be detected from the OMS database and then 2 periods would be obtained:

- The months prior to the incident were we consider the line to be damaged as it resulted in an unplanned secondary substation downtime caused by a hotspot event (without including the actual incident day as this would distort the data and the key factor of the algorithm is its predictive nature)
- The data related to the months after the incident was cleared, the corrective maintenance was performed and the secondary substation was now acting as designed with all damaged components replaced.

To detect which secondary substations had suffered hotspot events first the OMS database (ICDS) was accessed and all incidents for the past year, that were recorded as **unplanned outages** related to **medium voltage** or **several lines** (these were found to be the most successful markers) were extracted. The output was a list of all incidents with its unique code, a description of the incident and a date time. The description column was cleaned and normalised to ascii small characters with no double gaps or accents, this is important as the description is typically written in field by the operator that inspects the incident and could have errors. This description list was then **filtered by those that contained**

references to elements that are common consequences of hotspot events such as smoke, heat ... then a second filter was performed to those containing the desired location or element of such hotspot by finding those that also contained 'CT' (centro de transformacion), electrical box, low voltage. Finally 2 more filters were used to discard common incidents that occur in secondary substations that had similar descriptions as the desired ones but had a known different source, these are related to fuses, circuit breakers, cells, medium voltage lines ... Again, as these descriptions are written in the field, the word filtering had the complexity of using where possible word roots to avoid the effect of misspells.

This pre-extraction filter alongside the word filter reduced the size of the incident list from hundreds of thousands to about one hundred instances. It is possible some hotspot events were lost to some filter but several filtering techniques and different word combinations were used to loose as little hotspots whilst keeping the filter strict to avoid poor labelling, as this would greatly affect the performance of the models. Next, as the filters are not perfect, a by hand filter is then performed for the last tens of incidents by carefully examining the description and available information. This process could be further automated by the use of a large language model, but the use of such models is internally restricted by the company, so filtering by hand was the only option left.

Once the incidents have been filtered, another extraction is performed to obtain the list of the three most affected elements related to each incident. Being most affected those that had the longest outage time. This is an effective strategy as hotspot events will require corrective maintenance before reconnecting the secondary substation permanently but the low voltage lines and consumers can be reenergized from a different secondary substation or an emergency backup generator. Then for each incident, if the name of any of the three elements is included in the incident description, this tends to be the case for secondary substations as they have distinct names by which they are referred to by the maintenance crew that writes the descriptions, then that element is selected for that incident, if not then the largest duration element is the one selected. This results in a list of several secondary substations where hotpot events caused an unplanned outage.

This list is used to extract the SABT meters connected to these secondary substations, it is from these meters where the measurements will be sourced. The final extraction is using the final meter list and the incident date time, obtain the SABT measurements for the months before the incident and after the corrective maintenance. The current measurements are normalised by automatically finding in genesis the nominal current of such lines.

The basis for the statistical study is the recorded incidents with the measure-

ments for when they were healthy secondary substations and damaged ones but for the model training and creation two other categories were included. First the secondary substations that, whilst damaged due to hotspot events, they did not cause an outage (and so were not recorded by the OMS) but were recorded in the preventive maintenance logs. Finally a control population of secondary substations was included in the study, these had similar loading characteristics, region locations and types of ventilation than the other 2 categories but more were included, at a reason of 1:10 to 'simulate' the model performance where hotspot anomalies are uncommon.

Therefore, as a summary, the data used contains instances of both healthy and damaged lines and a time frame of about 7 months per substation. Damaged lines were detected via the OMS, the point were corrective maintenance was applied is recorded and from then on it is considered a healthy line once more. For the control population it is always considered as a healthy line.

There are 140,000 data points of damaged lines that eventually caused a significant incident (accounting for 8 secondary substations), a further 600,000 data points of these same secondary substations once the corrective maintenance takes place and they can now be considered healthy lines, 2,200,000 points of damaged lines that were repaired in routine maintenance inspections and caused no significant outage or incident recorded by the OMS (accounting for 20 secondary substations), 2,000,000 points of these same secondary substations once the corrective maintenance took place and they were now considered healthy lines, finally, a further 6,600,000 data points of a control population that shares the characteristics of the other groups, assumed to be healthy (accounting for 100 secondary substations). The time study tends to be of 2 months per type discussed here, of the 5 discussed types.

Statistical analysis

Several statistical analysis were performed in order to visually inspect the data and detect differences between healthy and damaged lines. Results will be presented in the next chapter but the tests performed and the methodology will be explained here.

After extracting the data it is read by python where several data analysis libraries are used to plot certain relationships. Most of the data, after the preprocessing mentioned, is then removed of erroneous values like nans, infinite values, those lines with no nominal current are removed, also those measures that exceed significantly from the nominal values (if these values were real circuit breakers and

fuses would have gone off before reaching these levels), the values are reshaped into floats, phase averages are calculated for some of the analysis and some minor data processing for python library uses.

The analysis performed were:

- Temperature measurement histogram
- Several current and temperature relationship pairplots and current squared vs temperature plots
- Phase balancing analysis plots
- Probability distribution KDE (Kernel Distribution Estimation) plots for temperatures and currents
- Bivariate KDE plots for current temperature relationships
- Line comparison for the same time period and same secondary substation in terms of current and temperature
- Time distributions for the evolution of line conditions

Economic analysis

To study the precision required to achieve economic break-even point, an economical study was performed. The study uses internal data or approximations for the study of the yearly costs and benefits of implementing a large scale model for hotspot anomaly detection. The costs are:

- The cost implied in visiting a secondary substation marked by the model as a positive for hotspot. In total 120€ for the team displacement, a simplification is made here assuming the team has enough time to add this to the calendar and requires no prioritization over a different work made by the team.
- The cost for the maintenance required for those secondary substations visited that had actual hotspots in them of 1200€. The cost is approximated at 30% of such cost as per technical expert recommendation, some hotspot maintenance does not require an equipment change but a clean-up, minimum element change or simply just mending the loose connection.

- The cost for the model execution based on the execution times and the costs for Azure's virtual machines [56] cost comparison of family D, selecting the most appropriate computational characteristics for each model in terms of price, vCPUs and memory. Nevertheless, this cost was finally excluded form the model due to its insignificance as it amounted to less than 50€ per training execution for the final combination of models (the most time consuming part of the model but with the least frequency)

The 'benefit' comes from the avoided cost related to the consequences of a hotspot event, set at 120000€ (approximated cost of replacing a secondary substation, also reduced to 35% as for the maintenance cost)

The combination of these items, with the estimated number of hotspot events and the number of secondary substations, results in a precision limit for the model so that economical feasibility can be granted. Although all direct costs and benefits are taken into account there is an indirect benefit of minimising public exposure to hotspot events that is not accounted for in this analysis due to its complexity and social bases. Also, there is a cost related to safety risks imposed by a hotspot event and its possible consequences, neither of these indirect social costs are in the scope of this project but must be taken into consideration for a full view on the economic side. Nevertheless, if economic feasibility is granted for a model, this indirect marketing benefit will help increase the appeal of such model.

3.2.2 Analytical criteria

After the first statistical review of the data, the analysis of the different features available and the state of the art research to find out what has been done and understand both the causes and consequences of hotspot events in secondary substations, an analytical criteria based model was developed. This model was intended to perform a further study of the data and the distinguishment between healthy and damaged lines and to base the decision making on physical phenomenon and expert knowledge making a more understandable model that also required less training and where more iterations could be carried.

This physically based model is recommended by some articles stated in Chapter 2 as an initial analysis and it also followed the ideas shared in i-DE of performing a fast MVP with threshold criteria. More than 20 criteria with physical bases on hotspot causes were analysed and thresholds devised. For this thresholds, were reasonable, statistical values were used such as deviation form a normal rather that hard coded values that could be more prone to errors of model expandability.

Once the criteria were selected, some combinations of criteria were analysed.

Combining criteria allows the model to capture slightly more complex relationships. Finally, as the study is carried over a period of time it is also desired to know for each criteria the optimum percentage of time this threshold must be surpassed, out of the total amount of studied time, so as to consider it a sufficiently anomalous behaviour and that the model should flag the secondary substation as possible hotspot containing.

These processes were carried following an agile work flow, first a couple of more evident hotspot cause related criteria were analysed and the thresholds developed to design the first model. Then, incrementally, more criteria were added as more relationships were detected and researched.

3.2.3 Machine learning modelling

The selection of the ML predictive model was guided by a combination of benchmarking studies [43] [44], expert recommendations, and practical considerations. A Multilayer Perceptron Regressor (MLP) was chosen due to its proven effectiveness in modelling non-linear relationships in multivariate datasets [57], their relative simplicity in implementation compared to more complex architectures, and their ability to capture dependencies, critical for anomaly detection [58]. Unlike recurrent architectures, MLPs do not rely on sequential memory, which aligns well with the nature of hotspot anomaly detection in power distribution grids, where abrupt changes in system behaviour are more indicative than long-term temporal patterns. The decision was also influenced by the need for a balance between computational efficiency and predictive performance. Although there are ML techniques that will perform better and more efficiently [59] [60] the proven capability of Neural Networks and their simplicity to train were key aspects in designing an MVP that proves the effectiveness of such models. As future work other ML techniques, with a better computational power can be assessed, but this model proves the validity of doing such a study.

Other key design decisions included the selection of input features, the size of the time window used for training, and the choice of performance metrics. These decisions were informed by domain knowledge and preliminary data analysis.

Two models were developed to compare between them and find out if a general model that is trained on all the data and is cheaper to train would be enough to detect the necessary cases or is a specific ML model created for every secondary substation would vastly outperform this initial one.

The input data to both models consists of multivariate time series collected from sensors installed in the secondary substations (SABT meters). Each data

point includes timestamped measurements of electrical parameters, environmental conditions, and operational metadata. Both models intake the date-time, phase currents, phase temperatures, zone code, outside temperature, ventilation type, secondary substation type. The models output include predicted values of the ambient temperature for each line as measured by the SABT meters of each line (which are related to the lines temperature) and anomaly flags based on the error thresholding mechanism. As training takes place with healthy state data, when this behaviour changes then an anomaly is detected. These outputs are designed to be interpretable and actionable for maintenance and operational teams to avoid the consequences of undetected hotspot events.

The generalized model, applicable to all secondary substations, is designed for scalability and rapid deployment. The steps the model undergoes are:

- Data is extracted and cleaned
- The healthy secondary substation data is extracted
- A random sample (10% trade off between execution time and precision of model optimization) of the new data frame is selected for model optimization
- Random train test split is performed (75/25 respectively, as per technical recommendations in reference documents, even though temporal splits are preferred for some data analysis [61])
- Several different model parameters are tried to decide on the optimum MLP architecture. Main parameters searched are model hidden layer sizes and learning rates, the model is retrained for an extensive list of parameters and the one with the least error is selected
- A new random sample is selected, now with a larger data set, double in size of the previous sample
- Another train test split is generated and the MLP with the optimum architecture, is retrained
- The model and standard deviation for the errors in training is stored as these will be used for the predictive algorithm
- The predictive algorithm is a model that predicts the ambient temperature the line should be recording with all the other features, the error is compared to the 3 times the training standard deviation, if the error is larger than the limit for 3 consecutive instances then the model flags the line as an anomalous line with a high probability of having a hotspot event. A more restrictive rule than either Nelson Rules or Western Electric Rules [62] for control charts.

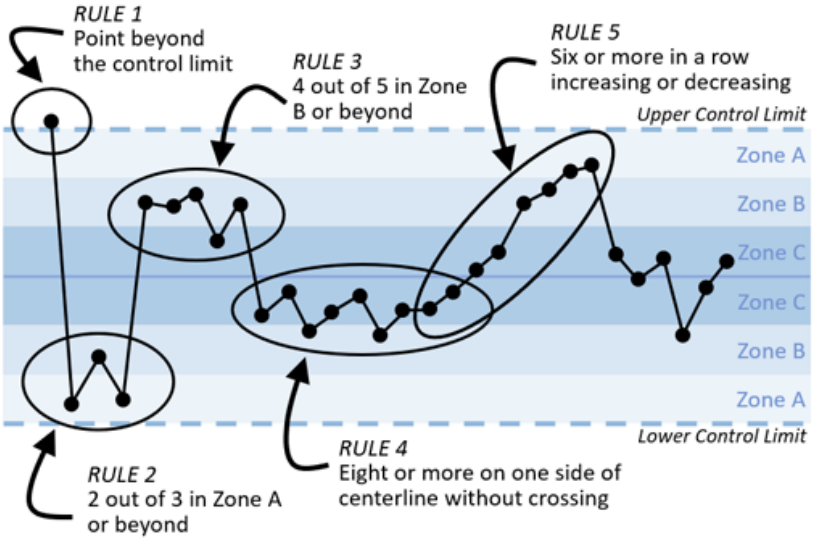


Figure 3.3: Main rules to detect out of control systems [63]

Then, a specific model tailored to each individual substation, offering potentially higher accuracy at the cost of increased training time and complexity, is developed. The basis of the model is similar as the general one, following the same steps except, now repeated for every secondary substation. Also, as now less data is used per model, the trade-off between precision and execution time now allows for a larger dataset to be trained on (percent wise). Each secondary substation follows the same steps, optimizes the architecture for that specific secondary substation, trains the optimal model on healthy state data and then detects if 3 consecutive errors between prediction and actual measurement lie above the 3 sigma training limit. As opposed to the general model described above this model thrives for a better precision over a sightly more computationally expensive process.

It is assumed that the input data collected from secondary substations is of high quality, meaning it is accurate, complete, and free from systematic errors or biases. It also assumes the correct labelling of healthy and damaged stage for training. Also, the model is based on the assumption that the healthy state of a substation remains statistically stationary over the training period and deterioration on that state is minimum. The anomaly detection mechanism is assumes that the prediction error follows a normal distribution. This statistical assumption justifies the use of the 3 sigma rule. While this is a strict approach, it may not hold in all cases, and deviations from normality could affect detection sensitivity and allow more secondary substations to be flagged as anomalous.

3.3 Anomaly detection in underground cable pits

Next, the methodology related to the predictive algorithm for underground cable pit hotspot detection will be presented, but first a view on the available measurements. As there is a lack of direct measurements, detection and differentiation is more complex than in the above case. A decision has been made to analyse the event logs from the SABT elements upstream from the incident but specially using the service quality events from downstream smart meters in the consumption points. This is due to the fact that line problems should manifest in the downstream section of the line more vigorously by unexpected downtimes or voltage deviations. Whilst this is not ideal, there is no other representative data sources for the line and statistically no visible difference, when looking at upstream data sources, between damaged lines and overloaded ones.

All this means that the methodology behind the model generation for the predictive maintenance algorithm is significantly different from the secondary substation case. Though as the information is recorded in the same databases some access and filter processes can be reused, these will be noted and changes explained, for the new algorithms designed a more in depth analysis will be performed.

3.3.1 Historic incident and statistical study

Historic incidents

Historic incident identification for model labelling follows a similar process to secondary substations, with the exception to data sources, features, parameters, filters and meter type. Nevertheless the broad method is obtain all of last year's incidents categorised where underground cable pit hotspot events could be classified, filter by word description, by hand filter, element identification with downtime duration, smart meter and SABT system meters related to the elements detected in the past step, measurement (or in this case event log) extraction, data preprocessing and model creation (with its subsequent iterations following the agile methodology to improve the outcome of such model).

As mentioned above, the event logs for the smart meters will be the main source of data for this part of the project these have been explained in this chapter on section 3.1. Knowing this is a key part of historic incident detection because it affects the data source we are concerned with and so the filters used in the detection of elements and meters.

To detect which underground cable pits had suffered hotspot events last year, the same OMS database must be accessed as before. The incidents are filtered by those that were **unplanned outages** or **OA** (Other Attentions) related to **supply nodes**, **electrical boxes**, **low voltage lines**, **medium voltage** or **several affected boxes**. The incidents that passed both filters were extracted and the output was a list of all incidents with its unique code, a description of the incident and a date time. The description column was cleaned and normalised to ascii small characters with no double gaps or accents, this is important as the description is typically written in field by the operator that inspects the incident and could have errors. This description list was then **filtered by those that contained** references to elements that are common consequences of hotspot events in underground cable pits such as smoke, heat, sulphated ... then a second filter was performed to those containing the desired location or element of such hotspot by finding those that also contained 'underground cable pit', underground LV line or . Finally another filter was used to discard common description keywords that appear in many incidents but not commonly in the desired ones, these are related to medium voltage. It was included in the first filter to avoid rapid discardment of incidents that could have been classified under that characteristic, but if specifically described in the description, MV is not a desired key word to find as then it would probably not be related to underground cable pits, as seen in many examples. Again, as these descriptions are written in the field, the word filtering had the complexity of using where possible word roots to avoid the effect of misspells.

This pre-extraction filter alongside the word filter reduced the size of the incident list from hundreds of thousands to about seven hundred instances. It is possible some hotspot events were lost to some filter but several filtering techniques and different word combinations were used to loose as little hotspots whilst keeping the filter strict to avoid poor labelling, as this would greatly affect the performance of the models. Next, as the filters are not perfect, a by hand filter is then performed for the last hundreds of incidents by examining the description and available information. This process could be further automated by the use of a large language model, but the use of such models is internally restricted by the company, so filtering by hand was the only option left.

Once the incidents have been filtered, another extraction is performed to obtain the most affected element related to each incident. Being most affected the element that had the longest outage time. This is an effective strategy few grid elements are disconnected during a hotspot event in an underground cable pit. Also, as hotspot events will require corrective maintenance before reconnecting the underground connection permanently meaning it should be the one with longest downtime duration. This results in a list of several lines (with the code of the

secondary substation they are connected to and the line number of that secondary substation) where hotpot events caused an unplanned outage.

This list is used to extract all the smart meters connected downstream of these lines, it is from these meters where the measurements will be sourced. The meters this section of the project is interested on are: SABT secondary substation supervisor meters 'LVS' and consumer downstream smart meters ('CN', 'T4', 'T4MI'). The final extraction is using the final meter list and the incident date time, obtain the SABT measurements for the 4 months before the incident. 4 months are considered to be enough data points to extract sufficient information for the statistical analysis and further model. All data points are aligned with their initial day being day 0 and their final day being day 120 as one day before the incident. It is done as such so that the events logged the day of the incident are not used in the preventive algorithm, as less than a day offers too little room for crews to mobilise and apply the desired maintenance.

The basis for the statistical study is the event logs mainly related to power quality events for the 4 months prior to the incident, but for the model training and creation another category was included. A control population of healthy lines that had no hotspot event were included in the study. But it was observed that healthy lines have no events at all and therefore provided no challenge in differentiating them from damaged lines. Nevertheless, there was a category of healthy lines that offered a significant challenge, these are healthy (meaning without hotspot events) that were overloaded during some period of the year, for example small coastal villages during summer where tourism multiplies the regions population by ten fold. These lines are known to have over and under voltages during these periods and so pose a significant differentiation challenge with damages (meaning with hotspot events) lines.

Therefore, as a summary, the data used contains instances of both healthy (no hotspot but constant voltage deviations due to overloading) and damaged lines and a time frame of 4 months per line. Damaged lines were detected via the OMS, and the healthy lines (used as control for a more challenging model creation) were detected by finding the periods where certain known lines were giving problems. Once more, note that a real control population would be lines with no events during the 120 day period, or just a few events if maintenance or another incident unrelated to the line occurred. But then the model would not be as specifically tailored to hotspot events, so the control population is taken from lines suffering from stronger than the norm voltage deviations.

For each line mentioned here all the downstream meters were extracted and form each a 120 day period prior to the incident was obtained. The data was first

cleaned to remove 2 faulty meters that periodically were sending more than 1000 daily events and were deemed faulty. Then for every remaining line, every day, all the events of all meters connected to this line are summed up and then averaged to the number of meters in the line. As an indicator of how damaged this line is as a whole as it is unknown the exact location of the line this underground cable pit hotspot event occurred. In total there is a list of almost 500 lines (450 of which are damaged, hotspot containing lines) with over 23,000 meters connected to those lines producing for the 4 month study around 1,800,000 filtered events to the event groups mentioned above that were of interest to this study.

Statistical analysis

The first event log analysis was a daily events plots for all lines, the incapability to detect patterns in this format incentivised a statistical study. Probability graphs were devised both for individual lines and the aggregation of lines, alongside a histogram. All in the pursuit of obtaining a path towards a possible pattern recognition that influenced model selection. These statistical analysis were once more developed in python with statistical libraries and now not only filtering for group of events but also for the specific event types that were of interest, ensuring the filter was properly applied to SABT meters (LVS) and downstream smart meters appropriately.

The analysis performed were:

- All individual lines KDE plot comparison
- Aggregated events KDE plot
- Histogram
- Individualized per line probability density plot

Economic analysis

To study the required precision of the model to achieved economic break-even a study was performed. The study uses internal data or approximations for the study of the yearly costs and benefits of implementing a large scale model for hotspot anomaly detection in underground cable pits. The costs are:

- The cost implied in visiting an underground cable pit flagged by the model as containing a hotspot, 120€ for the team displacement. A simplification is made here assuming the team has enough time to add this to the calendar and requires no prioritization over a different work made by the team. Another simplification is that this cost is fixed, but the model flags the whole line and then a further study must be performed to see the point of the line this hotspot event is, if not the crew must analyse every underground cable pit in the line until they find the damaged one, causing economic inviability.
- The cost for the maintenance required for those underground cable pits visited that had actual hotspots in them of 800€. The cost is approximated at 50% of such cost as per technical expert recommendation, some hotspot maintenance does not require an equipment change but a clean-up, minimum element change or simply just mending the loose connection.
- The cost for the model execution based on the execution times and the costs for Azure's virtual machines [56] cost comparison of family D, selecting the most appropriate computational characteristics for each model in terms of price, vCPUs and memory. Nevertheless, this cost was finally excluded form the model due to its insignificance as it amounted to less than 200€ per training execution for the final combination of models (the most time consuming part of the model but with the least frequency)

The 'benefit' comes from the avoided cost related to the consequences of a hotspot event, set at 1000€ (approximated cost of replacing an underground cable pit, also reduced to 90% as for the maintenance cost)

The combination of these items, with the estimated number of hotspot events, the number of lines and the number of annual lines with strong voltage deviations, results in a precision limit for the model so that economical feasibility can be granted. Although all direct costs and benefits are taken into account there is an indirect benefit of minimising public exposure to hotspot events that is not accounted for in this analysis due to its complexity and social bases. Also, there is a cost related to safety risks imposed by a hotspot event and its possible consequences, neither of these indirect social costs are in the scope of this project but must be taken into consideration for a full view on the economic side. Due to the high number of lines and the relative expensiveness of visiting a line compared to replacing a fully damaged one, economic unfeasibility could be possible for almost any realistic model with the actual data available. Nevertheless, these indirect social benefits regarding public image and safety risks, could significantly increase the economic benefits and should be further investigated to analyse the real economic outcome of the model.

3.3.2 Gaussian Mixture Model (GMM)

The model selection of the predictive algorithm for hotspot events was guided by the prior data and statistical analysis where a probabilistic method could be used favourably in the detection process. Although probabilistic clustering have many different flavours [64], GMM was selected as the model used for the characteristics of the data available with its unknown and varying cluster densities as well as its relative efficiency, simplicity and explicability. All required for the project in hand as long process times will complicate model adoption due to the amount of lines studied. In the State of the Art section, Chapter 2, the clustering probabilistic method of the Gaussian mixture model was explained. So next the methodology of implementation will be described.

The input data to the model, as mentioned above, is the event logs coming from the SABT system and consumer smart meters connected to the line analysed. It is only the number of events we are concerned on as the data stored for each type of event is different and not consistent, also the complexity of the lack of direct measurements will reduce the effectiveness of such measurements. If the model detects anomalous behaviour on the line, the probability of events occurring is increasing over time (the line has a deteriorating state), especially if this growth is exponential in the final section then the line will be flagged as anomalous.

The model does this flagging by following a series of steps analysing each of the line in the study independently:

- Data is extracted, cleaned of outliers (faulty equipment - exagerated amount of daily events) and dates are normalised so as to start in 0 and en in day 120 for very line
- A filtering process takes place to obtain only the groups and types of events desired
- There is a daily summation of events for the 120 days leading to the incident, if there is no event log for a certain day are provided a value of zero so as to not leave gaps in the readings and so improve model performance
- The events are averaged by the number of meters connected to the line to avoid high number of meters influencing the decision on the damage level of the line
- A loop is performed to try the 4 main covariance matrix types and up to 40 components (Gaussian normals) to obtain the best combination, that which explains the majority of the data

- The data is the processed by the selected optimum GMM
- Flag the anomalous results depending on their probability outcome from the model and study the number of required Gaussians and their characteristics. There is a difference, as will be seen in the statistical analysis result, in the last 30 days of the study. So comparisons between before the 90 day mark and after that mark are one of the key flagging methods

These are the basic steps, then several iterations where made both to data and flagging techniques, whose results will be explained Chapter 4. But the steps and methodology remains constant.

There are several limitations to this methodology, being the main one the lack of purpose measurements and sensors to detect pre-emptively this kind of hotspot anomalies, the model uses already set up data sources and leverages them to achieve the best possible outcome. Also, the lack of direct measurements makes it difficult to benchmark the model against alternative approaches. The model also assumes event frequency correlates with line degradation, which may not always hold true, as is the case for maintenance activity, noisy meters or certain underground cable pit sudden failures where the current data just isn't enough to capture any meaningful pre-emptive trend. GMM allows for simplicity, interoperability and computational efficiency but it also has some disadvantages mentioned in Chapter 2 and might not capture complex temporal dependencies or non linear patterns as a model with more feature might be capable of achieving. The 120 day window was based on statistical and physical analysis by trying different window size sensitivities but having a fixed value for all lines might not be appropriate. An important improvement to this model is starting the count (setting the 0) once a maintenance activity is carried out on the line but access to this information is not currently automatizable with the available datasets. As statistically there was a marked 30 day mark, thriving for model computational efficiency, very important point as almost 550.000 lines must be analysed via this model almost on a weekly or bi-weekly bases, the aggregation was made daily but then again this might not be the optimum for each case but rather a general optimum and more granularity could capture more complex temporal patterns. Also, while computationally efficient, the loop over covariance types and components may still become burdensome at scale, especially with thousands of lines and a general optimum covariance type could be imposed.

Chapter 4

Results and validation

4.1 Introduction

In this chapter the results for all 3 models will be presented for both hotspot event types secondary substations and underground cable pits. For each of these the statistic and economic studies results will also be analysed. Only the most significant results and graphs will be presented to prove the objective achievement and predictive maintenance algorithm functioning. Alongside some model iterations that were improved along the way to serve as reasoning behind the final model decisions taken.

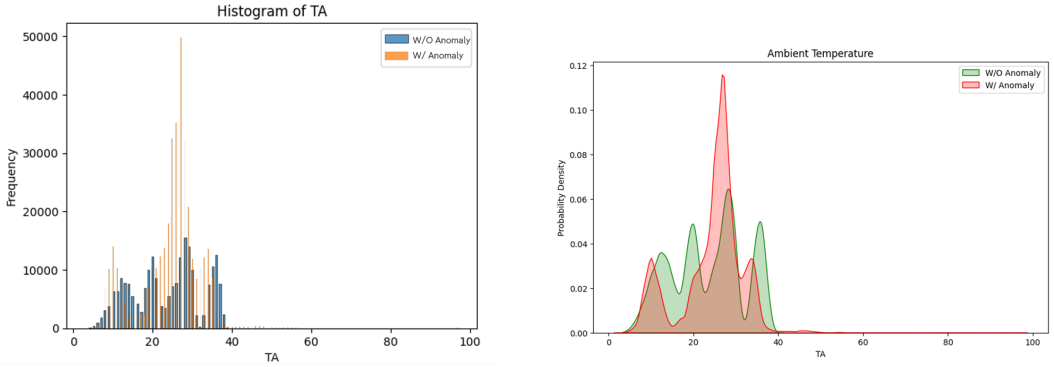
4.2 Anomalies in secondary substations

This section will now discuss the final predictive maintenance algorithm for secondary substations by analysing first the statistical and break-even economic analysis and then comparing each of the 3 final models, the analytical threshold model, the general ML regression model and the specialised ML regression model for each individual secondary substation.

The goal is to determine which model offers the best balance between predictive accuracy, interpretability, and operational feasibility. Each model will be evaluated not only on its technical performance but also on its potential for integration into the distribution company.

4.2.1 Statistical and economical analysis

The data that best should capture the increasing temperatures of the hotspot anomaly is ambient temperature, as mentioned above this ambient temperature is measured by sensors in the smart meter measuring the properties of the line and therefore, due to proximity, has certain relationship with the temperature of the wire. So, the first results from the statistical analysis is the histogram and probability density of this ambient temperature (TA) comparing the cases with and without hotspot anomalies in the secondary substation to see if there are any detectable differences between them:

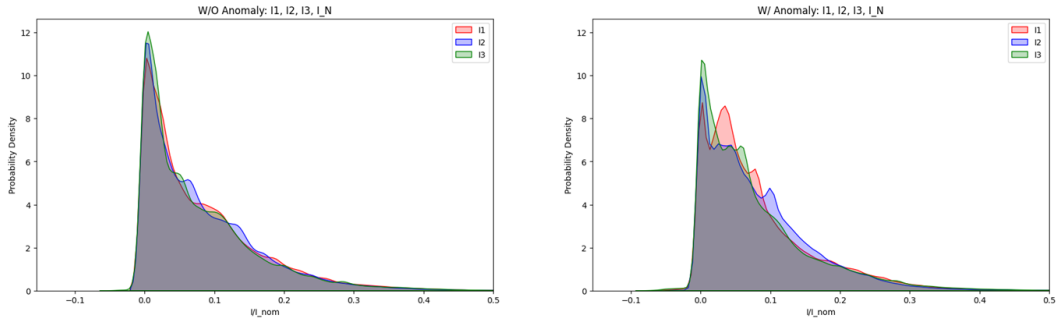


(a) Ambient temperature histogram

(b) Ambient temperature probability density

Figure 4.1: Statistical distribution of ambient temperature

As seen in the figure, there is a significant difference between substations with anomalies and substations without hotspot anomalies. First the shape of both is different, when there is no anomaly the differences between night and day and different loads can be seen but all are of a similar normalized probability. When the secondary substation has anomalies one peak is predominant and the load factors of the lines gains importance as temperatures are now generally higher. When there are anomalies there is a tail of very high temperatures, non-existent for the case when there are no anomalies, these are the actual lines that contain these hotspot events.

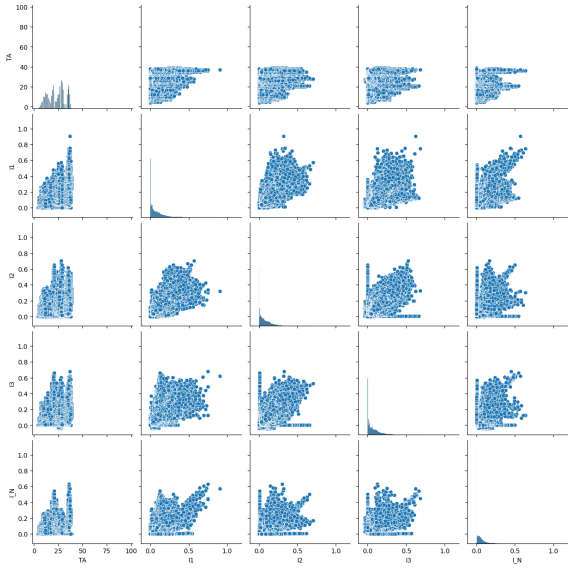


(a) 3 phase currents probability density for a healthy line

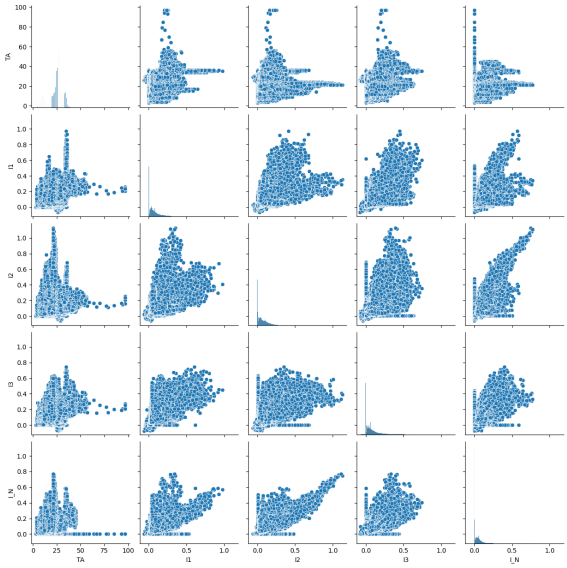
(b) 3 phase currents probability density for a damaged line

Figure 4.2: 3 phase currents probability density KDE plots

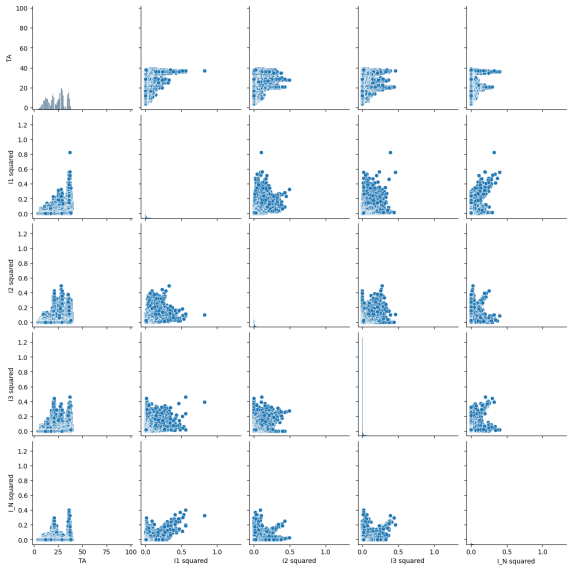
As seen, the damaged lines have higher phase unbalances and there is a clear difference in the behaviours of the phase currents, there are certain peaks where previously there was a descending monotonic curve. Wherever a monotonic curve is altered we can assume an anomaly happened [65].



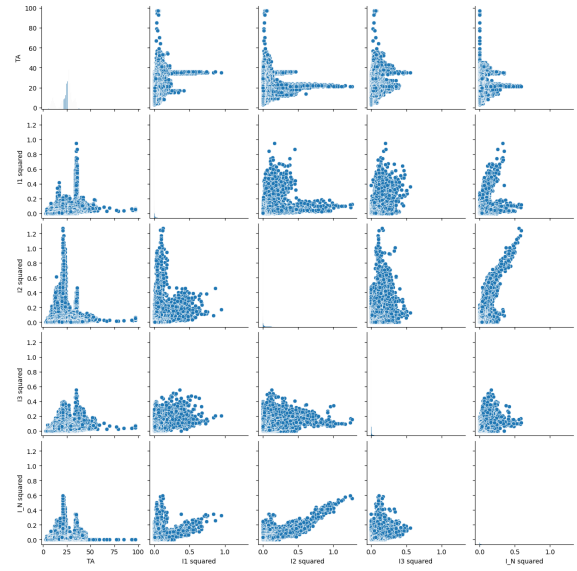
(a) healthy lines - current



(b) damaged lines - current



(c) healthy lines - current squared



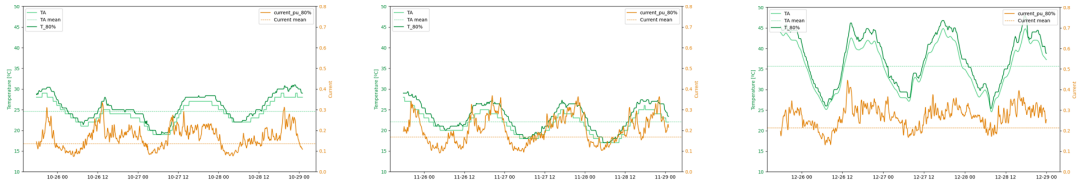
(d) damaged lines - current squared

Figure 4.3: Pair plot study for secondary substations

The pair plot analysis was made for several features for both healthy and damaged lines each analysing the relationship with the current and with the current squared. Features in order are: ambient temperature, current of phase 1, current of phase 2, current of phase 3 and current of neutral cable. As seen the current /

temperature relationship changes when an anomaly is present, where now higher temperatures can be achieved with smaller currents. It is also observed that phases are not balanced all the time as there is no direct correlation between them for neither damaged or healthy lines meaning unbalances either are unrelated to hotspot events or these events are not detectable by only looking at that feature. It is also observed that there is a higher relationship between current and ambient temperature for damaged lines, which is logic as a higher thermal degradation means more cable heat is reaching the sensor due to a poorer thermal insulation. Nevertheless, this relationship, though stronger, is not perfectly evident. With current squared these trends repeat but with a larger spread of data for damaged lines than for healthy ones, suggesting that extreme values become more common for damaged lines.

Once the feature relationships were analysed and the statistical analysis was observed, a temporal study was performed to try and find line degradation on a time axes to see if there are any significantly detectable patterns.



(a) Healthy line - October (b) Healthy line - November (c) Damaged line - December

Figure 4.4: Cable degradation with passing time

The line observed in figure 4.4 was from a secondary substation that suffered a critical failure on the first week of January due to a hotspot event. There is a clear gradual line degradation during the months prior to the incident where temperature suddenly rises the month prior to the incident although loading of the line remains almost constant during the 3 month study. The graph presents 2 temperatures, ambient temperature (TA) as measured by the sensor (this if damaged should be closely related to current) and an 80th percentile value for the calculated 3 phase cables at each point which is calculated following the formulation presented in 3.1. The value presented for the current is also the 80th percentile of the 3 phases for consistency. An in depth study found out a point where gradual degradation was no longer gradual but sudden, there was an incident which manipulated the line and accelerated the line degradation for the last 20 days before the critical failure, detecting these points in time are critical as preventive measures must act instantly once such situations are detected.

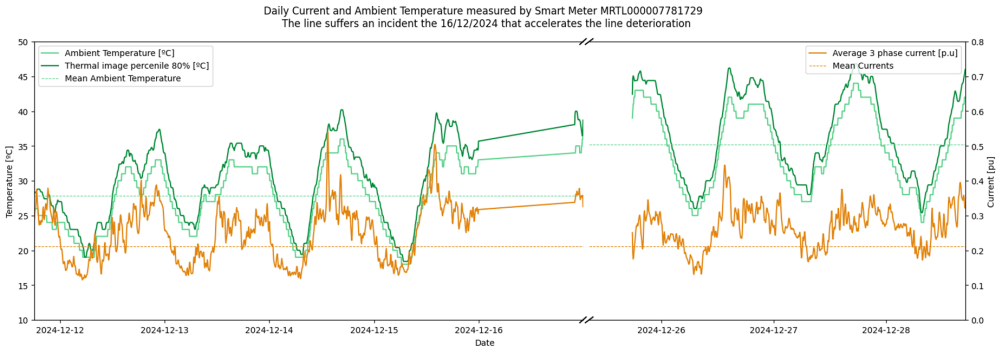


Figure 4.5: Point at which line degradation increased significantly

As seen, the damaged line has a higher average temperature for a smaller current as well as having a tighter relationship between the ambient temperature and the current explaining the thermal leakage. Another key finding is that thermal deterioration is gradual but, certain incidents significantly worsen the condition. Once this incident happens, the temperatures mean (dotted lines) increases significantly for a similar loading characteristic. The temperature leakage to the exterior of the cable is greater.

Another important analysis is how are other lines from this same substation behaving once a hotspot event evolves. Average currents and average measured ambient temperatures at each one of the 9 lines of a secondary substation were analysed:

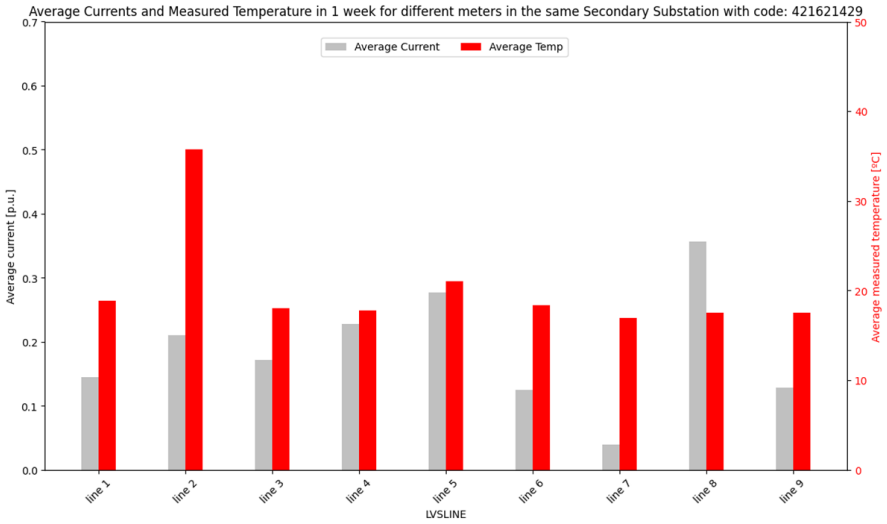


Figure 4.6: Gradient temperature between lines in a secondary substation

Whilst most lines maintain a constant ambient temperature measurement, independent of line loading, which implies this is the real average ambient temperature of the secondary substation, one line has a temperature reading almost 20ºC higher (almost the double from the others). Line 2 exhibits significantly higher ambient temperature averages despite lower current levels compared to Line 8 and other lines where ambient temperatures remain invariant of current. Possibly indicating thermal anomalies or potential damage in the line. Also proving that this 'ambient temperature' measurement is deeply affected by current once a hotspot has been developed due to the mentioned thermal leakage.

After all these findings a bases for al three models was established and feature selection had graphical and statistical backing. The remaining step was analysing the economic break-even point to obtain a theoretical model's minimum precision to obtain this.

The basis used for the economic model is that there are 100 hotspot events (as per the OMS system) in the 100,000 secondary substations every year, of which around 40% have the required SABT systems to collect the required data for the model. The cost for a crew to visit the flagged secondary substation is of 120€, for displacement costs and assuming no other task is ignored which would have its associated costs. The reapir process of a hotspot event depeds on the extent of the damage and the cause, is spans from the replacement of an electrical LV box to tightening a loose connection, therefore a cost of 1,200€ is applied (the maximum) for 30% of the cases (rendering the real averaged out cost for maintenance at 360€. The cost of corrective maintenance of a hotspot event, if not prevented, could be as large as replacing the entire damaged secondary substation or just a localised area if the transformer (the most expensive equipment) has not been damaged in the incident, therefore a cost of 120,000€ is applied for 35% of the cases. It is this last cost that, if avoided by flagging early the damaged substation, can be avoided and inferred as a cost reduction for the company, so it will be treated as a benefit. The cost inherit to model processing by virtual CPU is not included in this study as the values were generally insignificant compared to the rest of the results, nevertheless they will be included for every model when economic result of each model is presented. All the values are internally sourced from the distribution company.

The economic summary and the results are displayed here:

Category	Value
General Parameters	
Number of damaged secondary substations per year	100
Percentage of SABT in secondary substations	40.0 %
Total secondary substations	100,000
Probability of hotspot in secondary substations	0.001
Model Performance Limit	
True Positive Rate (TPR)	8.43 %
False Positive Rate (FPR)	2.92 %
True Negative Rate (TNR)	97.08 %
False Negative Rate (FNR)	91.57 %
KPI Limit (TPR/FPR)	2.89
Confusion Matrix Values Avrg.	
True Positives (TP)	3.37
False Positives (FP)	1166.77
True Negatives (TN)	38793.23
False Negatives (FN)	36.63
Model Metrics	
Precision	0.29 %
Recall	8.43 %
Operational Costs	
Cost per visit	120.00 €
Number of visits	1170.14
Total visit cost	140417.31 €
Repair Costs	
Cost per repair	1,200.00 €
Repair rate	30 %
Repairs required	3.37
Total repair cost	1,213.98 €
Incident Prevention Impact	
Cost replacement for secondary substations	120,000.00 €
Proportion of secondary substations damaged in incident	35 %
Incidents avoided	3.37
Investment reduction	141,631.28 €
Net Result	
Operational result	-0.01 €

Table 4.1: Summary of economic break-even study for secondary substations hotspot prevention model

4.2.2 Models performance

Now the results for each of the 3 individual models (analytical threshold model, general ML model, specific ML model) will be presented alongside the optimal model configuration for the best economic result. For each model a brief introduction on the model, findings and model optimization steps will be discussed.

Analytical model

This model applies expert knowledge based on the outcomes of the prior data and statistical analysis, alongside the knowledge based on the physical properties and causes of hotspot events in secondary substations.

More than 20 different physically explainable thresholds were investigated and compared. Such thresholds are intended to be statistical thresholds were possible to avoid hardcoded value thresholds that reduce scalability, nevertheless certain safety features are hardcoded to avoid infra detecting cases. Also, threshold combinations were analysed for model improvements, as certain high recall thresholds can increase precision by applying combinations of such criteria. Once a reduced selection of the highest performing thresholds was obtained, for each threshold the optimum percentage flagging was obtained. That is, out of the studied temporal window how many instances does the threshold need to be surpassed so as this to be considered abnormal behaviour and require a secondary substation flagging. This increases precision as certain criteria can be met sporadically but a repetition of such would imply, with a higher degree of accuracy, that the secondary substation is damaged and requires maintenance due to the hotspot event and as such must be flagged by the model. Depending on the threshold this value is different as high precision thresholds tend to require fewer instances for the model to flag the line whilst high recall and lower precision criteria would require a higher value of instances detected for abnormality to be considered.

A list of the 15 most significant criteria analysed is provided below:

1. Ambient temperature measurement exceeded a safety limit.
2. Ambient temperature measurement gradient inside a secondary substation, temperature difference between min and max of the lines, exceeds a 12^{o}C limit.
3. For a line, find out if its difference between measured ambient temperature and average secondary substation ambient temperatures over the average

current of the line squared (thermal relationship [66]) is in the higher quartile for the lines of the secondary substation.

4. High deviation between measured ambient temperature and average temperature for the substation for that time period.
5. Temperature slope compared to current squared slope over a 30-minute interval shows abnormal thermal evolution.
6. Temperature is above the 80th percentile for the substation, while at least two of the three phases are not above their respective 80th percentile of current.
7. At least two phases are above the 80th percentile of current, while temperature is not above its 80th percentile.
8. Z-score of temperature is higher than the Z-score of at least two of the three current phases.
9. Thermal variability (standard deviation) within the substation is significantly higher than the median variability across substations.
10. Current variability (standard deviation) within the substation is significantly higher than the median variability across substations.
11. Temperature-to-current ratio is in the upper quartile across all lines.
12. Phase imbalance: the difference between the highest and lowest phase current exceeds 10% of the average current.
13. Sudden temperature increase of at least 6°C within a 10-minute interval.
14. Ratio between the slope of ambient temperature and the slope of thermal image (80th percentile) is close to 1, indicating consistent thermal behavior.
15. Mean squared error between ambient temperature and thermal image (80th percentile) is below a threshold, indicating high agreement.

Next a table including the final models thresholds alongside their performance and optimal minimum required instances are displayed (if not provided it means that the criteria by itself has no optimal threshold and should be ideally used in a combination of criteria), also the threshold combinations used for the final model are included:

Table 4.2: Performance Metrics for Toggle Criteria

Criteria	Description	Precision	Recall	Optimal Threshold
1	Temperature exceeds a limit	100.0%	1.3%	0%
2	Temperature gradient inside a secondary substation	98.0%	26.5%	3.0%
3	Relationship between temperature and current squared	42.3%	1.13%	–
4	Deviation from mean temperature in that secondary substation	66.9%	9.6%	37.0%
10	High current standard deviation in that secondary substation	74.6%	58.3%	–
12	Phase unbalances	63.4%	87.3%	–
14	Close relationship between ambient temperature and thermal image	21.4%	8.42%	22%
A	Combination: Crit3 & Crit4 & Crit10 & Crit12	58.1%	0.12%	1%
B	Combination: Crit4 & Crit10	51.0%	16.0%	26%

An extensive search of combinations for the available data and test set, proved that the optimum model configuration is the and combination between criteria 2 & 14 & A & B each with the respective optimal thresholds. This final combination of physically explainable criteria accounts for the following model and economic results:

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	24	4	28
	n′	25	200	225
	total	49	204	253

Table 4.3: Confusion Matrix

Metric	**Value**
True Positives (TP)	24
False Positives (FP)	25
True Negatives (TN)	200
False Negatives (FN)	4
True Positive Rate (TPR)	85.71%
False Positive Rate (FPR)	11.11%
True Negative Rate (TNR)	88.88%
False Negative Rate (FNR)	14.29%
Precision	48.98%
Recall (Sensitivity)	85.71%
Specificity	88.88%
Accuracy	88.53%
F1 Score	62.34%

Table 4.4: Model Evaluation Summary

The result show economic feasibility of the model with the key KPI (TPR/FPR) above the economic break-even limit calculated in the previous section of 2.89. It is this value that affect most the economic result of the model so it will be this KPI with which each model will be compared. The computational cost of implementing

KPI	Result
True Positive Rate	85.71%
False Positive Rate	11.11%
TPR / FPR	7.71
Economic result	891,000 € / yr

Table 4.5: Economic results of the analytical model

the analytical threshold model in virtual CPUs from azure, taking the execution times from this project and extrapolating computation times for the rest of the secondary substations, would be the following:

- 0.47 processing minutes per secondary substation
- 313 total processing hours
- For an Azure Standard_D96ls_v5 VCPU the cost is 7.7571€/hr and has 96 vCPUs and 192 GiB (enough for the required task)
- for a total processing cost of 25.32€ per analysis (assumed negligible compared to the other costs and benefits)

ML General Model

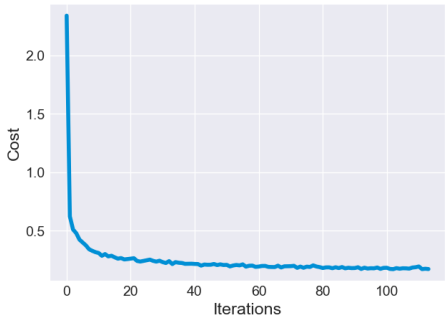
The model applies a neural network to predict ambient temperature with the different features explained in section 3.2.3, then the prediction error is analysed and if the 3 training sigma value is surpassed for 3 consecutive instances then the secondary substation is flagged. This model though less targeted to each secondary substation has a less computationally expensive training. The most effective models tested are included in the table below, as extensive search was required to obtain the appropriate model for the data available a trade off of 10% of the data was used for training, as many architectures were tested.

The simplest architecture, for computational efficiency, which obtains the best results was selected. TRP being a key indicator both for the economic benefit and the social one, undetected hotspot events must be minimised where possible. Therefore, the optimal neural network architecture obtained, form extensive searching of diverse architecture combinations, is a hidden layer size of 20 by 20 and a learning rate of 0.01.

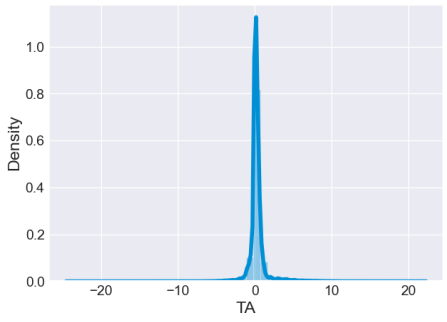
Once the model architecture is selected, the model it is trained with 20% of the data available, next some of the model training and test results are provided:

Table 4.6: Tested Model Architectures and Performance

Learning Rate	Hidden Layers	Training Data	TPR
0.01	(5, 5)	10%	0.86
0.01	(5, 10, 5)	10%	0.86
0.001	(10, 10)	10%	0.82
0.1	(10, 10)	10%	0.82
0.01	(10, 10)	10%	0.82
0.01	(20, 20)	10%	0.89
0.01	(20, 20, 20)	10%	0.89
0.01	(40, 20, 20, 40)	10%	0.86
0.01	(20, 15, 10, 5)	10%	0.79
0.001	(128, 64, 32, 16)	10%	0.89

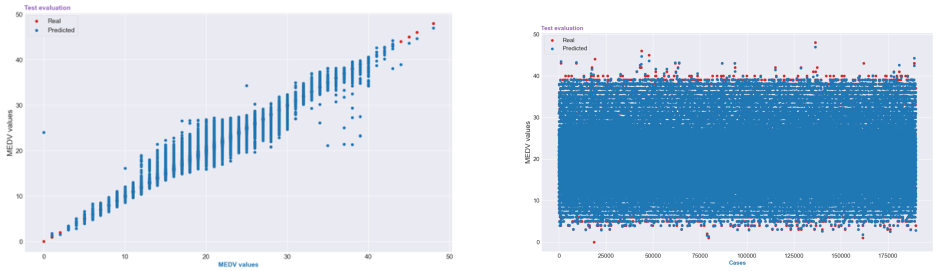


(a) Lost curve for the general ML model



(b) Error in testing for ML model

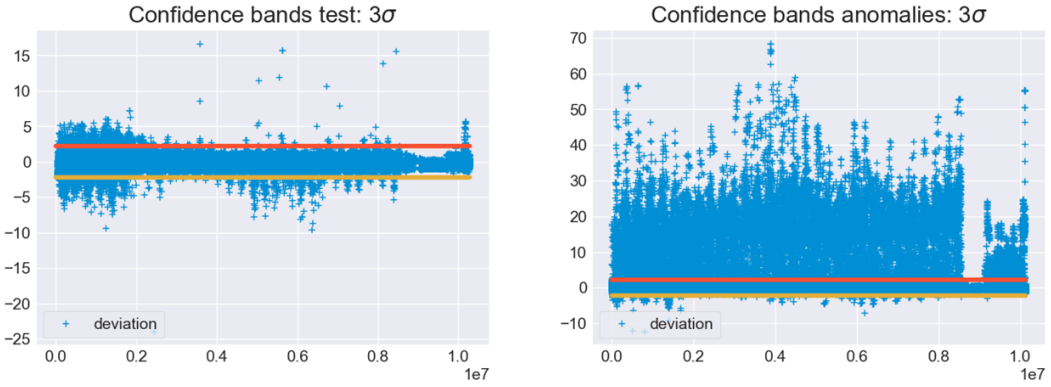
Figure 4.7: General ML model lost curve and error in testing



(a) General ML model train real vs predicted values (y axes)

(b) General ML model train real vs predicted values (x axes)

Figure 4.8: General ML model train real vs predicted values



(a) General ML model test error compared with 3 sigma limits

(b) General ML model hotspot containing substation error compared with 3 sigma limits

Figure 4.9: Model error for training with healthy lines and validating with damaged lines

As seen the difference in errors between healthy lines and damaged secondary substations that contained hotspot anomalies is significant. The model flags secondary substations that exceeds the upper limit for 3 consecutive measures. When tested over the 128 secondary substations explained in the Methodology section in Chapter 3 the model results and economic results are:

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	26	2	28
	n′	64	161	106
	total	63	71	253

Table 4.7: Confusion Matrix

Metric	Value
True Positives (TP)	26
False Positives (FP)	64
True Negatives (TN)	161
False Negatives (FN)	2
True Positive Rate (TPR)	92.86%
False Positive Rate (FPR)	28.44%
True Negative Rate (TNR)	71.55%
False Negative Rate (FNR)	7.14%
Precision	28.89%
Recall (Sensitivity)	92.86%
Specificity	71.55%
Accuracy	73.91%
F1 Score	43.99%

Table 4.8: Model Evaluation Summary

KPI	Result
True Positive Rate	92.86%
False Positive Rate	28.44%
TPR / FPR	3.26
Economic result	178,000 € / yr

Table 4.9: Economic results of the general ML model

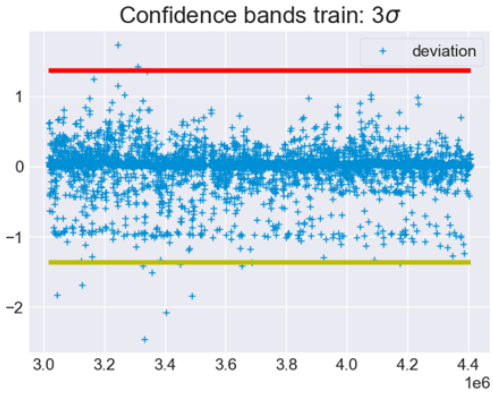
As seen the ML captures more complex patterns and is capable of improving on True Positive Rate, key to identify as many secondary substation incidents as possible. Nevertheless, this increase comes at a great FPR worsening. Resulting in almost half the ratio result than for analytical threshold model results. Which results in a lower economic result, although the result is still significantly viable at about 200,000€ per year, once more with no social and marketing benefits associated to a reduced number of hotspot incidents. The computational cost of implementing the analytical threshold model in virtual CPUs from azure, taking the execution times from this project and extrapolating computation times for the rest of the secondary substations, would be the following:

- 0.02 processing minutes per secondary substation
- 12 total processing hours
- For an Azure Standard_D4as_v5 VCPU the cost is 0.1375€/hr and has 4 vCPUs and 16 GiB (enough for the required task)
- for a total processing cost of 1.02€ per training (assumed negligible compared to the other costs and benefits)

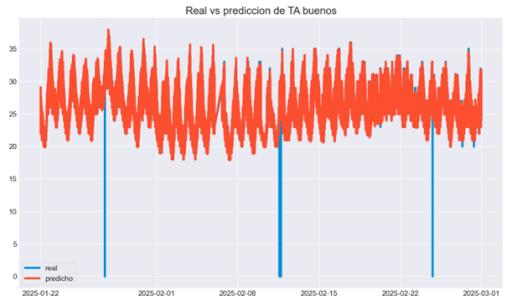
ML Specific Model

The specific ML model follows the same principle as the general ML model but it is only trained and tested with the data from one secondary substation and is used to detect anomalies in that secondary substation. It is a more complete model and can find the optimal architecture out from a list of possible architectures for each secondary substation. But, training is much more computationally expensive as an individualized model for each secondary substation must be trained and stored. As each model has its own optimal architecture and is selected automatically, here the result comparison of one of the models will be presented as a guide for the general results but the actual model performance will be displayed at the end.

First an example of a healthy line will be presented where the model predicts the ambient temperature for the line perfectly with very small error rates. Then a complex case study will be presented, where a model trained for when the secondary substation was healthy, gets damaged and a periodic maintenance detected the hotspot event and corrected it. Showing how when the line was damaged prediction was poor and errors were large but once the line was corrected and returned to a healthy state the model predicts accurately once more.

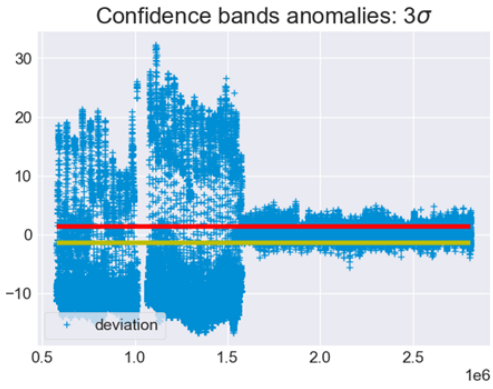


(a) Specific ML model train error compared with 3 sigma limits

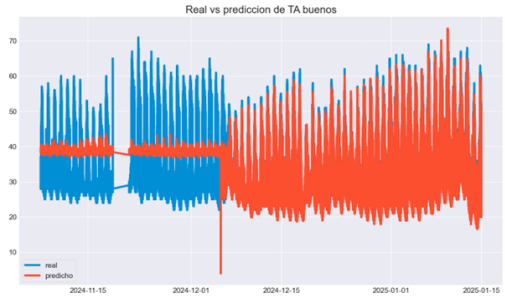


(b) Specific ML model ambient temperature prediction vs real measured value

Figure 4.10: Specific ML model predicts accurately for a healthy line



(a) Specific ML model test error compared with 3 sigma limits for a repair case



(b) Specific ML model ambient temperature prediction for a repair case

Figure 4.11: Model performance for a complex case where a damaged line gets repaired due to a preventive maintenance

As seen in the above cases this model, even for complex cases involving a secondary substation change of state, performs adequately. This is because now each secondary substation is trained for a healthy state of each secondary substation and is now detecting alterations of the secondary substation form this normal / healthy behaviour. Consideration must be taken to the complexity of training a ML model for each individual secondary substation, this computationally demanding step is required to obtain these results and is the most complex of the 3 developed models, therefore a trade off must take place to decide between the

three all factors considered.

In terms of model performance and economic results, a summary is now presented:

		Prediction outcome		
		p	**n**	**total**
Actual value	**p′**	18	10	28
	n′	31	194	225
	total	49	204	253

Table 4.10: Confusion Matrix

Metric	**Value**
True Positives (TP)	18
False Positives (FP)	31
True Negatives (TN)	194
False Negatives (FN)	10
True Positive Rate (TPR)	64.29%
False Positive Rate (FPR)	13.78%
True Negative Rate (TNR)	86.22%
False Negative Rate (FNR)	35.71%
Precision	36.73%
Recall (Sensitivity)	64.29%
Specificity	86.22%
Accuracy	83.79%
F1 Score	46.75%

Table 4.11: Model Evaluation Summary

KPI	Result
True Positive Rate	64.29%
False Positive Rate	13.78%
TPR / FPR	4.7
Economic result	407,000 € / yr

Table 4.12: Economic results of the specific ML model

The specific model outperforms economically the general model, there is less cases of hotspot detected but is more specific and reduces the false positives significantly compared to the previous model. This comes at an increased computational complexity. The computational cost of implementing the specific ML model in virtual CPUs from azure, taking the execution times from this project and extrapolating computation times for the rest of the secondary substations, would be the following:

- 0.176 processing minutes per secondary substation
- 117 total processing hours
- For an Azure Standard_D96ls_v5 VCPU the cost is 7.7571€/hr and has 96 vCPUs and 192 GiB (enough for the required task)
- for a total processing cost of 9.48€ per round of training, meaning training required for all secondary substations (assumed negligible compared to the other costs and benefits)

Optimum Model

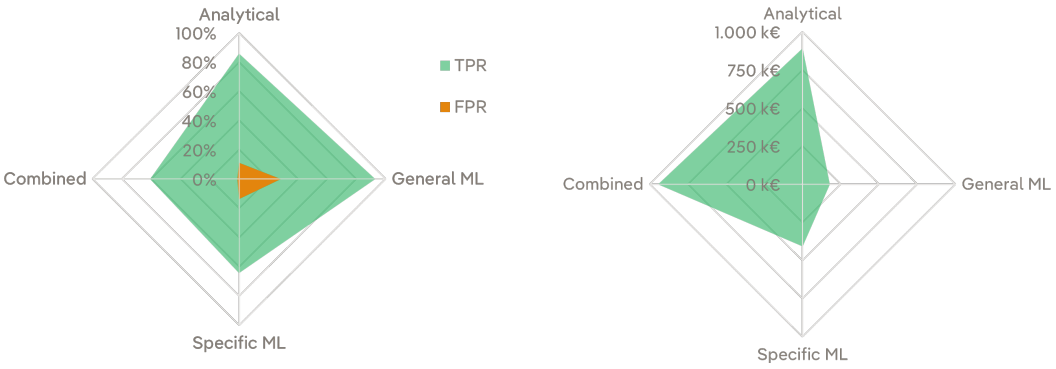
The economic optimum, once more state that this economic model does not take into account social cost of secondary substation unexpected downtime due to hotspot events and the safety risks concerned to such events, is a combination of the past 3 models. Due to the relative expensiveness of secondary substation visits the optimum economically is to combine all 3 models with ands, if all models are flagging a secondary substation then send the maintenance crew to fix the predicted hotspot event. This is, however, not the model with the highest prediction rate, several hotspot events will remain undetected, but the number of false positives is minimised and therefore the economic optimum is reached at this trade off.

KPI	Result
True Positive Rate	60.71%
False Positive Rate	1.33%
TPR / FPR	45.5
Economic result	944,000 € / yr

Table 4.13: Economic results of the combined model

As seen the benefits are maximised due to the significant reduction in the false positive rate, here the models flaggings are the most accurate but several hotspot events would remain unpredicted. In terms of computational costs:

- 0.8 processing minutes per secondary substation
- 533 total processing hours
- For an Azure Standard_D96ls_v5 VCPU the cost is 7.7571€/hr and has 96 vCPUs and 192 GiB (enough for the required task)
- for a total processing cost of 43.10€ for training (the highest of the prior but still assumed negligible compared to the other costs and benefits)



(a) Main KPIs summary

(b) Economic summary

Figure 4.12: KPI and economic summary for: analytical model, general ML model, specific ML model, combination of models

As seen the combination of models offers the best economic result, but the analytical model is of the same order of magnitude but has a significantly higher TRP, meaning more hotspot events will be detected. Therefore a further study must be made to internalise the social costs of allowing hotspot events to damage secondary substations and then a better decision for which model to select can be made.

4.3 Anomalies in underground cable pits

The objective of this part of the project is to create a model capable of preemptively detecting hotspot events in underground cable pits. This section will now discuss the model created for the case of predictive maintenance for hotspot events in underground cable pits. First a statistical analysis of the available data will be displayed to demonstrate the decisions taken, stated in section 3.3. Then an economic analysis will be performed alongside its limitations as model evaluator for this specific case.

4.3.1 Statistical and economical analysis

Once more, to understand the data selection as feature for the model, the complexity of this task must be stated. Underground cable pits have no direct measurements, the accessible data comes either from the SABT system upstream from the underground cable pit or the user smart meters connected downstream. As mentioned in Chapter 2, the causes and consequences of a hotspot event are mainly voltage deviations and possibly fault currents alongside temperature increases. Temperature gets discarded as viable measurement as there is no direct measure, and no indirect measure that can be extrapolated to underground cable pits. The lack of direct measures also complicates the specific capture of patterns related to hotspot event in the other 2 cases. This is why, cable health in this section will be measured by the number of power quality events the smart meters and SABT system are logging daily for each specific line. Meaning that a healthy line should have almost no events whilst a problematic or damaged line should have plenty events.

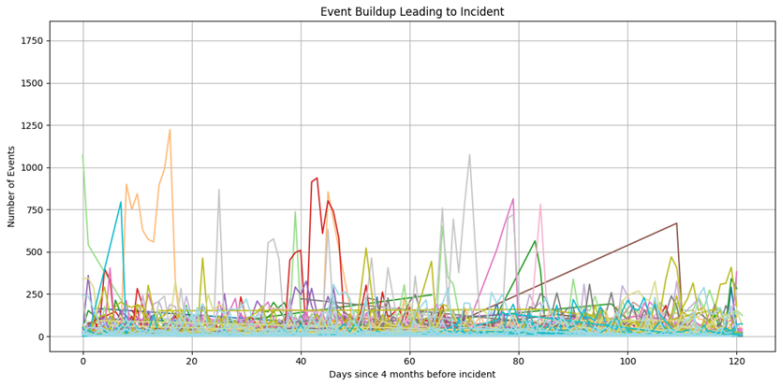


Figure 4.13: Number of events per line in the time leading an incident

As seen in figure 4.13, no pattern can be recognised as it could be done with the secondary substation case. Therefore no regression tool would be a meaningful tool for this instance and an analytical study would require of more information sources. Therefore a statistical analysis was performed to see if event probability increased as the date of the historic incident found was reached (day 120, the time window was selected after several statistical analysis as it was the one which could explain the most cases).

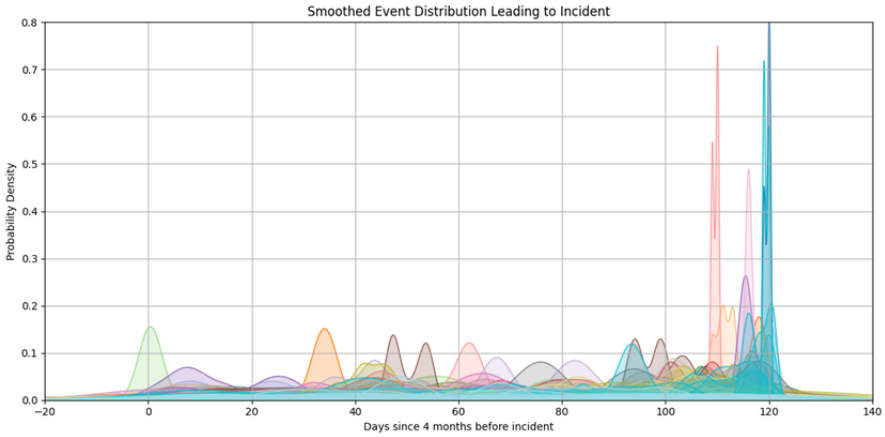


Figure 4.14: Probability distribution of number of events per day

After some prior data preprocessing to remove faulty smart meters, figure 4.14 now shows signs of a detectable pattern as the day of the incident closes in, where figure 4.13 did not. Around the 90 day mark (30 days prior to the incident) in several lines there is a significant increase of the probability of an event taking place. Stating the need for a statistical or probability tool to detect the patterns viewed. A further analysis was made with all of the lines to see if the supposed patter can be extrapolated to all other lines or just a few so a general combined event probability distribution and total histogram was analysed.

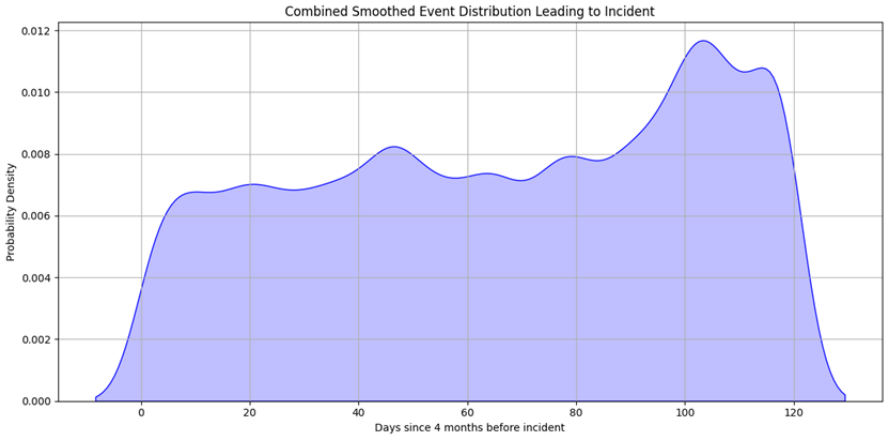


Figure 4.15: Combined probability distribution shows a constant value up to the 90 day mark prior to the incident where probability increases

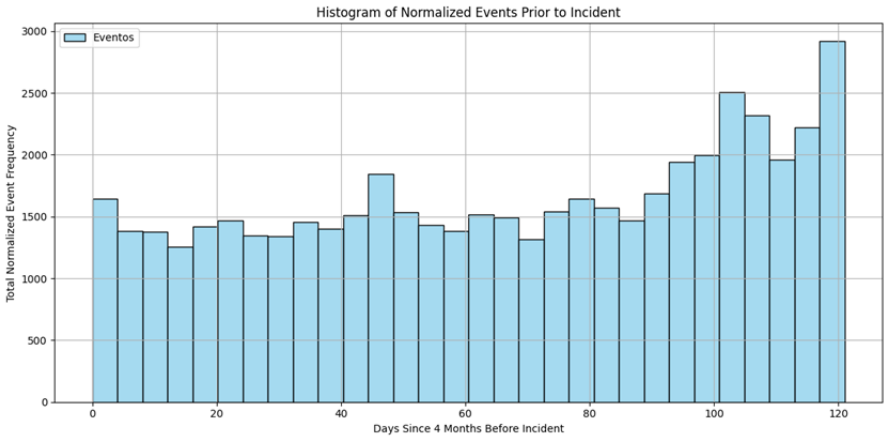


Figure 4.16: Histogram show a less smoothed version where the peaks after the 90 day mark and the constant distribution prior to it are even more clear

In the statistical analysis a precedent for a predictive algorithm is detected (which due to the complexity of the problem and lack of available data was not granted) and the type of model required is obtained. After this, an economic analysis was performed. Nevertheless, given the low CAPEX characteristics of underground cable pits and the high cost associated with visiting false positives relative to the modest corrective maintenance expenses, economic viability depends on a high accuracy of the predictive model, accuracy which is hindered by the limited availability of direct data for this specific case.

The economic summary and the results are displayed here:

Category	Value
General Parameters	
Number of damaged underground cable pits per year	500
Total UCPs	500,000
Probability of hotspot in secondary substations	0.001
Model Performance Limit	
True Positive Rate (TPR)	65.70 %
False Positive Rate (FPR)	0.21 %
True Negative Rate (TNR)	99.71 %
False Negative Rate (FNR)	35.30 %
KPI Limit (TPR/FPR)	315.47
Confusion Matrix Values Avrg.	
True Positives (TP)	307.33
False Positives (FP)	973.22
True Negatives (TN)	473551.78
False Negatives (FN)	167.67
Model Metrics	
Precision	24 %
Recall	64.7 %
Operational Costs	
Cost per visit	120.00 €
Number of visits	1280.55
Total visit cost	153,666.20 €
Repair Costs	
Cost per repair	800.00 €
Repair rate	50 %
Repairs required	307,3
Total repair cost	122,932.95 €
Incident Prevention Impact	
Cost replacement for secondary substations	1,000.00 €
Proportion of secondary substations damaged in incident	90 %
Incidents avoided	307.33
Investment reduction	276,599.14 €
Net Result	
Operational result	-0.01 €

Table 4.14: Summary of economic break-even study for underground cable pit hotspot prevention model

As seen, the required precision and KPI of TPR/FPR is significantly higher than for the secondary substation case. If distribution regulated remuneration was TOTEX or OPEX based even with a budget deficit this could be taken care off in exchange for the increased quality of service. As it is not the case yet, further analysis of the social costs involved are required to ensure true economic feasibility. As a preliminary analysis, expert knowledge from inside the distribution company feel social benefits of avoiding such hazardous incidents will be beneficial even at a direct economic loss.

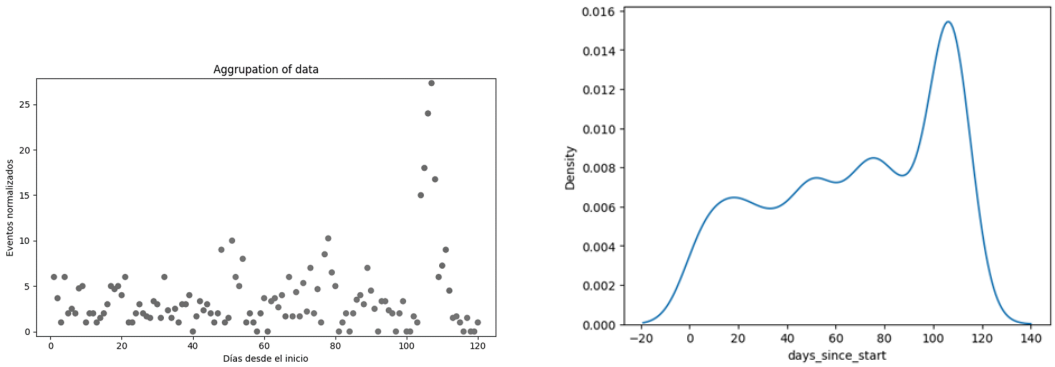
4.3.2 Model performance

Now the results for the underground cable pit predictive maintenance algorithm will be presented alongside the intermediate results that are used as a bases for the decisions made towards the final model.

GMM model

As seen in the statistical analysis a probabilistic model is best suited to capture the patterns detected. GMM represents data as a combination of multiple Gaussian distributions that maximise the probability of generating the known data.

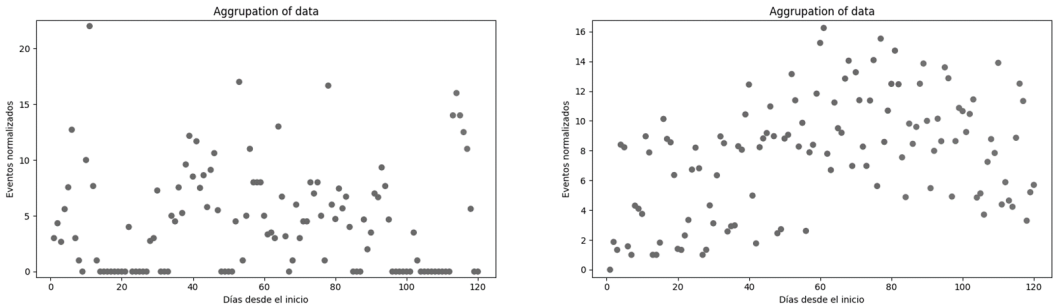
For the generation of the first iteration of the model 2 key aspects were analysed, first how was data imputed into the GMM and then which cases were of a higher difficulty for the predictor and would generate more false positives, which as observed in the economic analysis should be minimised to the minimum if economic feasibility is to be aimed. So first a single line, that resulted in a catastrophic failure on day 120 was analysed to see the most exaggerated pattern:



(a) Event log data points per day (b) probability distribution of events

Figure 4.17: Analysis of one single damaged line available data

Normal healthy lines tend to have no events during the day and are easy to differentiate (all 120 days have 0 events), that is why the model has been made to differentiate between damaged lines and overloaded lines which are the lines that are more difficult to classify between damaged and healthy. All FPR from now on are referenced to such overloaded lines which account for about 43,000 lines of the about 550,000 total lines (so around 8% of the lines). From all the overloaded lines, the ones that were only overloaded for a specified period, eg small coastal town in summer where tourism increases significantly cable loadings, were selected as they were deemed the most complex cases out of all the overloaded lines for model differentiation. The study was made as such to test the models limits, minimise FPR and get as close as possible to the economic break-even point. Here is a comparison between a damaged line and a overloaded one to show the complexity in differentiating them both:



(a) Event log data points per day for a damaged line

(b) Event log data points per day for an overloaded line

Figure 4.18: Comparison of event logs for a damaged line and a seasonally overloaded one

Around 60 different models and adaptations of the basic initial GMM were made to capture new features that previous ones did not or to try to solve the lack of data. The main steps will be presented as justifications of the final model.

The first step is to determine the number of components and covariance type selected for each GMM for each line. For the selection of the optimal number of components, a critical task that directly impacts model generalization and computational efficiency, the Bayesian Information Criterion (BIC) is employed as a penalized likelihood function that balances model fit against complexity. Formally, the BIC is defined as:

$$\text{BIC} = -2 \cdot \log(\hat{L}) + k \cdot \log(n),$$

where $\hat{L}$ denotes the maximized likelihood of the model, k represents the number of free parameters, and n is the number of observations. The logarithmic penalty term $k \cdot \log(n)$ introduces a complexity cost that discourages over-fitting by penalizing excessive parametrization. From an engineering standpoint, this criterion ensures parsimonious model design, aligning with principles of resource-efficient computation and robust statistical inference. Economically, BIC supports asymptotic consistency in model selection, favouring configurations that optimize predictive performance while minimizing unnecessary structural overhead.

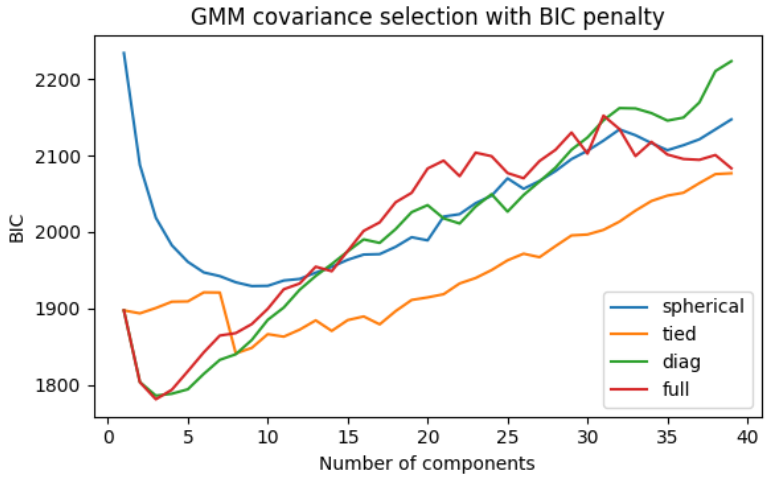
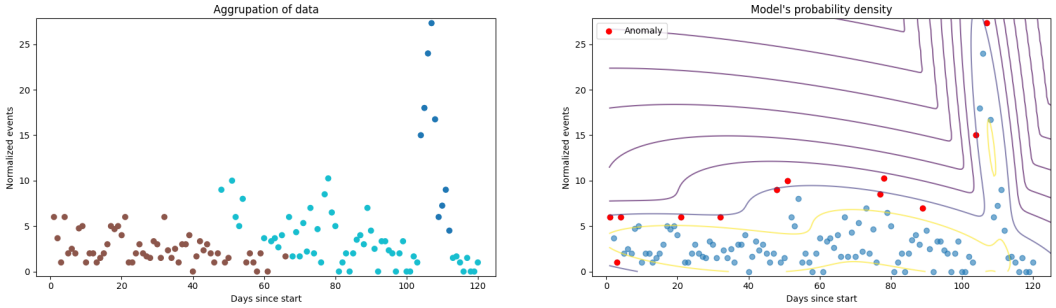


Figure 4.19: BIC penalised function to select the optimal number of components and covariance type

Then the data is aggregated into the selected minimum number of Gaussian normals and the 10% less likely points of the data are further analysed, in the fugure these are marked in red.

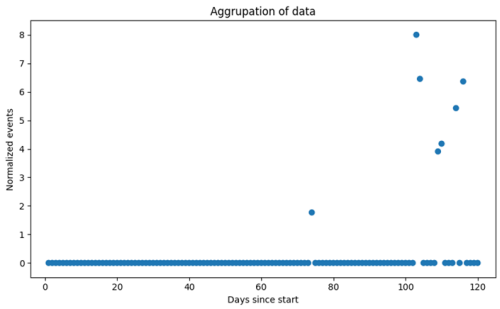


(a) Data aggregation into Gaussians

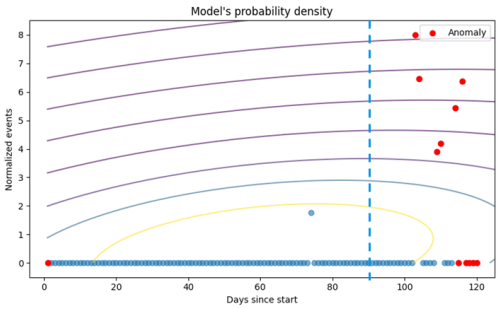
(b) Minimum number of Gaussians that explain the data points and anomaly flagging

Figure 4.20: Data aggregation into Gaussians and anomaly flagging

The first complete model (model 3), used the least probable points (furthest from the normal) to label as anomalies and then comparing anomalies before and after the 90-day mark. If the probability of these events was twice as high after the 90 day mark than before then the line was flagged as damaged. With this model the outcome obtained was a TPR of 28% and a FPR of 4%



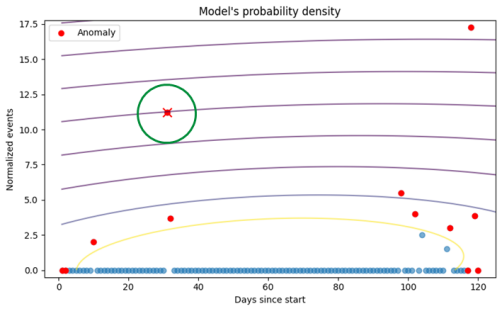
(a) Data spread example of first model



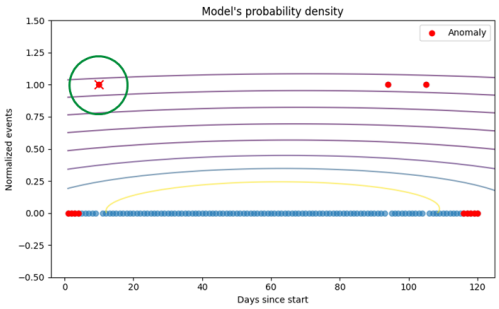
(b) Anomaly detection before and after the 90 day mark

Figure 4.21: First GMM model analysed, anomaly rate comparison

The next model (model 13) tries to eliminate errors sue to past incidents of the same line. For example if the line is disconnected for maintenance purposes, the ideal scenario would be to ignore those das and start the 120 day count from there, but as access to these logs is not available, an alternative is propose to eliminate these rare events. Using only the 90 percentile more probable points for the 'before' analysis, removes skews due to past field operations. This change improved the model performance to a TPR of 31% whilst remaining the FPR at 4%.



(a) Example of elimination of a problematic eventful day prior to the 90 day mark due to past maintenance

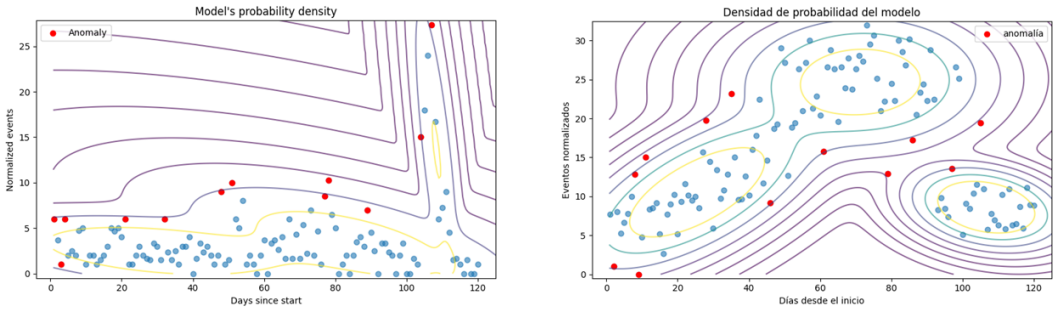


(b) Example of elimination of a problematic eventful day prior to the 90 day mark due to past maintenance

Figure 4.22: Second GMM model analysed, anomaly rate comparison when reducing effect from past incidents or maintenance events

For extreme cases the proposed model do not work, this is due to these cases having lots of events on the last few days. This great increase of anomalous points causes the model to adjust a special new Gaussian for this data and therefore are no longer considered anomalous as they can be explained by this new and

individual Gaussian. So the next proposed model (model 55) detects this increase in Gaussian number when modelling only the first 90 days and then extending to the full 120 days. If a new Gaussian is required to explain the last days then the line is flagged.

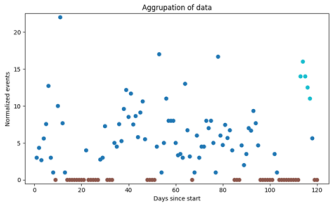


(a) Example of a case where a new Gaussian is needed to explain the last days of a damaged line

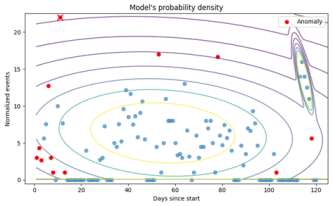
(b) Example of a case of an overloaded line which also requires a new Gaussian to explain the last days

Figure 4.23: Third GMM model analysed, detecting new Gaussian required for final data explanation

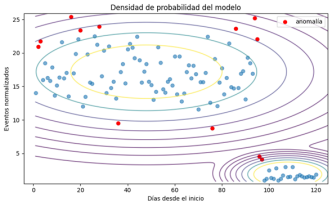
This nevertheless, poses a problem for FPR as it does not take into account if this new Gaussian has more events than the prior Gaussians. Sometimes the behaviour of the line changes but it is not a negative change, for example corrective maintenance, loading change ... It is important to distinguish between normal condition changes and condition changes where the line degrades into a damaged line whcih contains a hotspot event. Therefore the next model proposed (model 64) also measures the average mean number of events of the Gaussians prior to the 90 day mark and then, if a new Gaussian is required to explain the last few days, compares it with the mean number of events of this new Gaussian. If there is an increase in number of Gaussians and an increase in its average number of events then the line is flagged as damaged. This increased TPR to 39% and FPR was kept at 4%.



(a) Data Gaussian aggregation for damaged line



(b) Gaussian probability spread for the same damaged line



(c) Example of healthy line that changes condition due to maintenance and would not be flagged

Figure 4.24: Fourth GMM model analysed, detecting new Gaussian required for final data explanation and comparing the mean number of events

The final model developed that combines the advantages of each of the models is presented in the summary table below. It is a combination of the models explained above and, with the available data, was the best performing model capable to be obtained as it captures several different possible damaged line behaviours.

Table 4.15: Performance metrics of hotspot anomaly prediction GMM models for underground cable pits

Model Nº	**Model Description**	**TPR**	**FPR**
13	Rate of anomaly after $>$ 2 $\times$ rate before (using p90 probability for 'before')	31%	4%
55	Number of normal instances after $>$ number of normal instances before	73%	88%
64	Mean of normal instances after $>$ 2 $\times$ mean of normal instances before	39%	4%
68	Combination of models 55 and 64	36%	4%
100	Logical OR between models 13 and 68	46%	8%

The model is capable of detecting almost 50% of the incidents with less than 10% of false positives, therefore a predictive maintenance algorithm for underground cable pit hotspot events has been developed. Nevertheless, the key economic KPI is:

- TPR = 46%
- FPR = 8% * 43000/550000 + 0.2% = 0.83%

Resulting in a KPI of 55.73, which is smaller to the required KPI of 315 for economic break even. So the model would be 350,000€ in the red. This again without taking into account the social and safety related costs of such incidents developing in underground cable pits in urban areas, which as stated by the questioned experts are assumed larger than this value.

As for the computational costs, just for completion:

- 0.24 processing minutes per secondary substation
- 1800 total processing hours
- For an Azure Standard_D96ls_v5 VCPU the cost is 7.7571€/hr and has 96 vCPUs and 192 GiB (enough for the required task)
- for a total processing cost of 145.45€ for training, assumed negligible compared to the other costs

Chapter 5

Conclusion

This chapter summarizes the main findings of the thesis, reflects on the limitations encountered, and proposes some future research. It begins by revisiting the aim of the work and evaluating the extent to which the objectives have been achieved.

Conclusion on topic of the work

The primary aim of this thesis was to develop a predictive maintenance algorithm for detecting hotspot anomalies in low-voltage (LV) networks, specifically in secondary substations and underground cable pits. This objective was pursued through a hybrid approach combining analytical threshold-based models and machine learning (ML) techniques for secondary substations and GM model for underground cable pits. This was achieved leveraging data from both SABT systems and smart meters, alongside data form the other 2 main databases the GIS and the OMS for historic incident detection and data labelling.

The work successfully demonstrated that predictive maintenance for hotspot anomalies is not only technically feasible, both cases, but also economically viable, in secondary substations, when appropriate models are deployed. The thesis contributes to the ongoing digital transformation of Distribution System Operators (DSOs), aligning with regulatory trends and strategic goals such as increased grid reliability, reduced operational costs, and enhanced safety.

Summary of findings

The thesis addressed four core objectives:

1. Identification of historic incidents: A robust filtering and labeling methodology was developed to extract and classify hotspot-related incidents from OMS databases. This enabled the creation of a reliable dataset for model training and validation.
2. Development of predictive algorithms: Three models were developed for secondary substations: an analytical threshold model, a general ML model, and a specific ML model. Each was evaluated for accuracy, interpretability, and economic feasibility. The analytical model achieved the highest True Positive Rate (TPR) (85.71%), while the combined model offered the best economic performance 944,000€ / year.
3. Validation with field data: All models were validated against real-world incidents and field measurements. The specific ML model demonstrated strong performance even in complex scenarios involving maintenance interventions, but the combined model still offered the best relationship between TPR and FPR.
4. Extension to underground cable pits: A Gaussian Mixture Model (GMM) was developed to detect anomalies using event logs from smart meters. Despite the lack of direct measurements, the model achieved a TPR of 46% with an FPR of 0.83%, proving the feasibility of predictive maintenance in these challenging environments.

These findings emphasize the value of SABT and AMI data in enhancing grid observability and enabling proactive maintenance strategies. The thesis also provides a detailed economic analysis, showing that predictive models can yield significant cost savings and operational benefits, especially if social and safety-related impacts are considered.

Contributions to the field

This thesis makes a substantive contribution to the field of smart grid engineering and data-driven asset management. It introduces a novel hybrid methodology for hotspot anomaly detection in low-voltage networks, combining physically interpretable analytical models with machine learning techniques tailored to the operational realities of DSOs. By leveraging existing SABT and AMI infrastructure,

the work demonstrates that predictive maintenance is not only technically viable but also economically advantageous under realistic deployment scenarios.

The research also extends the frontier of anomaly detection into underground cable pits, a domain that has received limited attention due to its inherent data limitations. Through the use of event logs and probabilistic modelling, the thesis provides a framework for identifying degradation patterns even in the absence of direct measurements. This represents a significant step forward in enhancing grid observability and safety in urban areas.

Moreover, the thesis contributes an economic analysis framework that balances model performance with operational costs and benefits. This approach enables DSOs to make informed decisions about model deployment. The modularity and scalability of the proposed models further support their integration into agile development cycles and continuous improvement processes.

In alignment with strategic goals such as decarbonization, digitalization, and resilience, the work supports the evolution of DSOs into proactive system operators capable of managing increasingly complex and dynamic networks. It also aligns with key Sustainable Development Goals, reinforcing the societal relevance of the research. Overall, the thesis bridges the gap between academic innovation and industrial applicability, offering a robust foundation for future advancements in smart grid maintenance and reliability engineering.

Limitations of the study

Several limitations were encountered whilst performing this project.

The lack of direct measurements in underground cable pits constrained the model's accuracy. Event logs were used as proxies, which may not fully capture the degradation dynamics. For a more in-depth study, as for the secondary substation case, data availability of more complex behaviours of the line and other protection schemes would be useful.

While models were designed, where possible due to the agile workflow and time frame scope of the project, with efficiency in mind, large-scale deployment across hundreds of thousands of grid elements remains computationally intensive, though not economically unfeasible as calculated in the results section. Nevertheless, this project serves as an initial validation of the models and techniques and can be further improved with less computationally expensive models.

Incident descriptions in OMS databases required manual filtering due to inconsistent field reporting, limiting automation of labelling. The use of a large language

model can help automation of this section and would help with the detection of even more incidents and therefore of better training sets for the models.

The cost-benefit analysis had several assumptions such as excluding indirect social costs and reputational impacts, which could significantly affect the real world value proposition. The complexity of these aspects calls for a further in-depth study of these effects, specially for a regulated, natural monopoly as DSOs where service quality has strict criteria.

Recommendations for future research and industrial deployment

The findings of this thesis open several promising routes for future research and practical implementation. One of the most immediate recommendations is the enhancement of data granularity, particularly in underground cable pits. The current lack of direct measurements significantly limits the precision of anomaly detection models. Future deployments should consider integrating temperature and current sensors directly into these grid elements, enabling more accurate and timely detection of thermal degradation and electrical faults. Or further developing the decentralised detection strategies to flag automatically possibly damaged lines due to smart meter event logs that would then be analysed further with new models.

Another critical improvement lies in the integration of maintenance records and work orders into the anomaly detection framework. By incorporating these operational logs, models could reset anomaly counters after interventions, thereby avoiding false positives caused by residual effects of past incidents. This would also allow for a more dynamic and context aware model that adapts to the evolving state of the grid.

From a methodological standpoint, future research should explore more computationally efficient machine learning techniques. Techniques such as auto encoders, contrastive learning, or lightweight recurrent neural networks could offer similar or improved performance with reduced computational overhead, facilitating real-time deployment.

On the regulatory front, the current CAPEX-focused remuneration schemes present a barrier to the adoption of predictive maintenance strategies. A shift towards TOTEX or OPEX based models would better align incentives with the operational efficiencies offered by data-driven approaches. Researchers and industry stakeholders should collaborate to advocate for regulatory reforms that recognize the value of digitalization and proactive asset management to increase the current grid's capacity and reduce unexpected downtimes.

Finally state that, even after the successful completion of the project and the development of predictive maintenance algorithms for hotspot event anticipation in secondary substations and underground cable pits using data analytics, the project has opened several key investigation routs to further improve the models created. The last one which will be stated is the further study of social and safety implications of hotspot anomalies to quantify the reputational and public safety costs associated with these incidents, integrating them into economic feasibility models. Providing a more holistic view of the value proposition of predictive maintenance and support its prioritization in strategic planning.

Appendix A

SDG

The SDGs, Sustainable Development Goals [67], are a set of 17 goals created by the United Nations in order to provide a guide for peace, prosperity and sustainability of the human race. With an independence on the project or decisions made, it could be argued that if these blueprints are followed, this endeavour would be socially desirable.

The project is aligned with all goals focused on infrastructure, energy, innovation, reliability and sustainability. But, mainly with the following three SDGs:

7. Affordable and Clean Energy. By reducing interruptions and lowering maintenance costs through predictive maintenance, this project contributes to ensuring access to affordable, reliable, sustainable, and modern energy for all. Pre-emptive detection of anomalies helps avoid energy losses and improves the overall efficiency of the distribution network.

9. Industry, Innovation and Infrastructure. The project is deeply rooted in the modernization of electrical infrastructure. It leverages advanced metering infrastructure (AMI) and data-driven techniques to enhance the resilience and intelligence of the low-voltage distribution grid, promoting innovation in utilities.

11. Sustainable cities and communities. The current alternative to predictive maintenance is corrective maintenance, meaning that when secondary substations incur in an unexpected downtime, apart form the subsequent loss in reliability, once the users regain service it is mainly and temporarily via a diesel generator which increases pollution and then requires an environmentally expensive to make new secondary substation.

Bibliography

[1] Damiano Bracci and Simone Paglia. Advanced LV Management for Electrical Utilities. White paper, Gridspertise, 2023.

[2] CIGRÉ Task Force C6.04. *Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources*. Number 575 in Technical Brochure. CIGRÉ, 2014. Study Committee: Active Distribution Systems and Distributed Energy Resources.

[3] M. Arnold, W. Friede, and J.M.A. Myrzik. Challenges in future distribution grids - a review. *Proc. ICREPQ*, 20:1–6, 03 2013.

[4] Comisión Nacional de los Mercados y la Competencia - CNMC. Circular 6/2019, de 5 de diciembre, por la que se establece la metodología para el cálculo de la retribución de la actividad de distribución de energía eléctrica. BOLETÍN OFICIAL DEL ESTADO, n.º 304, Sección III, pág. 137528, December 2019.

[5] Enric R. Bartlett Castellà. Impulsar la digitalización del sistema eléctrico y el papel activo de los consumidores, February 2021.

[6] Insulated Conductors Committee. Ieee guide for early detection, mitigation, preventative measures, and response to smoke, fire, and explosions in underground electrical structures. *IEEE Std 2417-2022*, pages 1–69, 2022.

[7] Iberdrola. Buscamos alianzas para potenciar la detección de incidencias en nuestras redes de distribución, 2022.

[8] International Energy Agency. Unlocking Smart Grid Opportunities in Emerging Markets and Developing Economies. Technical report, International Energy Agency (IEA), February 2024.

[9] Ana Carolina Martinazzo Kanitz, Euler Ribeiro, Marcos Aurélio Izumida Martins, Kleber Duarte Tomaz, and Silvia de Francisci. Hot spot analysis in asset inspections in the electricity distribution area. In *2021 International Confer-*

ence on Electrical, Communication, and Computer Engineering (ICECCE), pages 1–5, 2021.

[10] Dave Anny. Impact on Predictive Maintenance and Reliability Engineering. *ResearchGate*, 2023.

[11] D.J. Allan. Fires and explosions in substations. In *IEEE/PES Transmission and Distribution Conference and Exhibition*, volume 1, pages 504–507 vol.1, 2002.

[12] Amaia Amantegui Escribano. Anomalías en CT relacionadas con el riesgo incendios, May 2025. Article published in the 'Gestión del Cambio' at i-DE (Iberdrola).

[13] IBM. What is advanced metering infrastructure (AMI)?, 2023. Accessed: 2025-07-21.

[14] Vida Rozite, Emi Bertoliand Brendan Reidenbach, and Kevin Lane. International energy agency (iea): Digitalisation. `https://www.iea.org/topics/digitalisation`. Last update on 12 July 2023.

[15] Jean-Philippe Vasseur and Adam Dunkels. *Interconnecting Smart Objects with IP: The Next Internet. Chapter 20: Smart Grids.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2010.

[16] Martin Högel, Oxana Dankova, Bas Sudmeijer, Maurice Berns, Eelke Kraak, Ferdinand Varga, Laura Villani, and Daniel Weise. Delivering the energy transition will come down to the wires. Whitepaper, Boston Consulting Group (BCG), February 2025.

[17] Chongqing Kang, Daniel Kirschen, and Timothy C. Green. The evolution of smart grids. *Proceedings of the IEEE*, 111(7):691–693, 2023.

[18] International Smart Grid Action Network. Isgan annual report 2024. Technical report, International Energy Agency (IEA), 2024.

[19] Ministerio para la Transición Ecológica y el Reto Demográfico. Plan Nacional Integrado de Energía y Clima 2023-2030 (PNIEC), September 2024. Versión actualizada del 24 de septiembre de 2024. Consultado el 21 de julio de 2025.

[20] European Commission. The EU Clean Energy Package, 2020.

[21] World Economic Forum, BloombergNEF, and Deutsche Energie-Agentur (dena). Harnessing Artificial Intelligence to Accelerate the Energy Transition. White paper, World Economic Forum, September 2021. Contributors: Claire

Curry, Jon Moore, Linda Babilon, Andreas Kuhlmann, Philipp Richard, Mark Caine, Dominique Hischier, Espen Mehlum.

[22] Iberdrola. Proyecto STAR: un referente mundial de eficiencia en inversiones en redes y contadores inteligentes, 2025.

[23] François Mirallès, Luc Cauchon, Marc-André Magnan, François Grégoire, Mouhamadou Makhtar Dione, and Arnaud Zinflou. Towards reliable detection of dielectric hotspots in thermal images of the underground distribution network. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3566–3574, New York, NY, USA, 2022. Association for Computing Machinery.

[24] Roberto Poyato. TERMOGRAFÍA EN SISTEMAS DE DISTRIBUCIÓN ELÉCTRICA. Nota técnica, Fluke Ibérica.

[25] Universidad Pontificia Comillas – Oficina de Comunicación. Expertos del sector energético requieren una mayor inversión en red eléctrica para que sea más resiliente y segura, March 2025. Consultado el 21 de julio de 2025.

[26] Antonio Martos Villar. La renovación de la CNMC atasca la retribución de las redes y complica la inversión de Iberdrola, Endesa y Naturgy, January 2025.

[27] Comisión Nacional de los Mercados y la Competencia. La CNMC inicia dos consultas públicas específicas sobre las circulares de metodología de la tasa de retribución financiera y la retribución de la distribución de energía eléctrica, May 2024.

[28] Comité para el Análisis de las Circunstancias que Concurrieron en la Crisis de Electricidad del 28 de abril de 2025. Versión no confidencial del informe del comité para el análisis de las circunstancias que concurrieron en la crisis de electricidad del 28 de abril de 2025. Technical report, Gobierno de España, Consejo de Seguridad Nacional, June 2025.

[29] Muhammed Fatih Pekşen, Ulaş Yurtsever, and Yılmaz Uyaroğlu. Enhancing electrical panel anomaly detection for predictive maintenance with machine learning and iot. *Alexandria Engineering Journal*, 96:112–123, 2024.

[30] Philipp zur Heiden, Jennifer Priefer, and Daniel Beverungen. Predictive maintenance on the energy distribution grid–design and evaluation of a digital industrial platform in the context of a smart service system. *IEEE Transactions on Engineering Management*, 71:3641–3655, 2024.

[31] R.E. Snodgrass and W.Z. Black. Mitigating the effects of explosions in underground electrical vaults. *IEEE Transactions on Power Delivery*, 20(2):1767–1774, 2005.

[32] IBM. What is a maintenance strategy?, 2024.

[33] ABB. 4 types of maintenance strategy, which one to choose?, 2024.

[34] Endesa. Maintaining the power distribution grid, 2025.

[35] Soesatijono Soesatijono and Mahros Darsin. Literature studies on maintenance management. *Journal of Energy Mechanical Material and Manufacturing Engineering*, 6:67–74, 05 2021.

[36] Philipp zur Heiden, Jennifer Priefer, and Daniel Beverungen. Predictive maintenance on the energy distribution grid–design and evaluation of a digital industrial platform in the context of a smart service system. *IEEE Transactions on Engineering Management*, 71:3641–3655, 2024.

[37] Martin Neumayer, Dominik Stecher, Sebastian Grimm, Andreas Maier, Dominikus Bücker, and Jochen Schmidt. Fault and anomaly detection in district heating substations: A survey on methodology and data sets. *Energy*, 276:127569, 2023.

[38] Rawal Keerti and Ahmad Aijaz. *5 Big Data Analytical Techniques for Electrical Energy Forecasting in Smart Grid Paradigm*, pages 101–126. 2023.

[39] Akash Sharma and Rajive Tiwari. Anomaly detection in smart grid using optimized extreme gradient boosting with scada system. *Electric Power Systems Research*, 235:110876, 2024.

[40] Young-Seob Jeong, JunHa Hwang, SeungDong Lee, Goodwill Ndomba, Youngjin Kim, and Jeung-Im Kim. Sensor-based indoor fire forecasting using transformer encoder. *Sensors*, 24:2379, 04 2024.

[41] Malin Bülund. Exploring integration of predictive maintenance using anomaly detection: Enhancing productivity in manufacturing. Master's thesis, KTH Royal Institute of Technology, 2024.

[42] Ahish Shylendra, Priyesh Shukla, Saibal Mukhopadhyay, Swarup Bhunia, and Amit Ranjan Trivedi. Low power unsupervised anomaly detection by non-parametric modeling of sensor statistics. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2020.

[43] Ming Zhou. Anomaly detection in smart distribution grids with deep neural network. Master's thesis, University of Alberta, 2022.

[44] Ming Zhou and Petr Musilek. Real-time anomaly detection in distribution grids using long short term memory network. In *2021 IEEE Electrical Power and Energy Conference (EPEC)*, pages 208–213, 2021.

[45] Yangfan Jiang, Jianfei Song, Xiaoyu Yang, Xiao Dong, Songyu Sun, Yibo Lin, Zhou Jin, Xunzhao Yin, and Cheng Zhuo. A parallel simulation framework incorporating machine learning-based hotspot detection for accelerated power grid analysis. In *2024 ACM/IEEE 6th Symposium on Machine Learning for CAD (MLCAD)*, pages 1–7, 2024.

[46] Mohit Choubey, Rahul Kumar Chaurasiya, and J.S. Yadav. Contrastive learning for efficient anomaly detection in electricity load data. *Sustainable Energy, Grids and Networks*, 42:101639, 2025.

[47] Indunova. Termografía: ¿Cuáles son las causas que originan puntos calientes en un sistema eléctrico?, 2020.

[48] Solidification Products International. How to Prevent Substation Fires, 2023.

[49] Practicus AI. Medium: The 5 clustering algorithms data scientists need to know, February 2018.

[50] Rahul Kaliyath. Day 11: Gaussian Mixture Model Clustering, July 2020.

[51] Cole Stryker. IBM: What is a recurrent neural network (RNN)?, 2024. Publicado en IBM Think. Consultado el 21 de julio de 2025.

[52] Vatsal Raval. When to use MLP, CNN or RNN?, January 2019.

[53] Joaquín Amat Rodrigo. Detección de anomalías con Gaussian Mixture Model (GMM) y Python, December 2020.

[54] IEC-CENELEC. *Measuring Relays and Protection Equipment - Part 149: Functional Requirements for Thermal Electrical Relays (IEC 60255-149:2013)*. AENOR, Avenue Marnix 17, B-1000 Brussels, 2013.

[55] Iberdrola Group, STAR project i-DE Project Team. Et funcionalidad supervision avanzada bt – edition 9.6. Technical report, Iberdrola, March 2022. Internal technical report, Edition 9.6 dated 09/03/2022.

[56] CloudPrice.net. Azure VM Comparison, 2025.

[57] Ali Al Bataineh, Devinder Kaur, and Seyed Mohammad J. Jalali. Multilayer perceptron training optimization using nature inspired computing. *IEEE Access*, 10:36963–36977, 2022.

[58] Lorenzo Mascali, Daniele Salvatore Schiera, Simone Eiraudo, Luca Barbierato, Roberta Giannantonio, Edoardo Patti, Lorenzo Bottaccioli, and Andrea Lanzini. A machine learning-based anomaly detection framework for building electricity consumption data. *Sustainable Energy, Grids and Networks*, 36:101194, 2023.

[59] Ibtissam Amalou, Naoual Mouhni, and Abdelmounaim Abdali. Multivariate time series prediction by rnn architectures for energy consumption forecasting. *Energy Reports*, 8:1084–1091, 2022. Technologies and Materials for Renewable Energy, Environment and Sustainability.

[60] Budi Santoso, Wiwik Anggraeni, Henry Pariaman, and {Mauridhi Hery} Purnomo. Rnn-autoencoder approach for anomaly detection in power plant predictive maintenance systems. *International Journal of Intelligent Engineering and Systems*, 15(4):363–381, 2022.

[61] Piero Danti and Alessandro Innocenti. A methodology to determine the optimal train-set size for autoencoders applied to energy systems. *Advanced Engineering Informatics*, 58:102139, 2023.

[62] Quality Gurus. Nelson Rules (and Western Electric Rules) for Control Charts, 2020.

[63] Advantive. Mastering Quality: Statistical Process Control 101. White paper, Advantive, 2025.

[64] Jiawei Han, Micheline Kamber, and Jian Pei. 11 - advanced cluster analysis. In Jiawei Han, Micheline Kamber, and Jian Pei, editors, *Data Mining: Concepts and Techniques (Third Edition)*, The Morgan Kaufmann Series in Data Management Systems, pages 497–541. Morgan Kaufmann, Boston, third edition edition, 2012.

[65] Oliver Urs Lenz and Matthijs van Leeuwen. Monotonic anomaly detection, 2025.

[66] Boletín Oficial del Estado (BOE). Regulation on low voltage electrical installations and complementary technical instructions (rebt). `https://www.boe.es/biblioteca_juridica/codigos/codigo.php?modo=2&id=326_Reglamento_electrotecnico_para_baja_tension_e_ITC`, April 2025.

[67] United Nations. Sustainable Development Goals, 2015.