Master's Degree in Industrial Engineering

Master Thesis

NB-PLC Frequency Bands (151-471 kHz) Feasibility Study for PRIME v1.4 Protocol

Author: Víctor Arias Blanco

Director: Javier Matanza Domingo

Director: Alberto Sendín Escalona

Madrid

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

NB-PLC Frequency Bands (151-471 kHz) Feasibility Study for PRIME v1.4 Protocol

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2019/20 es de mi 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.

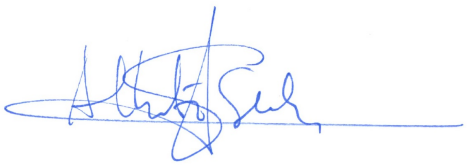
Fdo.: Víctor Arias Blanco Fecha: 30/08/2020

Autorizada la entrega del proyecto

LOS DIRECTORES DEL PROYECTO

Fdo.: Javier Matanza Domingo Fecha: 30/08/2020

Fdo.: Alberto Sendín Escalona Fecha: 30/08/2020



Master's Degree in Industrial Engineering

Master Thesis

NB-PLC Frequency Bands (151-471 kHz) Feasibility Study for PRIME v1.4 Protocol

Author: Víctor Arias Blanco

Director: Javier Matanza Domingo

Director: Alberto Sendín Escalona

Madrid

Dedication

To MY PARENTS,
whose effort, affection and education have allowed me to reach my goals. Always have been there, always will be there.

To MY GIRLFRIEND,
whose love has accompanied me since high school, being my support in the hard days and my joy in the good ones.

To MY FLAT MATES,
those responsible for me being able to bear 6 years at university without losing my mind too much.

To COFFEE,
who have made possible to finish this project in time.

NB-PLC Frequency Bands (151-471 kHz) Feasibility study FOR PRIME v1.4 PROTOCOL

Autor: Arias Blanco, Víctor.
Director: Matanza Domingo, Javier.
Director: Sendín Escalona, Alberto
Entidad Colaboradora: Iberdrola

Resumen — *La comunicación de banda estrecha por líneas de potencia (NB-PLC) es una de las tecnologías de acceso a los contadores inteligentes más comúnmente utilizadas. Con el avance de estas tecnologías, dispositivos y protocolos van aumentando las bandas de frecuencia disponibles. Este es el caso de la nueva versión del protocolo PRIME v1.4, cuyas nuevas bandas de frecuencia son el objeto de estudio de este proyecto.*

Este articulo trata de recabar la información existente del espectro de frecuencias entre 151 y 471 kHz para posteriormente compararla con un análisis propio de desempeño en campo. Tras el análisis, se llevará a cabo una agrupación de los contadores con patrones de comunicación similares, basados en el desempeño en cada banda de frecuencia. Tras dicha agrupación, se buscan propiedades de la topología de la red comunes para cada grupo.

Palabras Clave — PRIME, PLC, Banda FCC, Ruido, Agrupación, Clasificación, Contadores Inteligentes

I. Introducción

Los contadores inteligentes representan el primer paso hacia la digitalización masiva de la red de baja tensión. Esta red inteligente no sólo se denomina como tal por posibilidad de transmisión de consumos por parte de los contadores, si no por la posibilidad de lectura de variables de red, recepción de ordenes de conexión, desconexión o reducción de consumo que serán acciones necesarias para un funcionamiento avanzado de la red futura.

Con el despliegue de contadores inteligentes ha ido comúnmente asociada la instalación de tecnologías PLC para su comunicación. Esta tecnología ha ido desarrollándose debido a la aparición de objetivos que alcanzar o dificultades que solucionar. En este proceso apareció el protocolo PRIME (PoweRline Intelligent Metering Evolution) para la estandarización de esta comunicación en varias de las compañías de distribución.

Típicamente, en Europa se ha utilizado la llamada banda CENELEC-A que incluiría las frecuencias de 3 a 95 kHz, reservada para proveedores de electricidad. Posteriormente se añadiría una nueva banda de frecuencias que llegaría a los 148.5 kHz, pero ésta sería destinada a la comunicación interna a los domicilios. Con la aparición de necesidades de ancho de banda y de evitar posibles interferencias, la inclusión de bandas utilizadas comúnmente en otros continentes fue objetivo de los nuevos avances en los protocolos.

Una actualización del protocolo PRIME [1] fue presentada para permitir una mejor y más rápida transmisión de datos, la cual incluye la ampliación de seis bandas de frecuencia adicionales entre 151 y 471 kHz. Esta incorporación ampliaba en gran medida las posibilidades de comunicación añadiendo un mayor ancho de banda y franjas de frecuencia altas para disminución de ruido en la transmisión. El desempeño de estas nuevas bandas será el sujeto de análisis de este proyecto.

II. Definición del Proyecto

Este Proyecto enfrenta la tarea reunir el conocimiento existente de las bandas de frecuencias entre 151 y 471 kHz, y aplicarlo para analizar la calidad de la comunicación en los dispositivos de la red. La comunicación con estos dispositivos puede funcionar con distinta calidad en los diferentes canales, denominación otorgada a las bandas de frecuencia del protocolo. Identificar los parámetros de calidad y ayudar a esclarecer las razones de esas diferencias entre dispositivos y canales son tareas que llevar a cabo en este artículo.

III. Desarrollo del Proyecto

La investigación del estado del arte aporta algunos conceptos importantes que verificar en el posterior análisis. El primero es el incremento de la atenuación con la frecuencia, a causas mayormente del aumento en la impedancia. Este hecho generaría señales de menor intensidad en los canales altos cuando las líneas son más largas [2],[3].

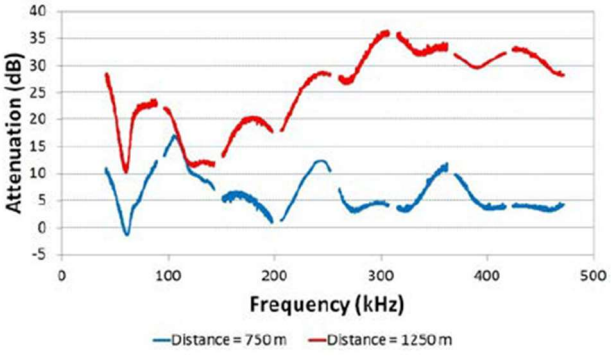


Fig. 1: Atenuación según la frecuencia para una linea rural

El segundo concepto es que existe un mayor nivel de ruido para cargas típicas en las frecuencias más bajas del espectro, aunque hay otras más localizadas que generan un ruido más homogéneo [4], [5]. Por último, se constata que interferencias puntuales debido al entorno en alguno de los canales puede reducir o inutilizar el canal que la sufra [6].

Tras analizar los datos existentes, se procede a adquirir datos propios con los que hacer un estudio de desempeño de dichos canales. Una vez se procesan las cifras, se examinan posibles tendencias que puedan resaltar de éstas y se extraen posibles indicadores del desempeño. La identificación de los canales con peor desempeño para cada terminal permite que éstos se puedan evitar, mientras que en los que sea óptimo, se priorizan. El siguiente paso en el desarrollo es la obtención de características topológicas que favorezcan los malos resultados en ciertos canales, para prevenir de ese riesgo en líneas con condiciones semejantes. El proceso de agrupación para identificar mala calidad de comunicación en ciertos contadores se realiza mediante un *clustering* con el método *kmeans*. Por otro lado, la identificación de posibles variables responsables de esos datos se llevará a cabo por árboles de clasificación.

IV. Resultados

Tras una investigación de la red de baja tensión, definiéndola en jerarquía y componentes y comprobar una homogeneidad elevada en los materiales, se adquirieron datos de contadores conectados a centros de transformación en diferentes localizaciones. De éstos se extrajo la disponibilidad, la relación señal ruido y un indicador de la intensidad de la señal para cada dispositivo y haciendo una media cada 10 minutos. Con esta información dispuesta para la visualización, se extrajo una tendencia alcista de la mediana para la disponibilidad, a pesar de que la relación señal ruido no se comportaba tan favorablemente en los extremos del espectro. El indicador de intensidad de señal redujo su importancia debido al cálculo indirecto de éste y a la dependencia con respecto al ruido, en el cual a valores bajos de la relación señal ruido, el valor de la intensidad de señal no era fiable.

Tras un cálculo de variables estadísticas diarias para cada dispositivo, éstas se introdujeron en un proceso de reducción de variables. Como el número de variables era complejo de representar y manipular, además de la existencia de una correlación entre las variables, se calcularon otras que eran combinación lineal de las originales. Con esta combinación lineal se reduce el número de variables a un par sin perder gran información sobre la varianza de los datos.

Las nuevas variables son requeridas para realizar una agrupación de los terminales en función del desempeño en la comunicación para cada canal, que está representado por éstas. La agrupación se hace por un método de *clustering* denominado *kmeans,* tanto en el conjunto de canales como canal a canal. El objetivo es identificar los contadores que presentan una mala calidad de comunicación en algún canal, o presentan un patrón específico entre todos ellos, y asignarle una etiqueta de grupo. En la Fig. 2 se muestra la agrupación según las variables generadas por combinación lineal de las originales. La imagen mostrada es para el séptimo canal y el conjunto se divide en dos grupos, uno con un desempeño aceptable y otro con uno inadmisible. Se puede sacar esa conclusión de estos valores porque los coeficientes de la combinación lineal son mayoritariamente positivos, por lo que altos valores de estas variables son altos valores de las originales.

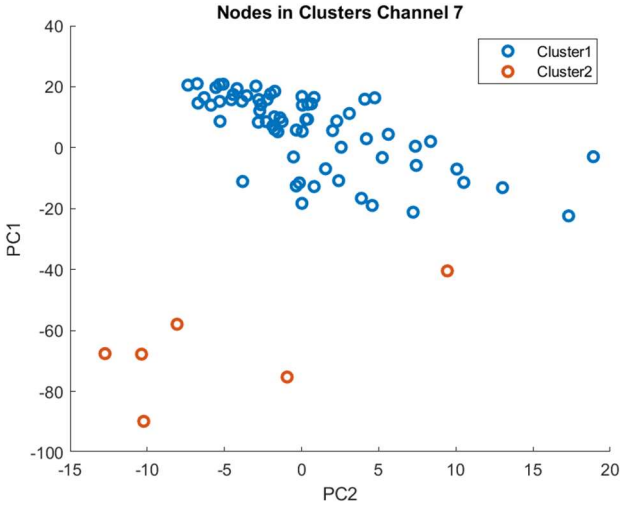


Fig. 2: Agrupación de terminales según desempeño por componentes principales

Por otra parte, en la Fig. 3 se refleja el patrón a lo largo de los canales de cada grupo si el conjunto de dispositivos se dividiese en tres. El valor mostrado es la puntuación de la primera variable combinación lineal de las originales. Se puede observar un grupo que tiene un desempeño bueno en todos los canales, otro que mejora cuando la frecuencia aumenta y otro que, al contrario, empeora en los canales altos.

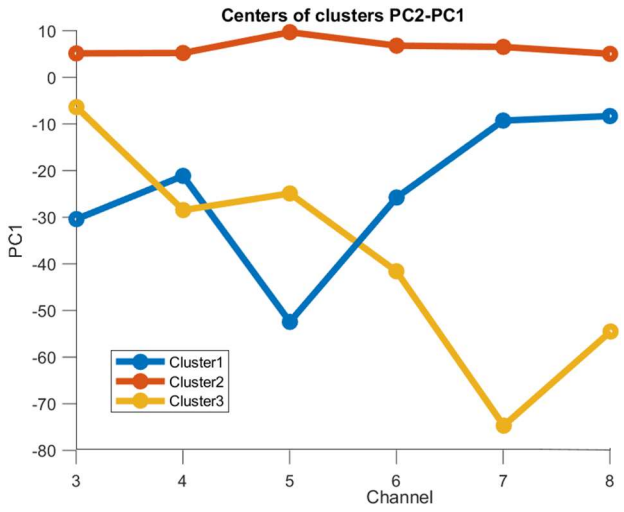


Fig. 3: Valor central de la 1ª variable de los grupos en cada canal

Una vez los dispositivos se han agrupado y etiquetado, se lleva a cabo una clasificación con unas variables distintas. Éstas nuevas variables tienen relación con la topología de la red, como el numero de clientes en la caja general de

protección, la potencia contratada o la distancia hasta el centro de transformación. Si un árbol de decisión, entrenado con parte de los datos extraídos, es capaz de clasificar los terminales en las etiquetas asignadas previamente con las variables topológicas, implica que existe una correlación entre ambos conjuntos de variables. Si existe tal relación entre las variables de desempeño en la comunicación y las variables topológicas seleccionadas, el árbol de decisión aportará que valores de éstas últimas genera una mala calidad de comunicación.

En la Fig. 4 se muestra la prueba de clasificación de un árbol de decisión, entrenado previamente con parte de los datos mostrados en la Fig. 2. Con el resto de las muestras no utilizadas en el entrenamiento se probó la eficacia del clasificador. En este caso el resultado es muy favorable al sólo errar un dispositivo, aunque la diferencia en el volumen de muestras por grupo puede sesgar la clasificación.

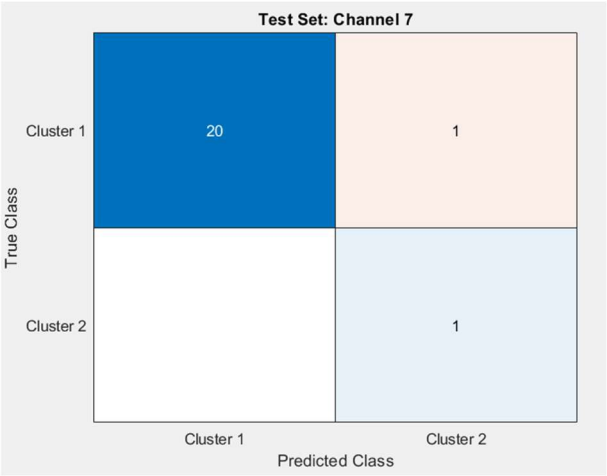


Fig. 4: Matriz de confusión de la colección de muestras de prueba

Las variables utilizadas principalmente para conseguir la separación entre grupos para el canal siete fueron la distancia al centro de transformación y el tipo de línea, si subterránea o aérea. Los malos datos serán más proclives en líneas más cortas o con líneas aéreas. En canales como el tercero las divisiones entre grupos no son clara lo que también impide la correcta clasificación, para lo que deberían incluirse otras variables en ambos procesos

V. Conclusiones

El análisis propio de los datos verifica algunas afirmaciones extraídas del estudio del estado del arte, como puede ser el mayor ruido presente en los canales bajos o la mayor atenuación de señal en los altos. Esas consideraciones son perceptibles en los valores de relación señal ruido, los cuales son más reducidos en los extremos debido a la mayor potencia del ruido en el inferior y al nivel mas reducido de señal en el superior. En cambio, la disponibilidad del canal está más vinculada al ruido y por tanto es monótona creciente, ya que, si el ruido es muy reducido, aunque la señal esté atenuada, la comunicación será satisfactoria.

En cuanto a la clasificación, se ha mostrada fiable en la separación según el desempeño en la comunicación, incluso identificando patrones específicos entre los canales. La excepción serían los canales de frecuencias más reducidas, en los cuales disponibilidad y relación señal ruido tienen una gran correlación y por tanto se pierde varianza en los datos, empeorando la agrupación.

La clasificación va ligada a la calidad de la agrupación, siendo más identificables los grupos que tengan diferencias más notorias entre los elementos integrantes. Además de esa dependencia, la clasificación es más sometido a contar con gran número de observaciones para ser fiable. Con todo ello, se aprecia claramente que los canales de alta frecuencia tienen mayor facilidad en la clasificación a pesar de contar con pocos elementos en el grupo de peor desempeño.

Por tanto, como resumen y como base futuros desarrollos, los principales frentes de ataque serían tres en orden de importancia: aumentar el número de observaciones, es decir, la cantidad de contadores analizados, para mejorar ambos procesos de agrupación y clasificación; añadir nuevas variables de clasificación que puedan reflejar las diferencias en desempeño, empezando por la identificación de una que refleje el ruido existente en la línea; por último, nuevas variables estadísticas para la agrupación, intentando mejorar la diferenciación en canales de frecuencias bajas, como medianas diurnas y nocturnas o varianzas.

VI. Reconocimiento

El autor quiere agradecer a Iberdrola su apoyo y aporte de información para el estudio.

VII. Referencias

[1] PRIME Alliance TWG, «PRIME v1.4 White Paper». Accedido: may 06, 2020. [En línea]. Disponible en: https://www.prime-alliance.org/wp-content/uploads/2014/10/whitePaperPrimeV1p4_final.pdf.

[2] I. Fernandez, A. Arrinda, I. Angulo, D. De La Vega, N. Uribe-Perez, y A. Llano, «Field Trials for the Empirical Characterization of the Low Voltage Grid Access Impedance From 35 kHz to 500 kHz», IEEE Access, vol. 7, pp. 85786-85795, 2019, doi: 10.1109/ACCESS.2019.2924253.

[3] I. Fernández et al., «Characterization of the frequency-dependent transmission losses of the grid up to 500 kHz», Madrid, Spain, jun. 2019, vol. 1146.

[4] I. Fernández, D. de la Vega, A. Arrinda, I. Angulo, N. Uribe-Pérez, y A. Llano, «Field Trials for the Characterization of Non-Intentional Emissions at Low-Voltage Grid in the Frequency Range Assigned to NB-PLC Technologies», Electronics, vol. 8, n.o 9, Art. n.o 9, sep. 2019, doi: 10.3390/electronics8091044.

[5] I. Fernandez et al., «Characterization of non-intentional emissions from distributed energy resources up to 500 kHz: A case study in Spain», Int. J. Electr. Power Energy Syst., vol. 105, pp. 549-563, feb. 2019, doi: 10.1016/j.ijepes.2018.08.048.

[6] I. Arechalde, M. Castro, I. Garcia-Borreguero, A. Sendin, I. Urrutia, y A. Fernandez, «Performance of PLC communications in frequency bands from 150 kHz to 500 kHz», en 2017 IEEE International Symposium on Power Line Communications and its Applications (ISPLC), Madrid, Spain, 2017, pp. 1-5, doi: 10.1109/ISPLC.2017.7897123.

NB-PLC Frequency Bands (151-471 kHz) Feasibility study FOR PRIME v1.4 PROTOCOL

Author: Arias Blanco, Víctor.
Director: Matanza Domingo, Javier.
Director: Sendín Escalona, Alberto
Collaborating Entity: Iberdrola

Abstract — *Narrowband Power Line Communication (NB-PLC) is one of the most used technologies for smart metering access. With the development of these technologies, devices and protocols increase the available frequency bands. This is the case of the new version of the PRIME v1.4 protocol, whose the new frequency bands are the subject of study of this project.*

This article seeks to collect existing frequency spectrum information between 151 and 471 kHz and then compare it with its own field performance analysis. After the analysis, a grouping of the counters with similar communication patterns will be carried out, based on performance in each frequency band. After that grouping, common network topology properties are searched for each group.

Index Terms — PRIME, PLC, FCC Band, Noise, Clustering, Classification, Smart Metering

I. Introduction

Smart meters represent the first step towards mass digitization of the low voltage network. This smart grid is not only referred as such by the possibility of transmission of consumptions by the meters, but also by the possibility of reading network variables, receiving connection/ disconnection orders, or the reduction of consumption, which will be necessary actions for advanced operation of the future grid.

The deployment of smart meters has been commonly associated with the installation of PLC technologies for communication. This technology has been developed due to the emergence of objectives to be achieved or difficulties to be solved. In this process, the PRIME (PoweRline Intelligent Metering Evolution) protocol appeared for the standardization of this communication in several of the distribution companies.

Typically, in Europe the so-called CENELEC-A band has been the one used, which would include frequencies from 3 to 95 kHz reserved for electricity suppliers. A new frequency band would then be added to reach 148.5 kHz, but this would be intended for in-home communication. With the appearance of bandwidth needs and avoidance of potential interference, the inclusion of commonly used bands on other continents was the target of new advances in protocols.

An update to the PRIME [1] protocol was introduced to allow for better and faster data transmission, which includes the expansion of six additional frequency bands between 151 and 471 kHz. This addition greatly expanded communication possibilities by adding higher bandwidth and frequency bands to reduce transmission noise. The performance of these new bands will be the subject of analysis of this project.

II. Project Definition

This project faces the task of gathering existing knowledge of the frequency bands between 151 and 471 kHz and applying it to analyze the quality of communication on the devices of the network. Communication with these devices can operate with different quality on different channels, a designation given to the frequency bands of the protocol. Identifying quality parameters and helping to clarify the reasons for those differences between devices and channels are tasks to perform in this article.

III. Project Development

State-of-the-art research provides some important concepts to verify in subsequent analysis. The first is increase of attenuation with frequency, caused mostly by the increase in impedance. This would generate lower intensity signals in high channels when the lines are longer [2], [3].

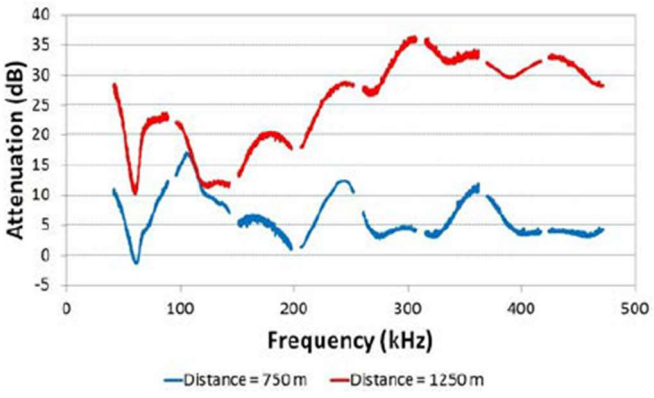


Fig. 5Attenuation through frequency in a rural line

The second concept is that there is a higher noise level for typical loads at the lower frequencies of the spectrum, although there are more localized ones that generate a more homogeneous noise [4], [5]. Finally, it is found that a spike of interference due to the environment on one of the channels can reduce or disable the channel that suffers it [6].

After analyzing the existing data, we proceed to acquire own data to make a performance study of these channels. Once the figures are processed, possible trends that can be highlighted are examined and possible performance indicators are extracted. Identifying the worst performing channels for each terminal allows them to be avoided, while optimal channels are prioritized. The next step in development is to obtain topological characteristics that favor poor results in certain channels, to prevent from that risk in lines with similar conditions. The grouping process to identify poor communication quality on certain counters is performed by clustering with the *kmeans* method. On the other hand, the identification of possible variables responsible for such data shall be carried out by classification trees.

IV. Results

After a low voltage grid research, defining it in hierarchy and components and checking a high homogeneity in the materials, data from meters connected to processing centers was acquired in different locations. From these was extracted availability, Signal-to-Noise Ratio (SNR) and a signal strength indicator for each device and making an average every 10 minutes. With this information arranged for visualization, an uptrend was extracted from the median for availability, even though the SNR did not behave as favorably at the extremes of the spectrum. The signal strength indicator reduced its importance due to indirect calculation of the signal and dependence on noise, in which at low values of SNR, the signal strength value was unreliable.

After a calculation of daily statistical variables for each device, these were introduced in a variable reduction process. Because the number of variables was complex to represent and manipulate, in addition to the existence of a correlation between the variables, others were calculated that were linear combinations of the originals. This linear combination reduces the number of variables to a pair without losing much information about the variance of the data.

The new variables are required to group the terminals based on the performance in the communication for each channel, which is represented by them. Grouping is done by a *clustering* method named *kmeans,* both in the channel set and channel-to-channel. The goal is to identify meters that have poor communication quality in some channel, or have a specific pattern between all of them, and assign a group tag to them. Fig. 2 shows the clustering according to the variables generated by linear combination of the originals. The image shown is for the seventh channel and the set is divided into two groups, one with acceptable performance and one with an inadmissible one. This conclusion can be drawn from these values because the coefficients of the linear combination are mostly positive, so high values of these variables are high values of the originals.

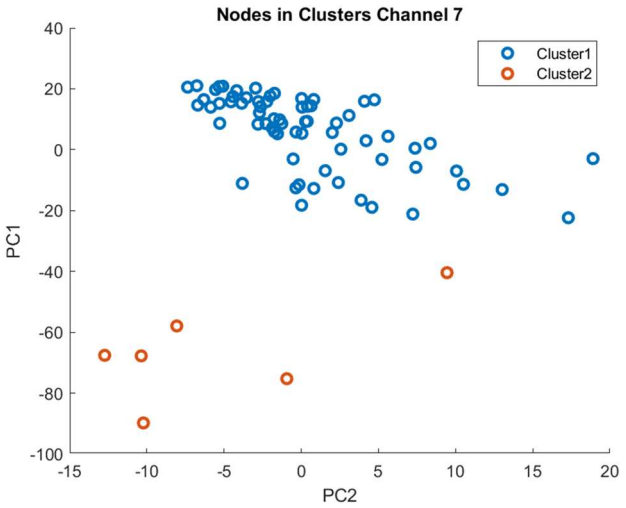


Fig. 6: Grouping terminals according to performance by main components

On the other hand, in Fig. 3 a pattern is reflected along the channels in each group if the device set is divided into three. The value shown is the score of the first linear combination variable of the originals. One group that performs well across all channels can be observed, another that improves when the frequency increases and another that, on the contrary, worsens in high channels.

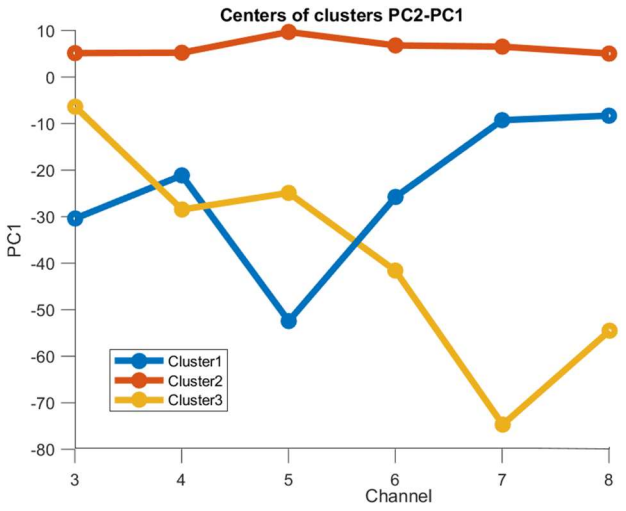


Fig. 7: Central value of the 1st variable of the groups in each channel

Once the devices have been grouped and tagged, a classification is performed with different variables. These new variables relate to the grid topology, such as the number of customers in the fuses box, the contracted power, or the distance to the secondary substation. If a decision tree, trained with some of the extracted data, is able to classify the terminals into the labels previously assigned with the topological variables, it implies that there is a correlation between the two sets of variables. If there is such a relationship between the performance variables in the communication and the selected topological variables, the decision tree will provide that values of the latter that generate a poor quality of communication.

Fig. 4 shows the classification test of a decision tree,

previously trained with some of the data shown in Fig. 2 With the rest of the samples not used in the training, the effectiveness of the classifier was tested. In this case the result is very favorable just by erring a device, although the difference in the volume of samples per group can skew the classification.

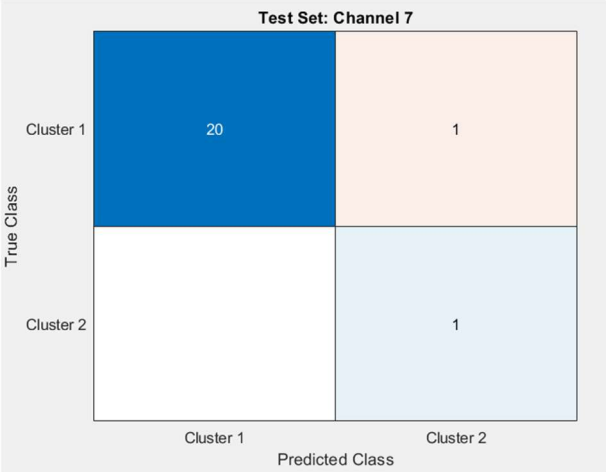


Fig. 8: Confusion matrix of test sample collection

The variables mainly used to achieve the separation between groups for channel seven were the distance to the transformation center and the type of line, whether underground or overhead. Bad data will be more likely on shorter lines or with overhead lines, according to the classification tree. In channels such as the third, divisions between groups are not clear which also prevents proper classification, for which, other variables should be included in both processes.

V. Conclusions

The data analysis verifies some statements taken from the state-of-the-art study, such as the increased noise present in the low channels or the increased signal attenuation in the highs. These considerations are noticeable in the SNR values, which are smaller at the extremes due to the higher noise power at the bottom and the lower signal level at the top. On the other hand, the availability of the channel is more linked to noise and is therefore monotonous and increasing, because if the noise is very low, even if the signal is attenuated, the communication will be satisfactory.

In terms of classification, it has been shown to be reliable in separation based on performance in communication, even identifying specific patterns between channels. The exception would be the lower frequency channels, in which availability and SNR have a high correlation and therefore variance in the data is lost, worsening the grouping.

The classification is linked to the quality of the clustering, with the groups that have the most noticeable differences between the members being more identifiable. In addition to this dependency, the classification is more subject to a large number of observations to be reliable. As a result, it is clear that high-frequency channels are easier in classification despite having few elements in the worst performing group.

Therefore, as a summary and as a basis for future developments, the main fronts of attack would be three in order of importance: increasing the number of observations, i.e. the number of meters analyzed, to improve both grouping and classification processes; add new classification variables that may reflect differences in performance, starting with identifying one that reflects the existing noise on the line; finally, new statistical variables for grouping, trying to improve differentiation in low-frequency channels, such as daytime and night medians or variances.

VI. Acknowledgement

The author wants to thank Iberdrola for his support and contribution in this this study.

VII. References

[1] PRIME Alliance TWG, «PRIME v1.4 White Paper». Accedido: may 06, 2020. [En línea]. Disponible en: https://www.prime-alliance.org/wp-content/uploads/2014/10/whitePaperPrimeV1p4_final.pdf.

[2] I. Fernandez, A. Arrinda, I. Angulo, D. De La Vega, N. Uribe-Perez, y A. Llano, «Field Trials for the Empirical Characterization of the Low Voltage Grid Access Impedance From 35 kHz to 500 kHz», IEEE Access, vol. 7, pp. 85786-85795, 2019, doi: 10.1109/ACCESS.2019.2924253.

[3] I. Fernández et al., «Characterization of the frequency-dependent transmission losses of the grid up to 500 kHz», Madrid, Spain, jun. 2019, vol. 1146.

[4] I. Fernández, D. de la Vega, A. Arrinda, I. Angulo, N. Uribe-Pérez, y A. Llano, «Field Trials for the Characterization of Non-Intentional Emissions at Low-Voltage Grid in the Frequency Range Assigned to NB-PLC Technologies», Electronics, vol. 8, n.o 9, Art. n.o 9, sep. 2019, doi: 10.3390/electronics8091044.

[5] I. Fernandez et al., «Characterization of non-intentional emissions from distributed energy resources up to 500 kHz: A case study in Spain», Int. J. Electr. Power Energy Syst., vol. 105, pp. 549-563, feb. 2019, doi: 10.1016/j.ijepes.2018.08.048.

[6] I. Arechalde, M. Castro, I. Garcia-Borreguero, A. Sendin, I. Urrutia, y A. Fernandez, «Performance of PLC communications in frequency bands from 150 kHz to 500 kHz», en 2017 IEEE International Symposium on Power Line Communications and its Applications (ISPLC), Madrid, Spain, 2017, pp. 1-5, doi: 10.1109/ISPLC.2017.7897123.

Index

Figure Index

Table Index

Chapter 1. INTRODUCTION

Telecommunications in the electricity distribution sector are getting more and more relevant with the progress of automatization and monitoring of the low voltage grid. The massive deployment of Smart Meters made in Europe and specifically in Spain accelerated the need of a generalized telecommunications connectivity of the whole grid. This process must be based on a very cost-effective solution due to the number of devices for its installation. One of the preferred solutions for connecting all these points is the Power Line Communication (PLC) due to the vast coverage of electricity grid and the low installation and maintenance costs.

PRIME is one of the main standardized protocols for narrow band PLC, growing intensively hand in hand with the expansion of the companies member of PRIME Alliance. Iberdrola, as the greatest exponent of the association, has been deploying the PRIME protocol extensively in Spain during the last decade with a good effectiveness and more sparingly in the rest of counties where the company is present. However, with the new objectives of transmission rate, BER and coverage, a new version of the protocol has been developed, PRIME v1.4. One of the main characteristics of this new version is the addition of more frequency bands for communication with the Smart Meters and other devices of the grid. The study of these new frequency bands (named channels in this document), with regards to performance, will be the center of the project.

The deep understanding of the performance of the band below 500 kHz (commonly but not very strictly referred to as Federal Communications Commission FCC frequency band in literature) in PLC will be the base for selecting the proper channel for each device, and so improve the communication performance indicators. Once meters can be differentiated by the performance, grouping them depending on those indicators, or even looking for grid properties on those meters that reflects a common performance pattern, will be possible.

1.1 ACRONYMS

DER: Distributed Energy Resources

EVM: Error Vector Magnitude

FCC: Federal Communications Commission

HCB: House Connection Box

LVP: Low Voltage Panel

LV: Low Voltage

MV: Medium Voltage

NB-PLC: Narrow Band Power Line Communication

PLC: Power Line Communication

PV: Photovoltaic

PCA: Principal Component Analysis

RSSI: Received Signal Strength Indication

SNR: Signal to Noise Ratio

SDG: Sustainable Development Goals

SS: Secondary Substation

1.2 TECHNOLOGY DESCRIPTION

This subsection is design for introducing some basic concepts of the projects that will not be explained afterwards but are needed to be understood for the proper comprehension of the text.

1.2.1 PRIME PROTOCOL

PRIME is a PLC protocol developed by the PRIME Alliance and used for the communication between data concentrators in the SS and the meters located at the customer`s property through the power cables.

There has been an update of the protocol, from PRIME v1.3.6 to v1.4 [1], in which some new features of the communication layer have been included. Besides changes in the

transmission load and robustness of the messages and the improvements in other layers, the main feature included for this project is the addition of three new frequency bands to communicate.

The frequency bands used in PRIME v1.4 are shown in Table 1. This project focusses its attention to the channels 3 to 8, whose performance is not yet totally acknowledged.

Table 1: Frequency Bands for PRIME v1.4

Channel	Frequency Bands (kHz)
1	42 - 89
2	97 – 144
3	151 – 198
4	206 - 253
5	261 – 308
6	315 - 362
7	370 – 417
8	424 - 471

In the PRIME protocol, there are different devices involved in the communication. In the SS, where the Data Concentrator 8DC) is located, there is a base node that transmits the messages from the DSO to the customers' meters and receives the data in the other way. The customer's meter will be the service node, that in this paper will be called simply node. There is other definition for a function that a service node can carry out, it is the switch node; it is a service node that, because the signal from the base node to farther meter in the same feeder would be too weak, the switch node, appointed by the base one, will retransmit the message to the terminal node with a stronger signal power. The terminal switches will not take part on this project since the nodes selected do not need one for communicating, since they are not that far away.

1.2.2 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is used in two circumstances: the first one is when the objective is to reduce the number of variables, for that reason new ones are created by being linear combinations of the originals; the second one is for changing the main axes to others that makes more visible the differences between the values represented. Both objectives are achieved at the same time.

For a more detailed explanation, those new axes are the vectors in which the data variance is maximized. These vectors can be used for reducing the number of variables of the system without losing too much information or simply rotate the axes for a better representation of the data studied.

In the case of this project, since the objective is to cluster some data into different groups from an indefinite number of variables, the purpose of the PCA will be to reduce the number of variables without losing much of the explained variance. The variables used are partially correlated, so with one or two principal components the variance explanation will be higher than 90-95%.

The system used for the PCA analysis has been Matlab, with the "pca" function, as well as for the clustering previously mentioned with the "kmeans" method. The complete script of the PCA, clustering and classification program is on the Annex II, sections 9.2 and 9.4.

Chapter 2. STATE OF THE ART

The use of PLC communication, and more specifically the FCC band, in the LV grid is not a totally mature subject. This gap of knowledge can be reduced doing a literature review and connecting the different types of analysis, extracting higher level conclusions useful for the sector.

This state-of-the-art analysis is organized in four sections that summarizes: impedance and attenuation, noise, communication performance previous research, and offers conclusions. Impedance and noise sections extract valuable concepts from the literature while the communication performance tries to validate those concepts with some field trials. Finally, some applicable conclusions are obtained from the review.

2.1 IMPEDANCE AND ATTENUATION

This are the most studied characteristics of the LV grid for the FCC band. Both impedance and attenuation are related and have impact on the transmission power needed to reach the most distant nodes and the waveform received. Impedance amplitude and phase change with frequency, typically with an increasing trend in amplitude. Despite this fact, peak values can be reached in lower frequencies because of resonances. Figure 1 shows the access impedance for three transformer stations, two urbans and one rural.

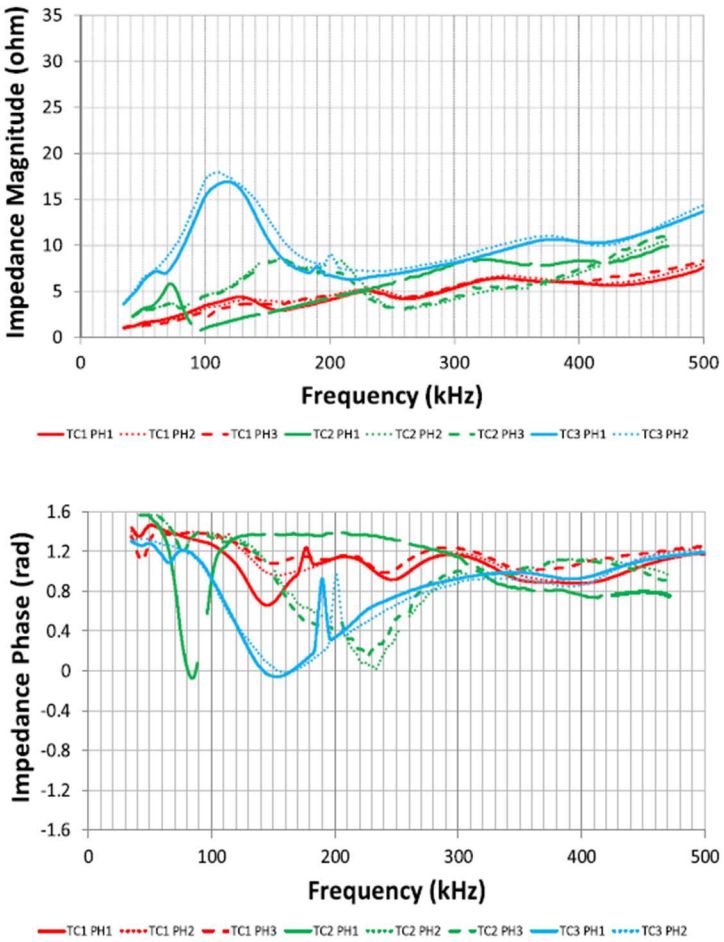


Figure 1: Access impedance in transformer stations for three locations: Urban-1 (TC1), Urban-2 (TC2) and Rural-1 (TC3) [2]

As it was commented before, the amplitude's general trend is to increase with frequency, while the phase is mainly inductive. The second urban SS has a big difference between its phases PH1 with respect to PH2 and PH3. This difference should come due to the number of outgoing feeders and the loads connected to them, since the access impedance is calculated as the parallel between the access connection of each feeder, which reduces the impedance value.

The rural area will suffer the highest impedance since the number of feeders is reduced and their length is longer. The effects of the distance from the transformer station are shown in Figure 2 , taken from reference [3]. This paper focusses its attention on one rural area and how much the distance and the noise in the communication channel affect

performance in each channel. This figure highlights how the distance affects the attenuation for all frequencies but predominantly the upper channels.

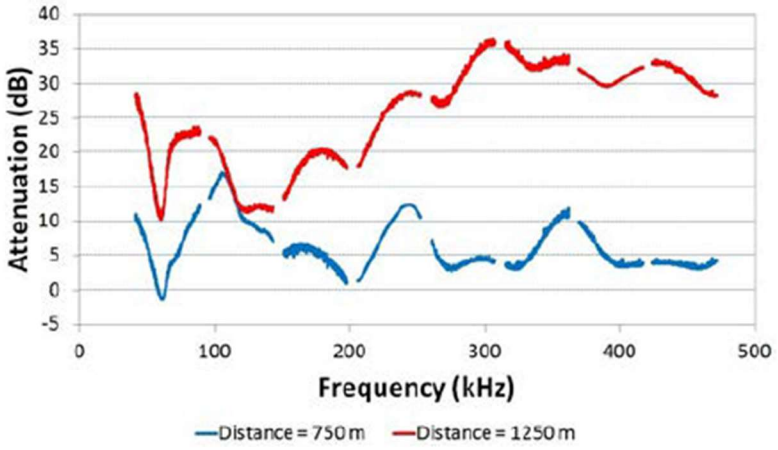


Figure 2: Transmission Losses in Rural Area [3]

Although the distance affects more the attenuation of the upper part of the spectrum, there is not a clear correlation between attenuation and frequency. The non-constant attenuation level will distort the signal, that along with the noise, will cause difficulties for the proper communication. In Figure 3, there is not much general attenuation, compared to Figure 5 that is the one transmitted, but the distortion is suffered in most of the channels, being the upper ones the less affected.

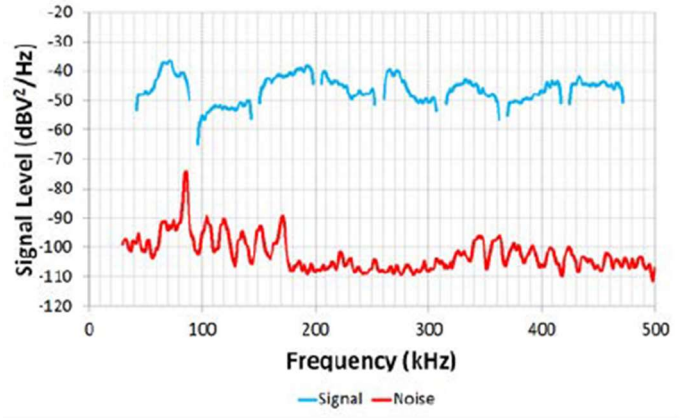


Figure 3: Signal and Noise level at 750m distance [3]

Signal in Figure 4 is much more attenuated, being the SNR of the upper channels low. In this case middle channels, like the third one will be more optimal. Channel 1 is neither attenuated but it is more distorted.

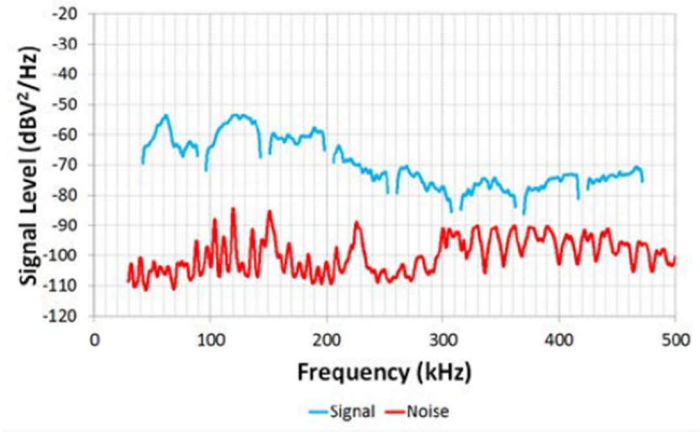


Figure 4: Signal and Noise level at 1250m distance [3]

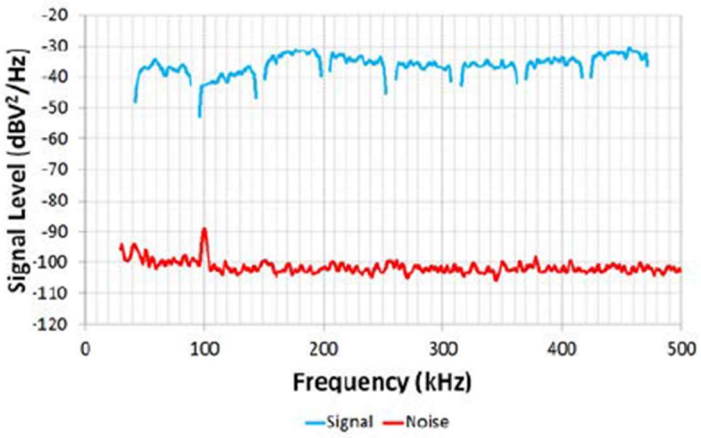


Figure 5: Signal and Noise level at transmission point [3]

Another concept to consider is the variability in time of the impedance. In reference [2], there is a comparison between the access impedance in two different days at the same spot. Figure 6 shows how, for a common urban line, exists a high relative variance in amplitude and phase, that imply a need for adaption to the changing grid values.

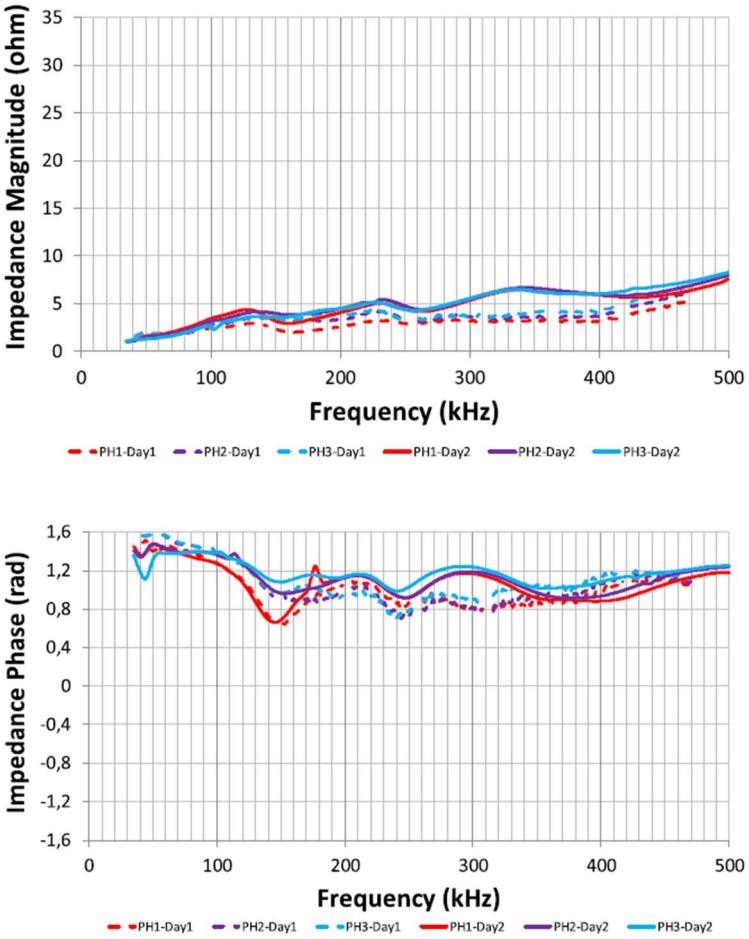


Figure 6: Grid impedance variability for two different days [2]

2.2 NOISE

Noise suffers a strong time variability as well. In Table 2 and Table 3 there is noise power data from transformer stations and customer meters, respectively, located in Doha, Qatar . These tables show noise power measures along the different days of the week and for every channel.

Table 2: Average noise power at transformer stations [4]

Channel	All CH	CH 1	CH 2	CH 3	CH 4	CH 5	CH 6	CH 7	CH 8
Total	27.46	34.45	27.69	28.40	28.29	28.32	24.48	22.77	21.58
Day	27.30	36.78	32.13	30.74	33.31	31.39	26.26	23.96	23.26
Night	27.42	33.24	25.44	25.92	27.64	27.48	23.05	22.12	21.69
Weekday	27.54	33.38	25.52	25.64	28.03	27.35	23.54	22.43	21.57
Weekend	27.21	33.03	25.15	25.30	28.43	27.11	22.42	21.62	21.69

Table 3: Average noise power at customer meters [4]

Channel	All CH	CH 1	CH 2	CH 3	CH 4	CH 5	CH 6	CH 7	CH 8
Total	16.01	27.49	17.00	12.78	11.55	12.28	12.89	12.69	12.56
Day	16.09	27.29	17.37	12.79	11.69	12.49	12.98	13.08	12.95
Night	14.16	26.22	14.45	10.43	9.54	9.95	11.20	11.72	12.23
Weekday	15.97	27.23	16.78	12.56	11.45	12.24	12.95	12.79	12.53
Weekend	16.16	28.04	16.92	13.02	11.97	13.12	13.82	12.84	12.54

There are two main insights extracted from the data: there is a descendent trend of noise power, reducing with the increase of frequency, reaching -13 dBuV difference between the 1st and 8th channels; the time of the day or the day of the week means an offset in the trend.

In Figure 7, it is represented the average noise power with the signal attenuation in two scenarios: low and high attenuation. With this graph, the two main cases of selecting upper of lower channels can be shown clearly.

In the scenario of lower attenuation, since the signal power remains constant for all channels, the noise power becomes predominant in the study of the receiving signal. As shown before, noise power is greater in lower channels so the upper ones will be more likely to achieve a better performance. On the other hand, with strong channel attenuation, it becomes more significant than the noise and so, lower channels are then more likely to perform better.

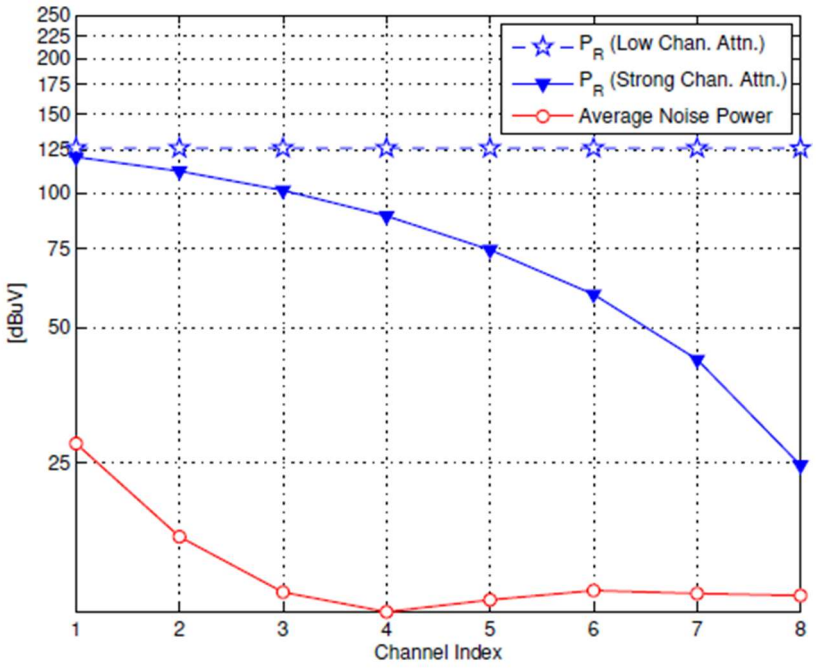


Figure 7: Variations on average noise and signal power at customers meters [4]

In order to support the conclusions reached with the data shown above, in reference [5] there is data that shows a similar trend to the previous one. Figure 8 shows an almost decreasing monotonous noise spectrum, very similar to the previous. Also, in Figure 9 can be appreciated that exist a broad variability depending on the time and location along the line.

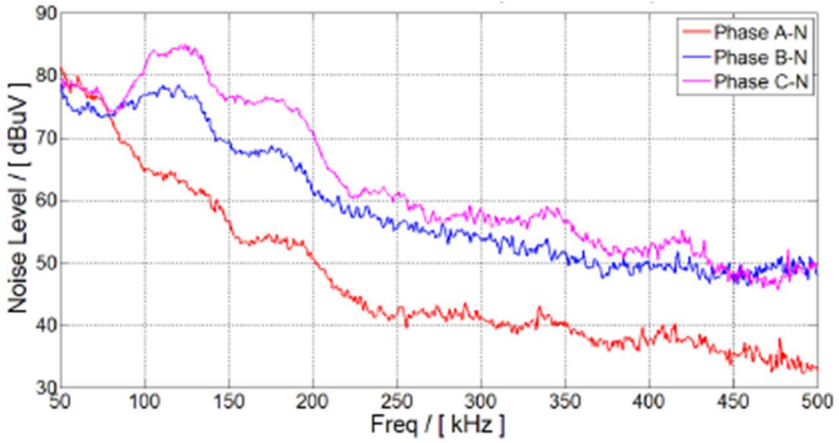


Figure 8: Noise power spectrum measured at the transformer [5]

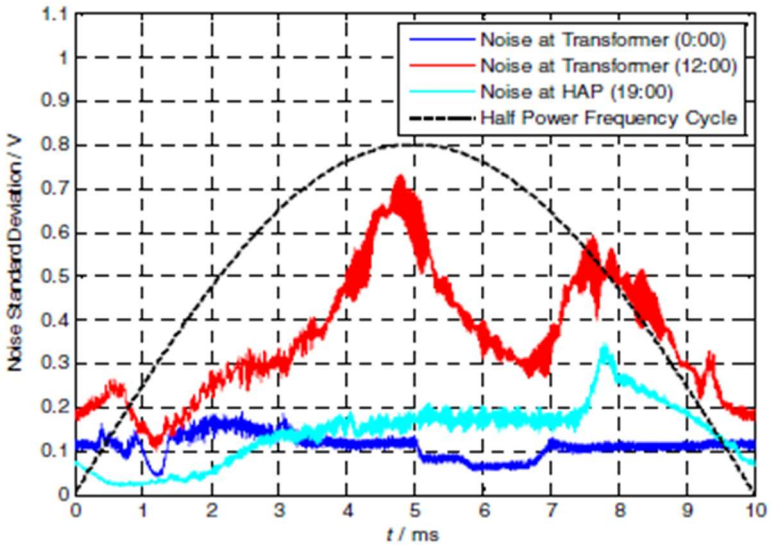


Figure 9: Standard deviation of the noise over half power frequency cycle [5]

About noise topic, there are studies of non-intentional emissions from different devices. Reference [6] focuses in the emissions of different distributed energy resources, which are being deployed all along the LV grid. Among them, there are two that have a relevant noise power in the upper part of the spectrum, that usually is mostly free of high noise power. These two are the hydropower pump and the wind turbine, which emission spectrum is shown in Figure 10 and Figure 11 respectively.

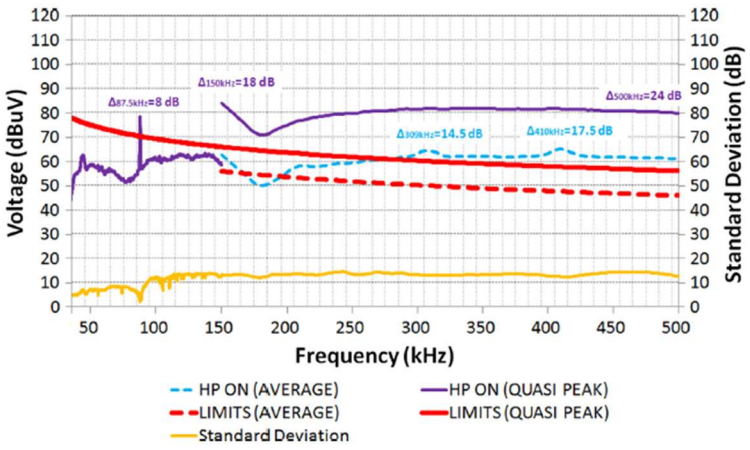


Figure 10: Voltage levels and standard deviation of non-intentional emissions generated by hydro pump [6]

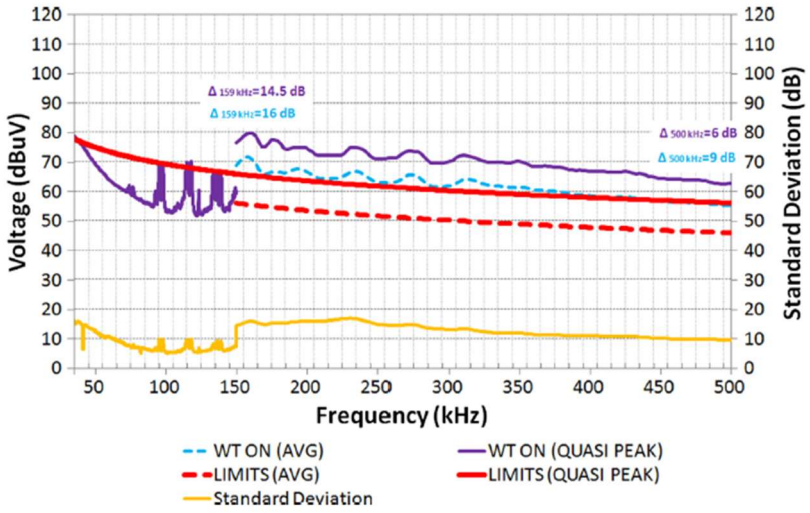


Figure 11: Voltage levels and standard deviation of non-intentional emissions generated by wind turbine [6]

Both graphs show a mostly constant noise power in the whole spectrum, which will worsen the upper channel performance since they usually have higher attenuation. In this paper, there are more DERs that can be problematic in other parts of the spectrum, like the PVs that generates spikes in low-mid frequencies, but it is a more common type of noise.

2.3 COMMUNICATION PERFORMANCE

To verify the hypothesis about the communication drawn for the impedance and noise data, some papers has done empirical trials of communication. In reference [7], they have sent a hundred packets for every channel in each phase, counting the packets collected and the ones received correctly. Besides, they measure the noise and calculate the SNR and the EVM for each channel and phase. SNR (Signal to Noise Ratio) expresses the difference in dB between the power of the signal received and the noise power level. EVM (Error Vector Magnitude) is the error between the ideal constellation point of the modulation and the point received in dBs.

The packet data and noise of Figure 12 have been obtained at a fuse box, 63 meters away from the transformer station. The number of packets received correctly is good in most pairs

of channel and phase, being worse or even unable to communicate when the noise graph suffers from significant spikes.

Channel	Phase	Mean SNR (dB)	Mean EVM (dB)	Packets received	Packets OK
3	L1	15.52	-12.52	100	100
3	L2	15.51	-12.51	100	100
3	L3	15.33	-12.33	99	99
4	L1	15.66	-12.66	100	100
4	L2	15.03	-12.03	100	100
4	L3	14.72	-11.72	99	99
5	L1	15.74	-12.74	100	100
5	L2	15.88	-12.88	99	97
5	L3	8.59	-5.59	99	99
6	L1	24.43	-21.43	100	100
6	L2	3.79	-0.79	7	0
6	L3	0.00	0.00	0	0
7	L1	25.40	-22.40	100	100
7	L2	21.35	-18.35	100	99
7	L3	17.12	-14.12	100	97
8	L1	24.55	-21.55	100	100
8	L2	20.75	-17.75	100	95
8	L3	13.10	-10.10	100	99

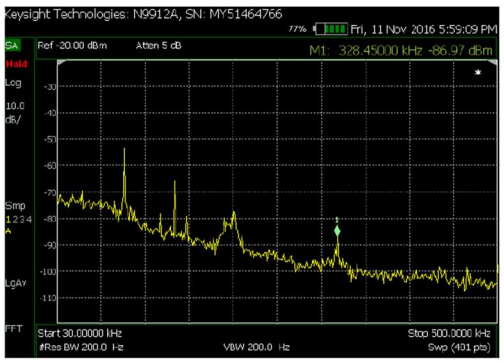


Figure 12: Performance data and noise spectrum from Gernika Lorateguia 11 [7]

Most channels are communicating satisfactorily, except for two phases of channel 6 that was not capable to communicate. The SNR is relevant and correlated with the proper reception when substantial differences appear, while not so useful to distinguish smaller changes in performance. Taking this example, the upper channels, except for the sixth, will be optimal to communicate since they are receiving an adequate signal level and lower noise power.

Figure 13 shows the data from a fuse box further from the transformer station, about 208 meters. In this location, the noise power graph is more favorable, but the signal attenuation makes difficult to communicate in the upper channels.

Channel	Phase	Mean SNR (dB)	Mean EVM (dB)	Packets received	Packects OK
3	L1	14.84	-11.84	100	100
3	L2	15.03	-12.03	100	100
3	L3	15.08	-12.08	100	99
4	L1	11.16	-8.16	100	100
4	L2	14.23	-11.23	100	100
4	L3	14.18	-11.18	100	98
5	L1	10.86	-7.86	98	91
5	L2	15.26	-12.26	99	99
5	L3	13.40	-10.40	99	99
6	L1	-	-	0	0
6	L2	-	-	0	0
6	L3	-	-	0	0
7	L1	-	-	0	0
7	L2	4.73	-1.73	13	0
7	L3	4.45	-1.45	42	2
8	L1	4.81	-1.81	82	3
8	L2	4.23	-1.23	64	0
8	L3	4.58	-1.58	3	0

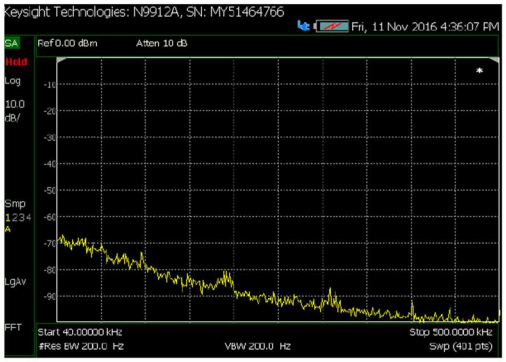


Figure 13: Performance data and noise spectrum from Zugazagoitia 6 [7]

All the channels and phases above channel 5 are unable to receive messages consistently, which verifies the assumptions about the adjustment of channel depending on the length of the feeder.

2.4 STATE OF THE ART CONCLUSIONS

- Impedance study reflects that longer distances worsen attenuation in upper channels when distance is increased.
- Noise levels tend to be higher in lower channels when the line is affected by typical loads, other can generate a more colored noise.
- Urban and rural lines can be a proper approach of classification since rural ones tend to have less loads and longer lines while urban lines are shorter and heavy loaded. For this reason, in rural areas lower channels would be more suitable while in urban will be the upper ones.
- Besides the previous statement, papers reflect the importance of the particular location and environment over the previous division, since the frequency specific noise levels and the fluctuation of impedance and noise over the day an week could be more relevant than the urban or rural classification.

- The main conclusion is the need for adaptation, since the communication have to adapt to the particularities of the grid at each point and time for reaching the better performance levels.
- Nowadays, the best option to identify the suitability of the channels is to carry out impedance, noise, and performance analysis during a period of time and not just punctually.
- Another question to bear in mind is the throughput, besides robustness, taking into account that upper channels have higher throughput [8].

Chapter 3. PROJECT DEFINITION

3.1 MOTIVATION

I-DE Redes Eléctricas Inteligentes, S.A.U. has renewed its determination on achieving a more intelligent and connected grid. With this objective, I-DE with other companies in the PRIME Alliance developed a Narrow Band Power Line Communication (NB PLC) protocol. This protocol allows an improved communication with the deployed smart meters, being able to accomplish a higher transmission of information with lower error rate. The newer version of the PRIME protocol (v1.4) increases the number of channels available from the initial channels 1 originally to the 1-8, increasing the frequency range up to 471kHz, although channel 2 is not applicable in Europe due to the limitations of the UNE-EN 50065-1-2012 [9]

The knowledge gap of the performance of PLC communication in these new channels (3-8) is the main motivation for this project. With the proper understanding of the PLC situation, it is possible to focus on the keys aspects to achieve the best communication possible.

With the performance data, it will be able to cluster the different feeders depending on how well they work in each channel, communicating in the best one for each group. Besides the ratios of performance, it would also be important to match the grid topology with the channel selection preference, in order to extrapolate to other feeders which topology is known but not able to test the channels.

3.2 OBJECTIVES

The main objectives of this projects are: first gathering the maximum information possible of the frequency spectrum that is subject to study and the context of the communication; then do an autonomous performance analysis of the channels for different location, trying to extract some pattern of low performance from different topology variables.

1. Deep study of the FCC Band for PLC
2. Topological study of Iberdrola's grid
3. NB-PLC channels classification
4. Correlation study between channel performance and network topology

3.3 METHODOLOGY

To accomplish the final objectives, a set of tasks should be carried out. These tasks will be more specific with the advance of the project, focussing on extracting practical knowledge. In the following subsections, each of these tasks will be defined to have an overall understanding of the process followed.

3.3.1 STATE OF THE ART RESEARCH

The tasks involved in the state-of-the-art research will be the different topics addressed in when looking for information about the FCC band characteristics.

- **PRIME v1.4**: it is the base for the project, know exactly the changes of the protocol from the PRIME v1.3.6
- **Impedance:** reading of papers about the main property that affects attenuation and signal power.
- **Noise:** study of the non-intentional emissions in the FCC spectrum that affects the proper reception of messages
- **Performance:** search of field studies of communication in these bands to compare with the one done in this project.

3.3.2 GRID RESEARCH

The grid research tasks involve two goals, the description of the grid topology characteristics but also the acquisition of the data used in subsequent tasks.

- **Topology:** an extensive knowledge of the LV grid topology would help to identify the elements that can affect communication, which also will be helpful for the future possible correlation between topology and performance.
- **Data Acquisition:** obtaining information and determine how useful it is for the objectives of the project.

3.3.3 RESULTS ANALYSIS

The result analysis jobs are the three main tasks of the own research, these will be proper presentation of the data already acquired and the two data driven methods carried out for achieving the objectives, the clustering and the classification.

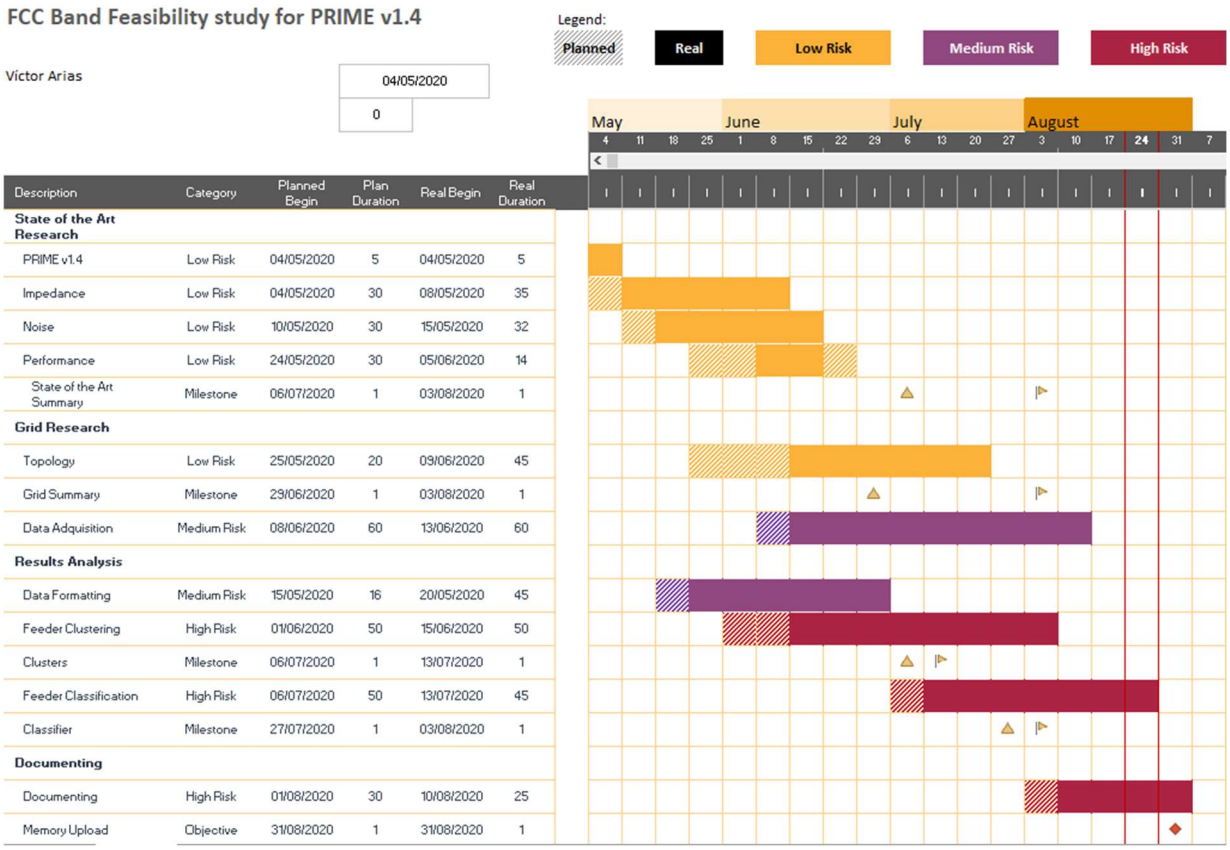
- **Data Formatting:** automation of the raw data conversion to a human-understandable information, as well as format it to be readable by the analysis programs.
- **Feeder Clustering:** program different methods to find the proper clustering of the elements measured
- **Feeder Classification:** combine the clustering with the topological information to search a possible correlation that will make possible the classification.

3.3.4 DOCUMENTING

Writing the memory will be the final task, and an important one. This is because it works as a summary of all the progress done along the project, from the information sources to the results analysis, but also about the insights extracted. However, it is not a final document since it gives the next steps to take for the continuity of the development.

3.4 PLANNING AND ECONOMIC ESTIMATE

The chronogram showed in Figure 14 displays the evolution of the project, being able to see the differences between the planned beginning dates and duration with the accomplished ones. Generally, it exists a time delay between the planned and real beginning dates of the tasks due to some mismatch on the time needed for the completion of initial jobs.



Description	Category	Planned Begin	Plan Duration	Real Begin	Real Duration
State of the Art Research					
PRIME v1.4	Low Risk	04/05/2020	5	04/05/2020	5
Impedance	Low Risk	04/05/2020	30	08/05/2020	35
Noise	Low Risk	10/05/2020	30	15/05/2020	32
Performance	Low Risk	24/05/2020	30	05/06/2020	14
State of the Art Summary	Milestone	06/07/2020	1	03/08/2020	1
Grid Research					
Topology	Low Risk	25/05/2020	20	09/06/2020	45
Grid Summary	Milestone	29/06/2020	1	03/08/2020	1
Data Adquisition	Medium Risk	08/06/2020	60	13/06/2020	60
Results Analysis					
Data Formatting	Medium Risk	15/05/2020	16	20/05/2020	45
Feeder Clustering	High Risk	01/06/2020	50	15/06/2020	50
Clusters	Milestone	06/07/2020	1	13/07/2020	1
Feeder Classification	High Risk	06/07/2020	50	13/07/2020	45
Classifier	Milestone	27/07/2020	1	03/08/2020	1
Documenting					
Documenting	High Risk	01/08/2020	30	10/08/2020	25
Memory Upload	Objective	31/08/2020	1	31/08/2020	1

Figure 14: Project chronogram with tasks and their duration

The schedule highlights with different colours the risk of the task for the attainment of the final objective. The state-of-the-art tasks are low risk for two main reasons, the first the existing time after them until the memory delivery; and the second is the low dependence with the following tasks. The data acquisition and formatting jobs are not high risk partially for the time after them but mainly due to the need of a partial conclusion but not a dependency of a total completion. On the other hand, the clustering and the classifier are high risk since they do not have time to recover from delays and are the main objectives of the project.

3.4.1 ECONOMIC ESTIMATE

Due to the project conception and the teleworking context, the only two cost categories will the hardware renewal and the labour time. The first category consists of just one expenditure, the purchase of a computer with a cost of 1200€. On the other hand, the labour time will be

the sum of working hours of a junior engineer along the whole project. A simplified operation will be full-day job 5 days a week for 18 weeks:

$$Labour\ Cost = \frac{8\ h}{day} \cdot \frac{5\ days}{week} \cdot 18\ weeks \cdot 12€/h = 720h \cdot 12€/h = 8640€$$

$$Total\ Cost = Labour\ Cost + Harware\ Renewal = 8640€ + 1200€ = 9840€$$

Chapter 4. PROJECT DEVELOPMENT

4.1 GRID TOPOLOGY

In the distribution grid there is a main criterion to divide itself that would be the voltage of the line, being high voltage above 10 kV and low voltage at 400 / 230 V. Besides, high voltage grid divides itself in medium (below 20 kV) and high (30 kV).

LV grid begins where the transformer sets the voltage level lower than 400 V and ends when the customer consumption is reached. There are different devices involved in the operation of the LV grid.

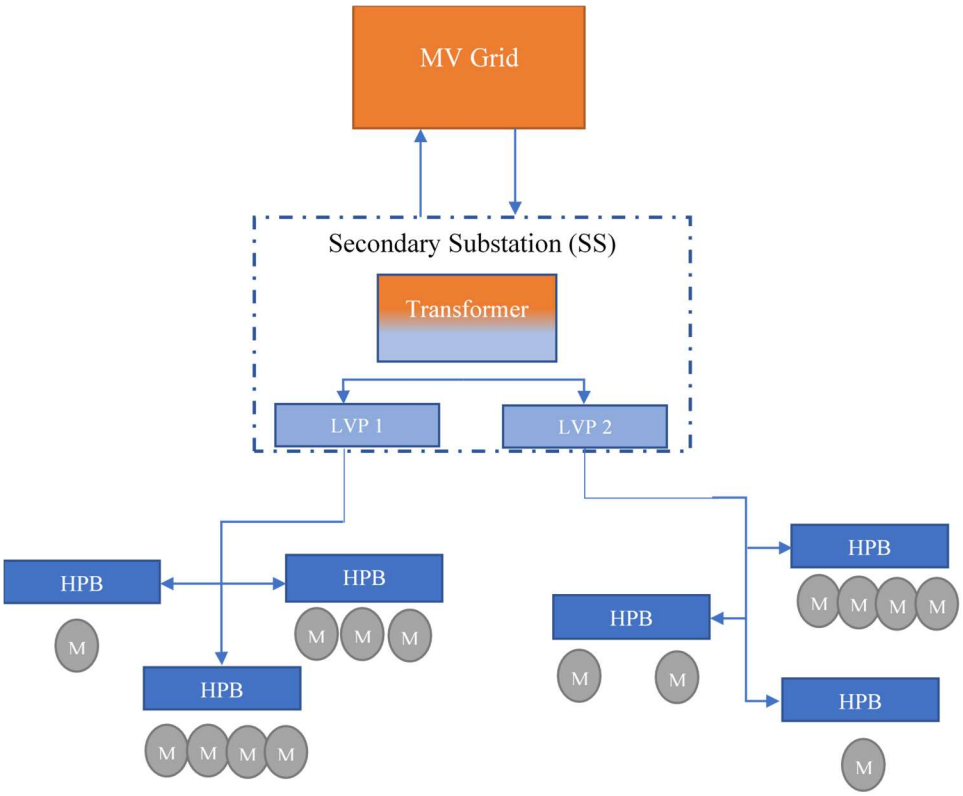


Figure 15 LV Grid General Description

The first section of the LV grid is where the voltage is lowered from the MV grid, that is the transformer inside the SS. Then the Low Voltage Panel (LVP) comes where is also located a basic protection cabinet is also located. Still in the SS, there are the remote management and communication cabinets, connected to the low voltage panel.

Once the feeders are routed out from the transformer station, they are channelled to the house connection boxes. The last device to be connected to the grid would be the meters for the final customer that could be in the house connection box, although typically are located inside the customer building.

4.1.1 CONNECTION TO MV/HV

The connection between the MV/HV and the LV grid is done by a SS. This SS will include the protection systems, the transformer, metering systems, the LVP and communication panels. The urban SS are usually connected in a ring, being able to be fed from two different lines, achieving the n-1 safety goal. There are other sections of the grid that have a more radial shape, being unable to be served in case of a line fault.

In the SS, after the entrance and exit units for MV, the switchgear unit protects the medium voltage line from any trouble in the substation or below. The type of protection will depend on the load and the type and ownership of the SS. The transformer is the following device, where is the real connection between both grids, MV and LV.

4.1.2 LOW VOLTAGE PANEL

Once the voltage is lowered by the transformer, it is followed by the connection to the low voltage panel. There are two models of LVP, one for panels for compact outdoor transformer stations and other for indoor, this last one will vary depending on the allocated current.

Table 4: Normalized LVP elements [10]

Designación	Corriente asignada A	Tensión asignada V	N° Salidas	Tensión soportada a frecuencia industrial Valor eficaz kV		Tensión soportada a impulsos tipo rayo Valor cresta kV	Código
				partes activas y masa*	partes activas	partes activas y masa (*)	
CBTIC-EA-ST-SL-400	400	440	2(**)	10	2,5	20	5044060
CBTC-EAS-ST-SL-400	400	440	3(***)				5044062
CBTC-EAS-ST-SL-630	630	440	4(***)				5044063
CBTC-EAS-ST-SL-1000	1000	440	5				5044064

The main functionalities of the LVP are entry-sectioning function, bus-bar function, protection function, auxiliar entry function, and function of feeding and controlling of supervising and remote management devices. LVPs will oversee the protection of the transformer station from faults in the feeders as well as allowing other connected systems to operate by feeding and allocating their devices.

Some of the devices that are connected but not part of the LVP are the remote management, communication, or automation systems [11]. All the represented in Figure 16 is in a SS with three LVP and broad band PLC communication with the other substations.

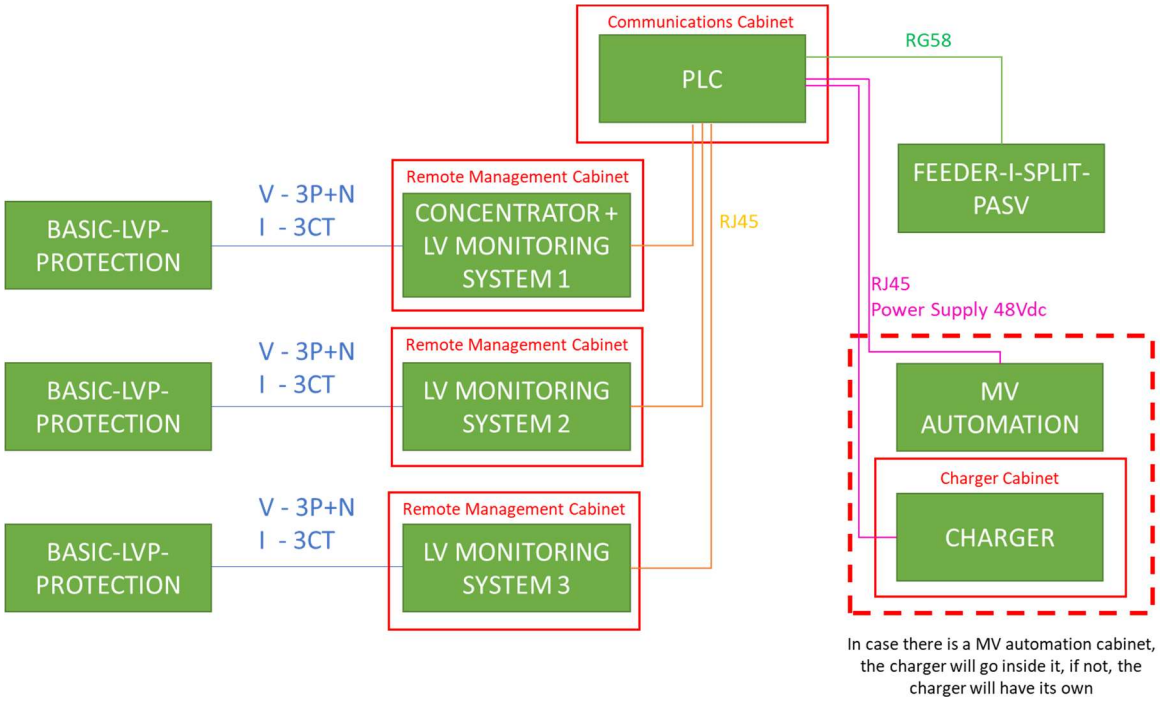


Figure 16:Cabinets connections example acceptable for three LVPs and PLC communication

4.1.3 FEEDERS

The main characteristics of the LV feeders are the type of line, conductor and isolation material and number of cables in these feeders.

The main criteria to divide feeders would be the type of line, being the two possibilities overhead and underground. This document will focus in 400/230V lines, either overhead or underground.

4.1.3.1 Feeder Characteristics LV grid

1. Standard voltage: 230/400 V
2. Nominal frequency: 50 Hz
3. Earthing system: Neutral connected directly to ground
4. Grid cable isolation: 0,6/1 kV
5. Maximum short-circuit current: 50 kA

4.1.3.2 Underground feeders

Most of urban feeders are underground, these cables are aluminium unipolar with cross-linked polyethylene isolation, always using four conductors (3 phases + neutral).

Table 5: Underground feeder types [12]

Cable Type	Phase Conductor (mm^2)	Neutral Conductor (mm^2)	Isolation
RZ1(S)	AL 50-95-150-240	Al 50-95-150	XLPE
RZ1(AS)	AL 50-95-150-240	Al 50-95-150	XLPE

Cables of 150 and 240 mm^2 are for the underground distribution grid, while 95mm^2 cables are used in low and uniform load density and for branching-offs from the feeder and service connections.

4.1.3.3 Overhead feeders

Overhead cables for LV are isolated single-pole stranded conductors that rest on the front of the buildings. In case it is not possible, they would rest on supports.

Table 6: Overhead feeder types [13]

Cable Type	Phase Conductor (mm^2)	Neutral Conductor (mm^2)	Isolation
RZ-25	3x 25 Al	29.5 Alm	XLPE
RZ-50	3x 50 Al	29.5 Alm	XLPE
RZ-95	3x 95 Al	54.6 Alm	XLPE
RZ-150	3x 150 Al	80 Alm	XLPE
	16 Al	16 Al	XLPE
	25 Al	25 Al	XLPE
	3x 16 Al	16 Al	XLPE

Conductors size 150 and 95 mm^2 are used for distribution grid. Specifically, 150 mm^2 conductors are applied for areas where feeders cannot be built underground or for heavy loaded or long lines. For singular points away from the planned outline, and only if can cope

with the demand, 3x25 or 3x 50 mm^2 conductors could be installed. Bipolar and tetrapolar cables are utilized for branching-offs for the House Connection Box (HCB) of fuses box.

4.1.4 HOUSE CONNECTION BOX

It is the device dedicated to accommodate the protection for the feeder. It defines the beginning of the customer property.

If the power load needs it, more HCBs could be installed for the same building. The type of HCB will be determined depending on the feeder characteristics, the load forecast for the feeder and the location. In case there is more than one feeder, each line will be protected independently by one HCB [14].

The HCB will be in the front wall of the building or in the limit of the property, always with permanent access from public road, although the box and the devices inside are owned by the customer. The protection used in the HCB are LV fuses, particularly blade contacts fuses.

The HCB will be connected to the meters' concentrator, a closed premises where all the building customers have their meter and it is only destined to that purpose.

4.1.5 METERS

This is the border device of the grid which is controlled by the distribution company. It receives signals for operation and sends information about the energy consumption of the customer, as well as other grid measures such as voltage and current. This communication is made by narrow band PLC, with the PRIME v1.4 communication protocol. The data mentioned before is sent to the concentrator located in one of the remote management cabinets of the secondary substation, where will be accessible by upper communication devices [15].

The location of the meters inside a building will be in a closed space designated for this purpose, as mentioned in the previous section. This space will be as close to the entrance as possible and must assure a set of conditions to be enabled as meter concentrator, like draining, air circulation or lightning. All the meters of the building will be in this place,

except if the building has enough floors or customers to set another place in another plant. Another exception to this meter centralization will be if there is only one customer, in this case the meter could be set in the house connection box.

4.2 DATA ACQUISITION AND FORMATTING

After the compilation done about the specifications of the grid topology and the research of the required frequency bands, it is time to acquire some data to do an independent analysis. First, the information needed is the performance indicators in each communication channel. Then, some characteristics of the HCB will be necessary, along with some others of the feeder which it is connected to.

4.2.1 PERFORMANCE RESULTS

To detect the optimal channels to communicate to each node, there are some indicators that could be measured and be distinctive of a proper transmission. These indicators should be representative of a correct communication but also be available nowadays. The variables used are the following:

- **Availability:** it reflects the probability to send a message and be received properly. It is an indirect measure, since it is obtained from the number of correct messages sent for acquiring other variables.
- **Received Signal Strength Indication (RSSI):** it represents the relative strength of the signal received from the ideal, the measure is relative to the device manufacturer that will set the scale and then it is referred in decibels referenced to one milliwatt. The higher the value, the better the signal. Sometimes is represented in negative, being 0 the best value possible; in this case, it has been used in positive decibel. This value is related to the attenuation, a characteristic revealed important in the state-of-the-art research.
- **Signal to Noise Ratio (SNR):** This is another decibel measure reflecting the power of the signal received compared to the noise affecting the channel. It is a necessary

indicator observing the importance of noise in the papers read about this specific topic.

The values used are ten minutes means of the three indicators for the time of a day. The devices selected located in the HCB and with direct communication to the secondary substation, avoiding interference of switch nodes.

4.2.1.1 Formatting Performance Data

The data selected is received but it is vast amount of values from which information needs to be extracted. The first goal of the formatting process was to make this amount of numbers readable and understandable by a human. To achieve the target, it could be done manually since at the beginning there were not too many files which values had to be extracted. However, the number of files will be growing in the time this project develops and continuing in the future. For this reason, the automation of the whole process, from extracting the data to the formatting and presentation of the information, was done. The steps automated tracked by the script are the following:

1. **Get data from csv files:** each device belonging to a SS has a file for each indicator.
2. **Calculate summing variables:** sum up the ten minutes values to variables that represents the total amount of variables. These can be the mean, median, variance…
3. **Save all the new information in an Excel file:** Write the variables in sheets of an .xlmx file for an easier access.
4. **Format and present the information:** with the information extracted and calculated, it will be presented in charts and the channels will be ranked. These last step helps for the analysis of the results.

The script has been programmed in Python with the help of visual Basic libraries for the most complex excel tools. In Annex II can be seen the whole script for further information.

4.2.1.2 Script Formatting

After having reached an appropriate formatting for being presented and explained to the people interested in these results, there is the need of a format to be properly taken by the clustering script.

The correct arrangement of files and folders to be iterated and get the desired data from the original set of files were done by another python script. Once all folders and files are settle in a correct order, a Matlab function was programmed for iterating through all the files and doing the specified operation to the 10-minute measures of the nodes in every variable and channel. The calculus of every node was concatenated and returned as a table to the clustering script that call the function.

4.2.2 HCB AND FEEDER CHARACTERISTICS

For most part of the project there is no need for more data, although in the classification of devices with topology characteristics, these will be necessary. The explanation of the data of topological characteristics is simple, since the only processing that must be made is to select the devices used in the process. The device name is the last four values of the MAC address, with them the whole MAC address can be found. The last step would be to select from the list the addresses found and take the rest of characteristics of the file. Those characteristics will be the following:

- **DistanceSS:** feeder length from the SS to the device.
- **TypeBTLine:** if it is overhead, underground, or mixed line.
- **COD_CT:** the company's internal code of the SS.
- **COD_CALLE:** the company's internal code of the street where the device's building is located.
- **NUM_CLIENTES _CGP:** the number of customers in the HCB
- **POT_CONTRATADA_CGP:** the contracted power of the HCB
- **NUM_CLIENTES _CT:** the number of customers in the SS
- **NUM_CLIENTES _TRIF_CT:** the number of customers with three phase consumption in the SS
- **POT_CONTRATADA_CT:** the contracted power of the SS
- **COORD_X:** longitude value of the device
- **COORD_Y:** latitude value of the device

- **COD_LINEA:** the company's internal code of the line where the device is connected.
- **COD_BRIGADA:** the company's internal code of the brigade that do the maintenance of this device.

4.3 CLUSTERING OF DEVICES' PERFORMANCE

The figures showed in sections 4.3 and 4.4 are for the better comprehension of the text, not suited for the analysis of the values presented. The values used for analysis will be displayed and commented in section 5.3 and 5.4 of the results analysis chapter.

Because it is simpler to use fewer variables to group, especially if the originals are related to each other as supposed, a Principal Component Analysis (PCA) is done for each channel. From this analysis it comes out that there is a component that explains almost all the variability of the sample (more than 80%) and that adding another component would explain 10% more but would double the number of variables. Although it is a low explanation increase by a large increase in complexity, could be easier to cluster with two dimensions than with one.

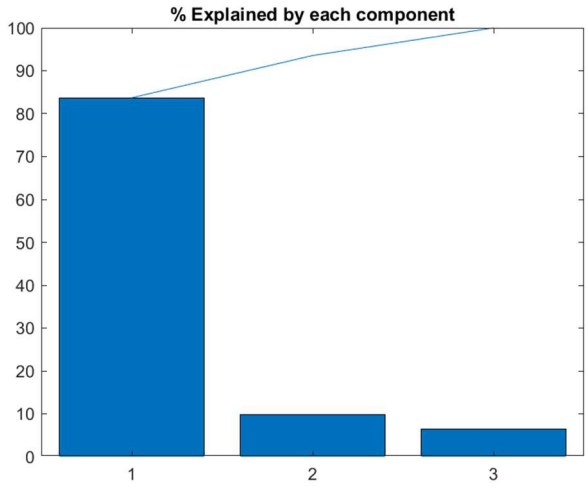


Figure 17: Example of PCA variance explained by each component

This first component is a linear combination of the three performance variables (Availability, RSSI, and SNR) with the coefficients of each shown in Figure 18:

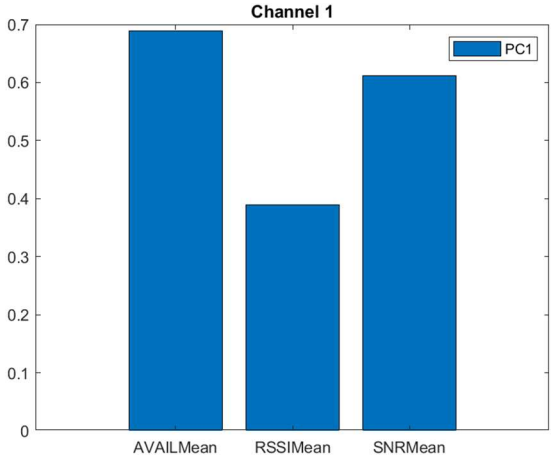


Figure 18: Example of PCA coefficients of the first principal component

The data entered had to be normalized, since each variable had a different range of action. Therefore, the final values that will be used to group the different elements will be 6, 1 per channel and which is the projection on the axis is defined by the coefficients previously obtained. By representing these values now, the nodes would be identified by these values (each line is an element).

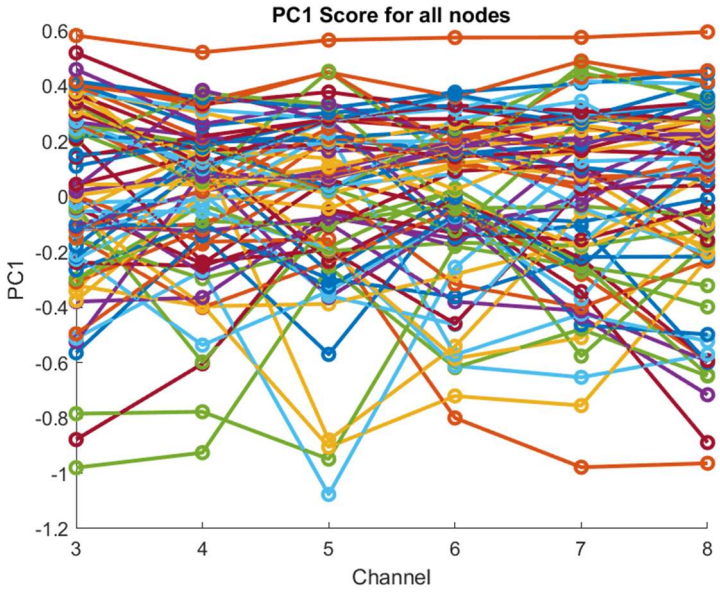


Figure 19: Example of PCA score in each of the six channels for all the devices measured

Already with simplified but representative values, we proceed to group. To estimate what the number of groups would be most appropriate, the quantization error of doing so with

each number of groups should be calculated. The optimal tends to be at the elbow of the curve, where the error suffered is not too large and the groups are not too many to not generating groups with too few elements.

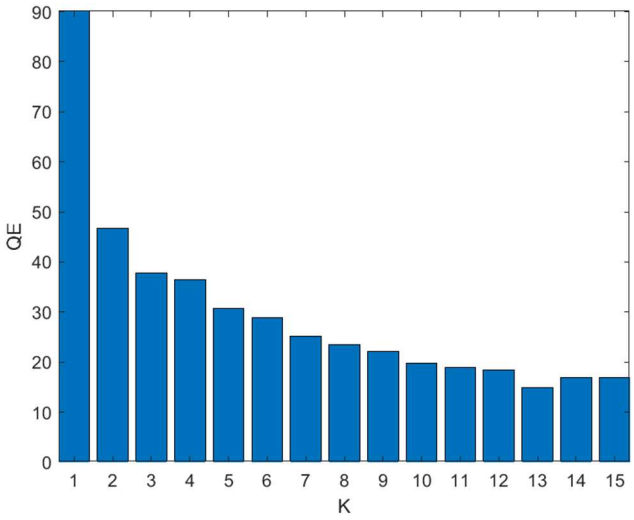


Figure 20: Example of clustering quantization error depending on the number groups desired

By selecting 4 groups, the mean value is represented in Figure 21, although it is not exactly an average value but the center of the Euclidean distances.

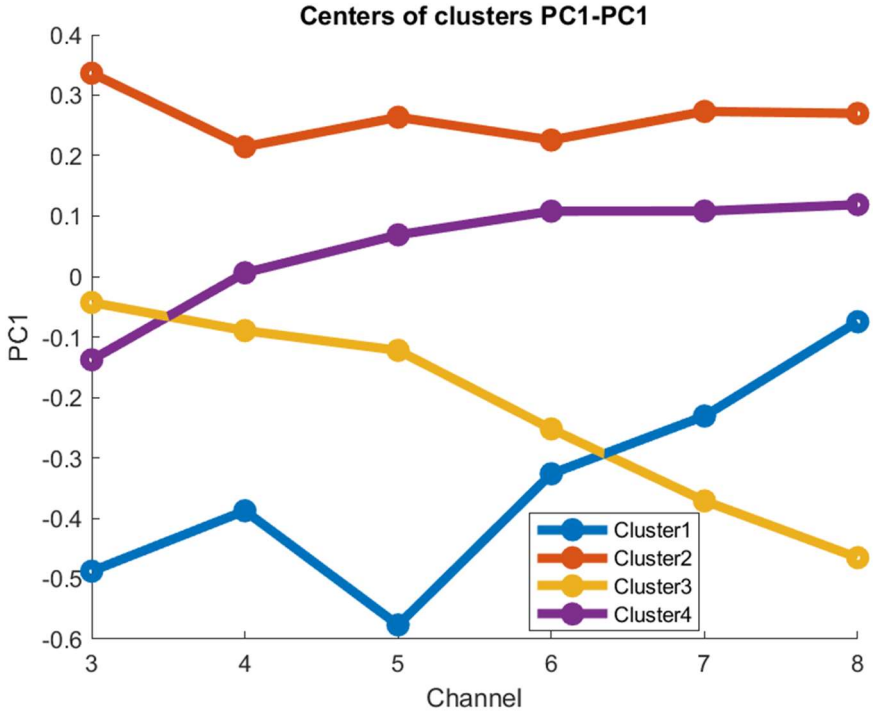


Figure 21: Example of cluster's centres for each channel if four clusters are selected

The values of those centers would be representing how well each channel works in each cluster. From this section we will get to which group belongs each node, trying in the classifier to reach this same result, but through the topological variables.

The sequence explained in this section are the steps required for clustering these variables, although several iterations needs to be done no reach the desired groups. The changes of the iterations can be from the input data, from means to variances of those indicators, to the number of clusters selected, going through the number of components used. The final version of this iteration will be explained in the subsection 5.3 of the results analysis chapter.

4.4 CLASSIFICATION OF FUSES BOXES

The first step is to select the topology variables that you want to include in the classifier. Subsequently, a random separation by code of the nodes is made into two groups, one for classifier training and one for testing with a 70-30% ratio. Once you have both groups, you train a decision tree with the previously split training group. A decision tree has been selected

for this explanation as it is one of the most representative and viewable methods. A cross validation of ten folds is then done, which is to modify the order and selection of that group to be more generic and not overfit a specific dataset. The next step is to check the average importance of the predictors, variables used in the classification. What it takes for important depends to a large extent on how nodes are grouped together and the distribution of groups.

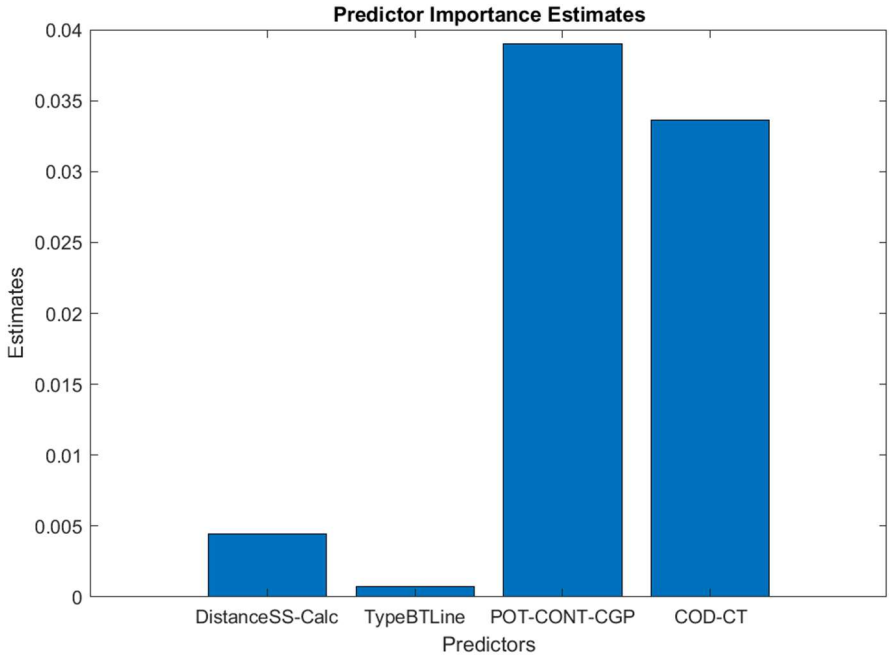


Figure 22: Example of mean importance of predictors used

With the model already trained, the error and accuracy of the model is calculated, which will be done with the training data and then with the test group data. In addition, you can display the decision tree and calculate the number of selected items in each division.

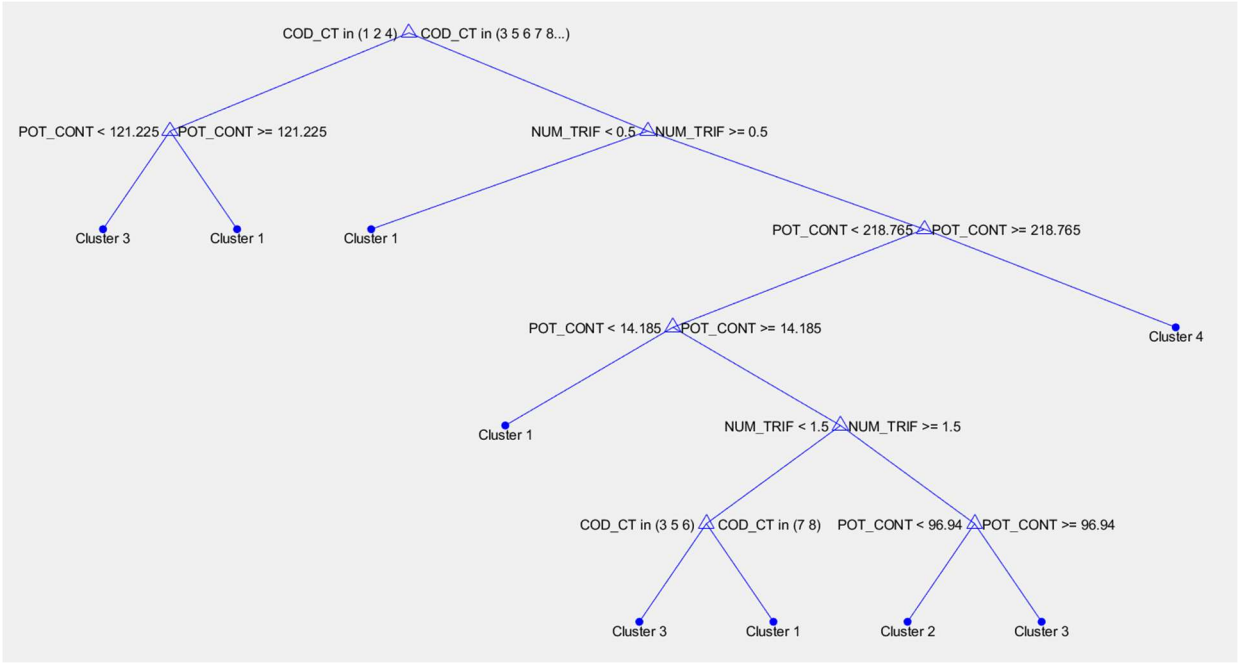


Figure 23: Example of trained decision tree

The COD_CT splits representing the index of the vector of unique values.

Finally, the model correct and errored data is represented for the two groups, train and test, in a confusion matrix. The diagonal represents the correct data, which is the match of the vertical axis that represents the true group, and of the horizontal axis that is the predicted group.

Figure 24: Example of confusion matrix

The optimal classification model and the analysis of the classification results will be commented in the subsection 5.3 of the results analysis chapter.

Chapter 5. RESULTS ANALYSIS

5.1 GRID TOPOLOGY

From the research, apart from knowing how all the LV grid works in the specific company and which are the elements taking part on the PLC communication, there are some other useful insights to be extracted.

The conductor material is the same for all the feeders, the main change between them is the diameter of the cable. For this reason, the properties of the line are mostly the same, which relegate the importance of other characteristics different from the length of the line to the background.

Other information obtained from the description and hierarchy of the grid is which are the shared elements between common nodes since they would be affected by the same incidences and characteristics. This dependence can take part on a performance pattern so relevant for the subsequent steps of the project, that will try to correlate it with a pattern of grid characteristics.

The topology variables used in the classifier will be selected taking into account the expected relevance in the classification based on the research of the LV grid done previously.

5.2 DATA ACQUISITION AND FORMATTING

From the original performance data of the csv files, two type of MS Excel files were created. The first type was a summary file for each SS, where the mean values of availability, RSSI and SNR of every channel of each device connected to that transformer station was showed literally and graphically. In Figure 25, Figure 26 and Figure 27 can be observed the charts of every variable distinguished by channel and device.

Figure 25: Availability summary of the devices in a specific SS

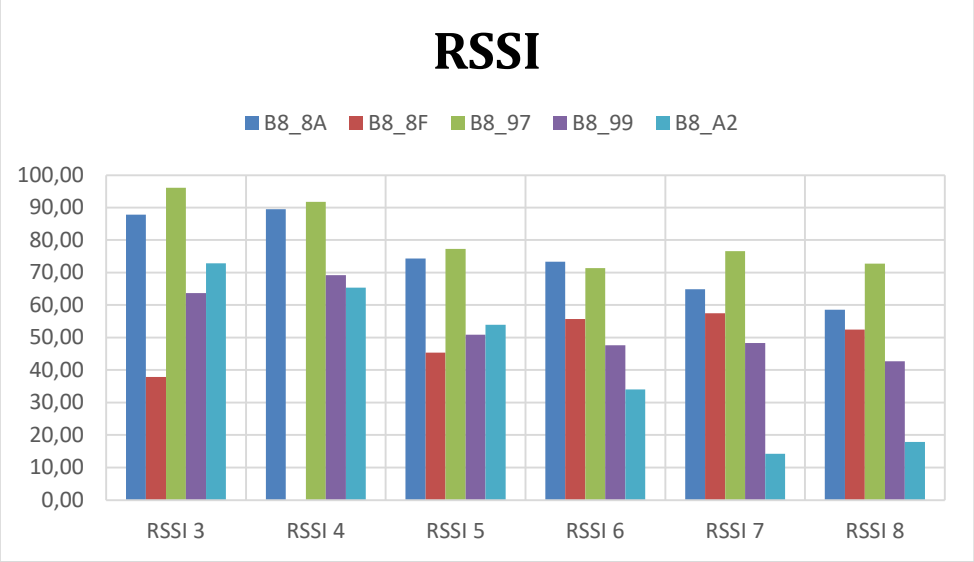


Figure 26: RSSI summary of the devices in a specific SS

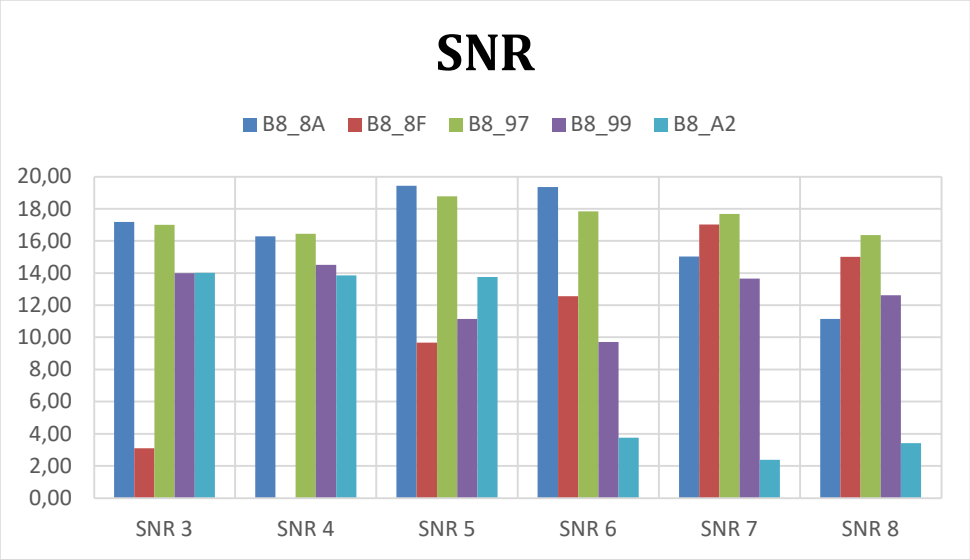


Figure 27: SNR summary of the devices in a specific SS

The second type of file is a summary of all the SS analysed. The mean devices are calculated and shown as the SS results. The comparison between all the transformer stations is presented in a chart comparing also by channel. In Figure 28 is represented the availability chart as an example of the final representation.

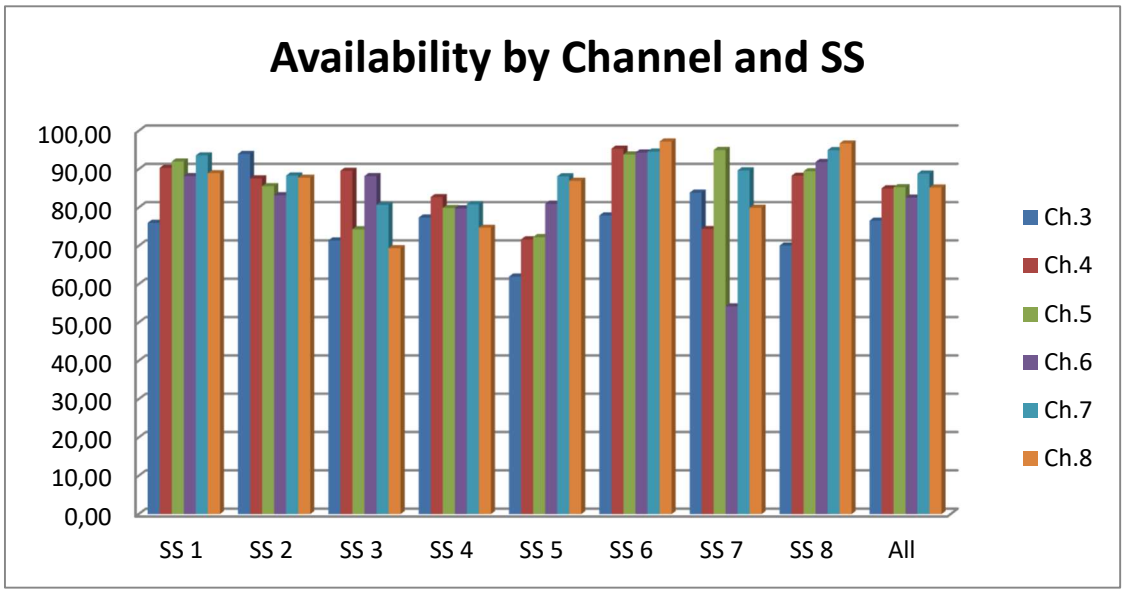


Figure 28: Availability summary chart by channel and SS

As a last simplification of the processed data, and with the purpose of figuring out which are the channels with more availability in mean values, the SS data mean has been calculated resulting in a value for each channel.

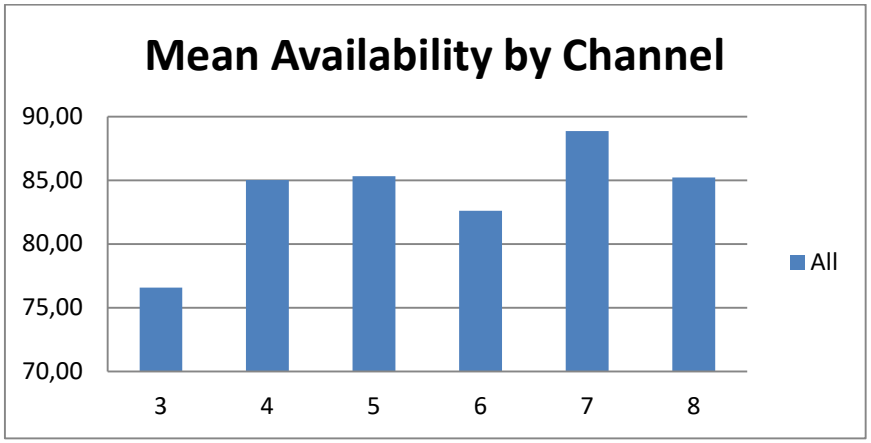


Figure 29: Mean availability by channel

With this dataset, a tendency of better performance in the upper channels is signified not also in median values but also in consistency since the dispersion in the 7th and 8th channels is low. In RSSI variable, the values are higher in the first channels due to the reduced attenuation suffered in the lower part of the frequency spectrum. The SNR tends to depend less on the channel than on the location, that is, all the channels of a specific SS has worse SNR values than in other if the presence of noise is dominant.

5.3 CLUSTERING OF DEVICES' PERFORMANCE

This subsection is the one with more variables involved and the one that can be reached from more approaches. It is going to be divided in three sections: the first one does a PCA to the values used in each channel separately but then using the components of all the channels for clustering; the second one does the same with the PCA but also clusters each channel individually; the last section will include the other approaches that have been tried.

5.3.1 CHANNEL COMBINATION CLUSTERING

As mentioned in the introduction of the clustering section, this approach does a PCA to the performance variables used in each channel, that in the final version has been six:

- Availability mean
- SNR mean
- Availability median
- SNR median
- Availability standard deviation
- SNR standard deviation

The RSSI data has been avoided due to a high dependency and inconsistency when the SNR levels are low. In this case the variables have not been normalized. From these six variables, two principal components have been selected, reaching the vast majority of variance explained. For example, for the sixth channel the variance explained has been 89.83% from the first component and 6.04% from the second, as showed in Figure 30.

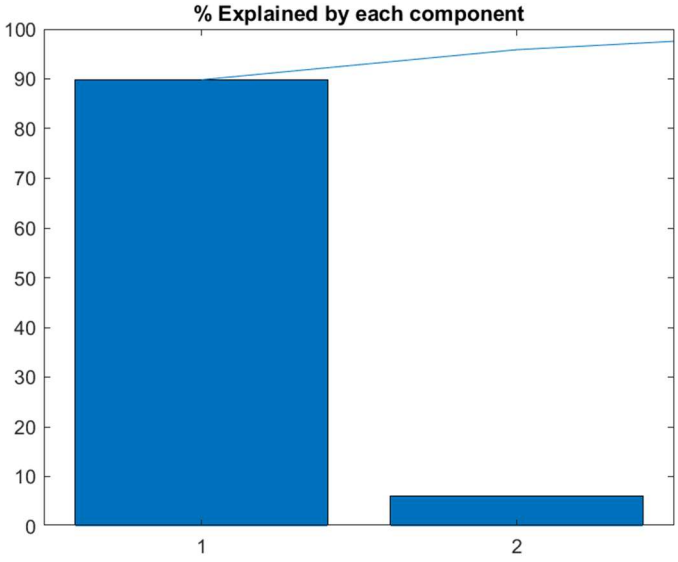


Figure 30: Channel Combination, PCA variance explained for 6th channel

The coefficients of the sixth channel for the two principal components are showed in Figure 31, although it varies from one channel to the other. These coefficients will be necessary for approximately reverse from PC scores to the original variables, not exactly because reducing the number of used variables the not-explained variance information is lost.

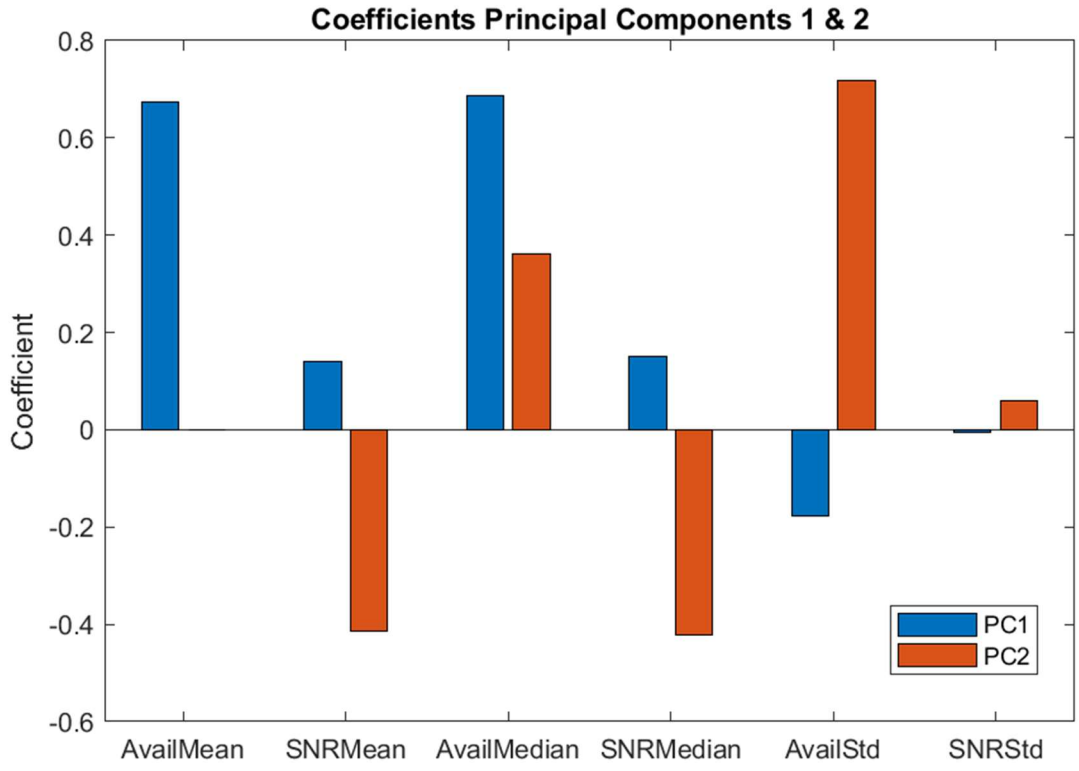


Figure 31: Channels Combination, coefficients of principal components 1 and 2

The first component is more related to the availability mean and median while the second is more general, although with a big influence of the availability standard deviation.

With the scores of the nodes in these new variables, all the terminal values can be represented in a three-dimensional plot, with PC1 and PC2 values in z and y axes respectively and the six channels in x axe. Each colour represents a channel for a better presentation in Figure 32.

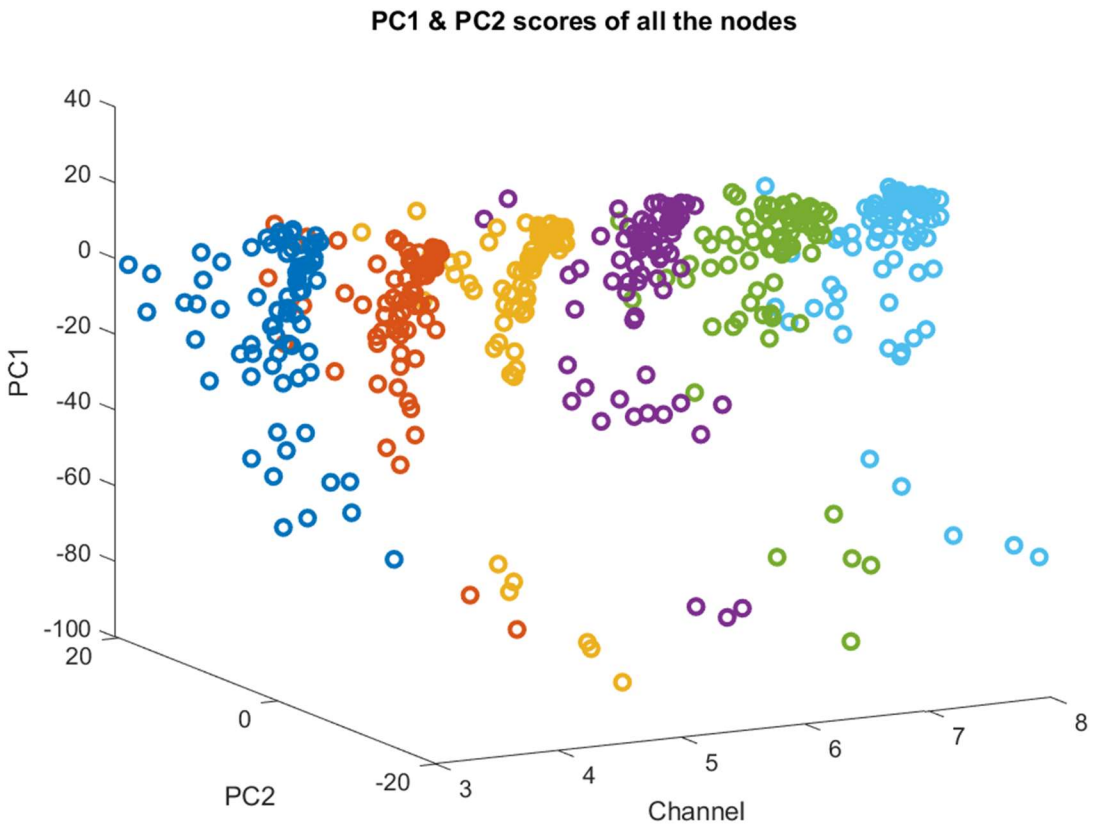


Figure 32: Channel Combination, scores of principal components 1 and 2 for all the nodes

Once the PCA is done, it is time for clustering the nodes by the scores previously represented. The first decision to be made is the number of clusters for an adequate division. The number selected is three, because four makes groups with too few elements while two did not divide in bad and good performance, which would be an interesting option.

With the number of clusters selected, all the components needed for the clustering are presented. The clusters' centres can be represented with different axes and several will be shown in the following figures. The first two shows the first step, the principal components score, PC1 in Figure 33 and PC2 in Figure 34.

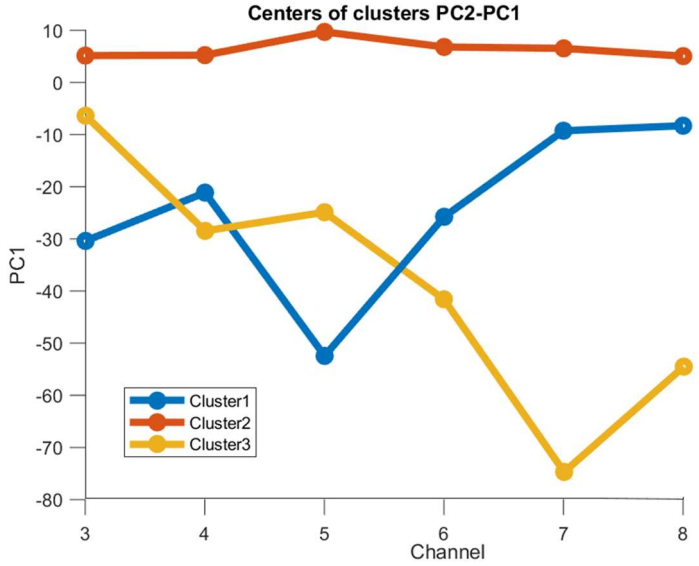


Figure 33: Channels Combination, PC1 score of clusters' centres

Although it is not easy to extract real considerations from principal components scores, since they are combination of different meaningful variables, PC1 was principally related to the availability mean and median. Can be assumed that those two variables figure will be like this one.

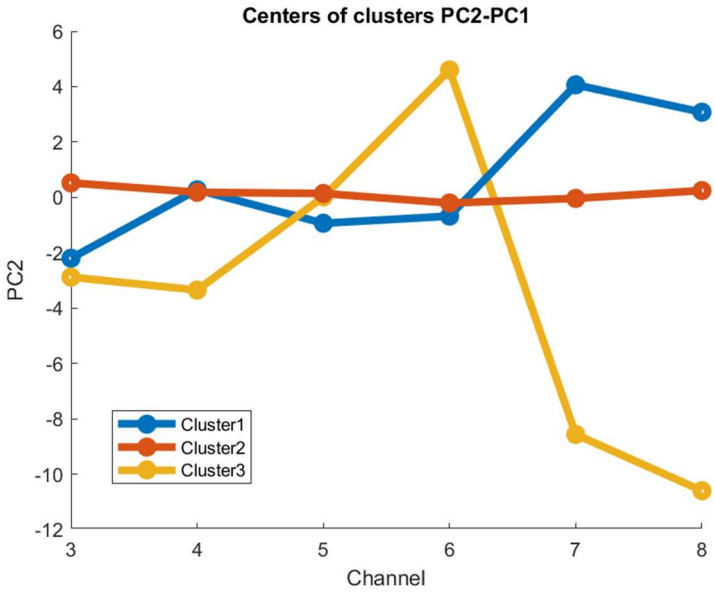


Figure 34: Channels Combination, PC2 scores of clusters' centres

PC2 was more related to SNR and the standard deviation of the availability so it is not so easy to comprehend the meaning directly. However, for that reason the values can be reversed with the coefficients calculated previously, getting the original variables which have a single definition. Figure 35 and Figure 36 represents the median values of availability and SNR respectively, since they are the more interesting variables corresponding to channel performance.

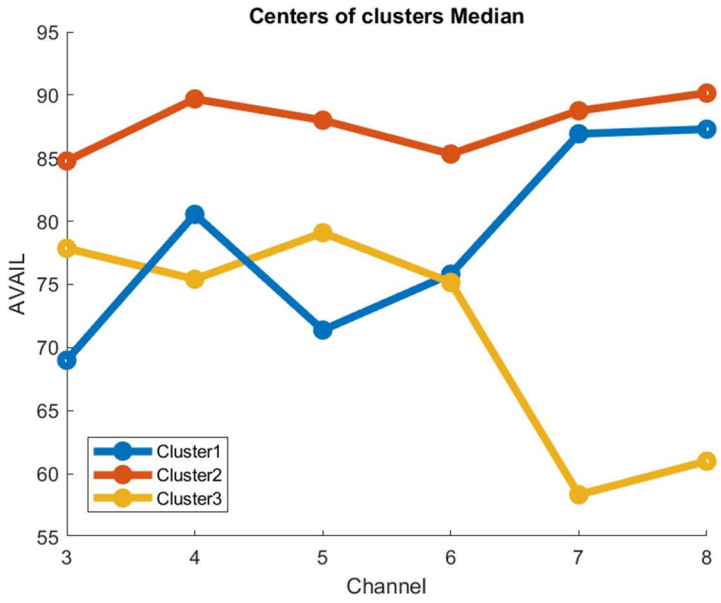


Figure 35: Channels Combination, availability median of clusters' centres

This figure is very similar to Figure 33, changing the y-axe scale, due to the tight relationship between PC1 and median availability. From Figure 35 we can extract that the total mount has been divided in three clear groups: one that works better in upper channels, the second that works the best on every channel, and the third that works better in lower ones although not even really good.

Figure 36 has a remarkable insight, the groups with better median performance in availability do not have better SNR, although in the first two at least the intra cluster tendency matches the availability trend. The third cluster has the better SNR in upper channels that corresponds with a drastic drop in availability, which is counter intuitive and is to be examined more deeply.

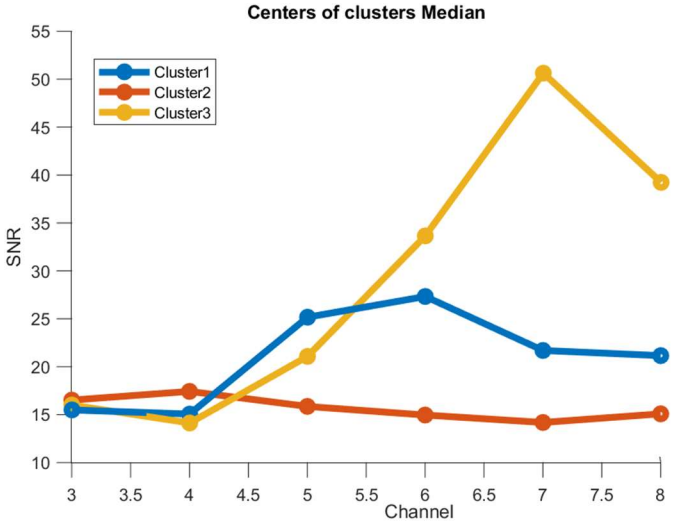


Figure 36; Channels Combination, SNR median of clusters' centres

The last figures have been the centres of the clusters, that are easier to represent and understand, although the real nodes attached to each of these groups are different between them. As an example of the nodes' values inside the clusters, in Figure 37 is represented the availability median values in each cluster.

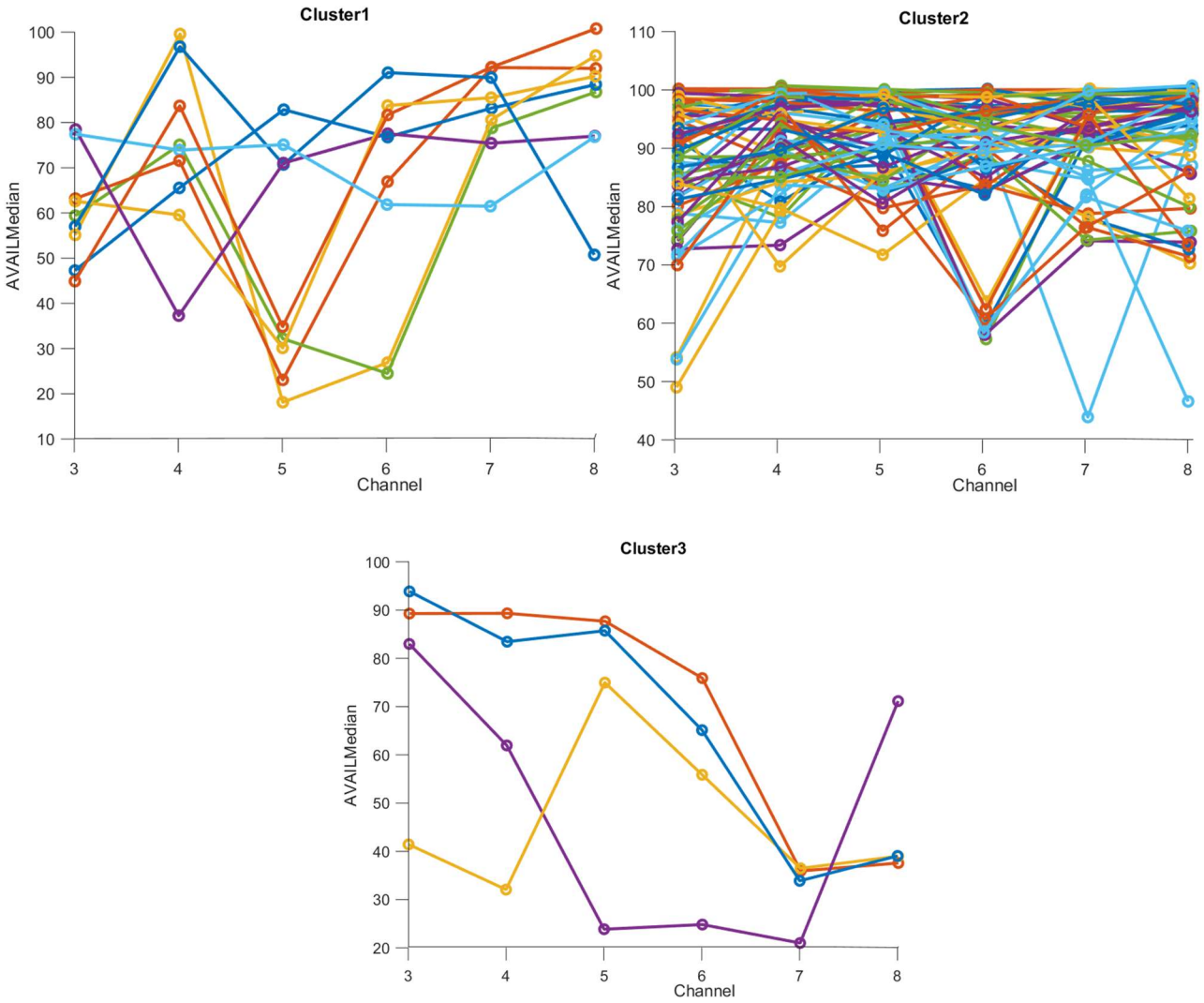


Figure 37: Channel combination, avail median of the nodes in each cluster

After this clustering there is a cluster assigned to each node so it can be identified with the typical cluster performance without the variables needed for the clustering. This information will be passed to the classification script that will try to match these labels with some intrinsic features of the topological variables.

5.3.2 CHANNEL BY CHANNEL CLUSTERING

The other of the main approaches is the one that uses the same PCA as the previous method, extracting the principal components for each channel, but know the clustering will be applied to one channel at a time. This is, there are six clustering processes (one per channel), that

targets to identify bad node performance in a specific channel, not depending if it is better or the same on other ones.

Since there would be six times more figures and information than in the previous section, for a better presentation and comprehension of the method there are a couple channel clustering commented.

In Figure 38 the division has been made more dependent on the PC1 scores, what becomes a median availability division in Figure 39, where the variables can be located.

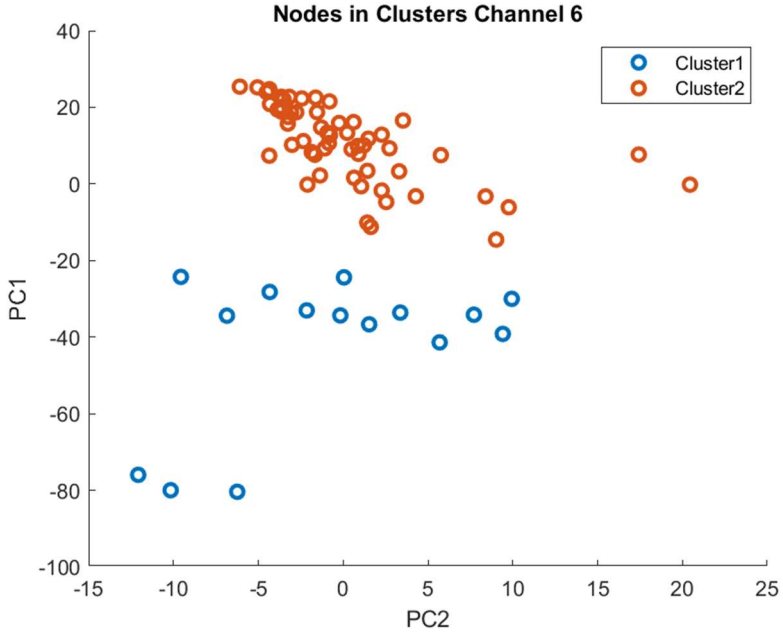


Figure 38: Channel by channel, channel 6 principal components of the nodes in each cluster

In channel six, a division in the median availability around 70% was found. This can be understood as the border between of good and bad performance. It could be divided in three clear clusters, but the division between a good performance and a bad one is useful for the company for selecting the best channel in each situation.

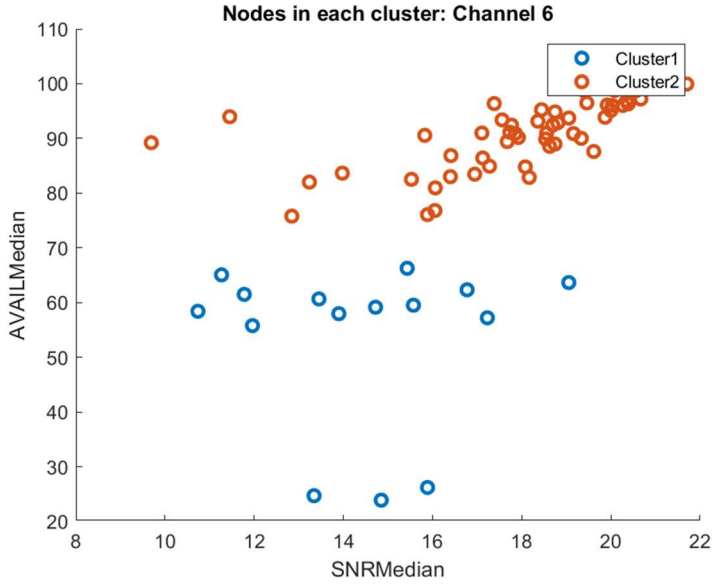


Figure 39: Channel by channel, channel 6 Avail and SNR medians of the nodes in each cluster

Channel seven has a similar division, observed in both Figure 40 and Figure 41, that will help for in that binary groups, although in this case the number of nodes in the lower availability group is reduced what will complicate the future classification due to the low number of samples.

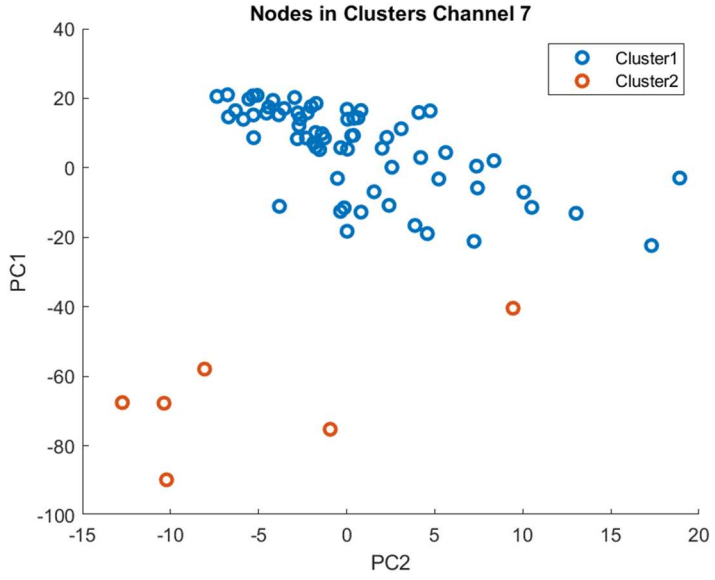


Figure 40: Channel by channel, channel 7 principal components of the nodes in each cluster

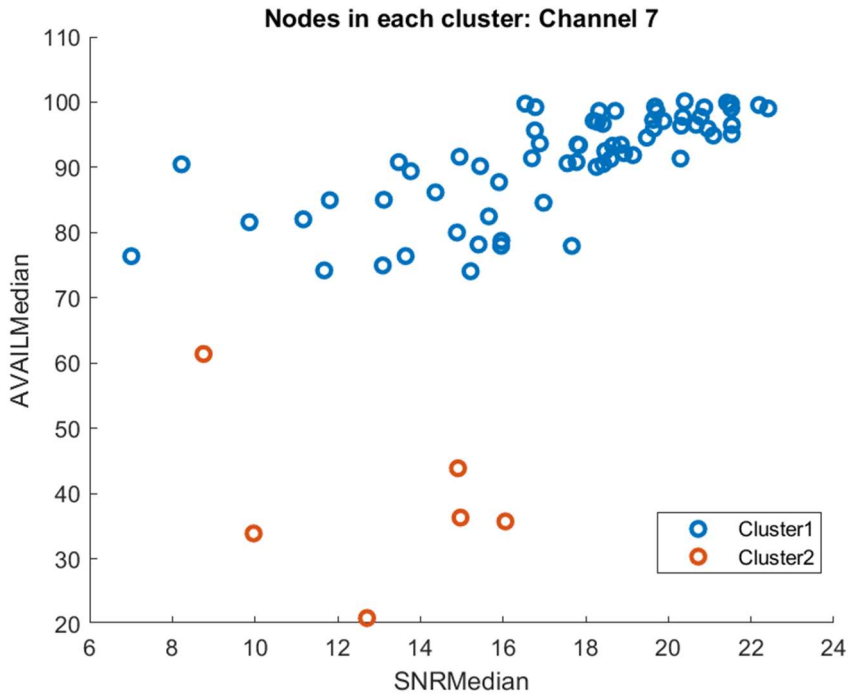


Figure 41: Channel by channel, channel 7 Avail and SNR medians of the nodes in each cluster

Upper channels have a reduced correlation between SNR and Availability that helps on dividing in the two clusters, while lower channels availability and SNR creates almost a line with less variance, which worsen the division and subsequent classification.

5.3.3 OTHER APPROACHES

The number of different variables in contact for the clustering process is vast. The main aspects that have been considered are:

- The usage or not of the PCA
- Which and how many variables introduced in the PCA
- Normalizing or not the variables
- How many components to take
- The clustering method
- How many clusters to select
- Clustering all the channels at the same time or one by one

After performing several iterations with most of the parameters, the two options explained in the previous sections have been selected for being more useful in the company practice, easier to extract insights from and performing better than the others with the existent data.

5.4 CLASSIFICATION OF FUSES BOXES

In this chapter the first two sections of the clustering have been maintained to comment each of them separately. For each clustering process the classification method is the same, the samples are divided in training and testing sets, that are used for verifying the classification tree. The tree method has been selected as the default one as it has been proved to be the best one for this kind of classifiers.

5.4.1 CHANNEL COMBINATION CLUSTERING

In Figure 42 is represented the importance that each predictor achieved in the training process of the tree. At the start of the study, all the available variables were used for classification, although with multitude of iterations the constantly less important and the ones that overfitted the model were set aside.

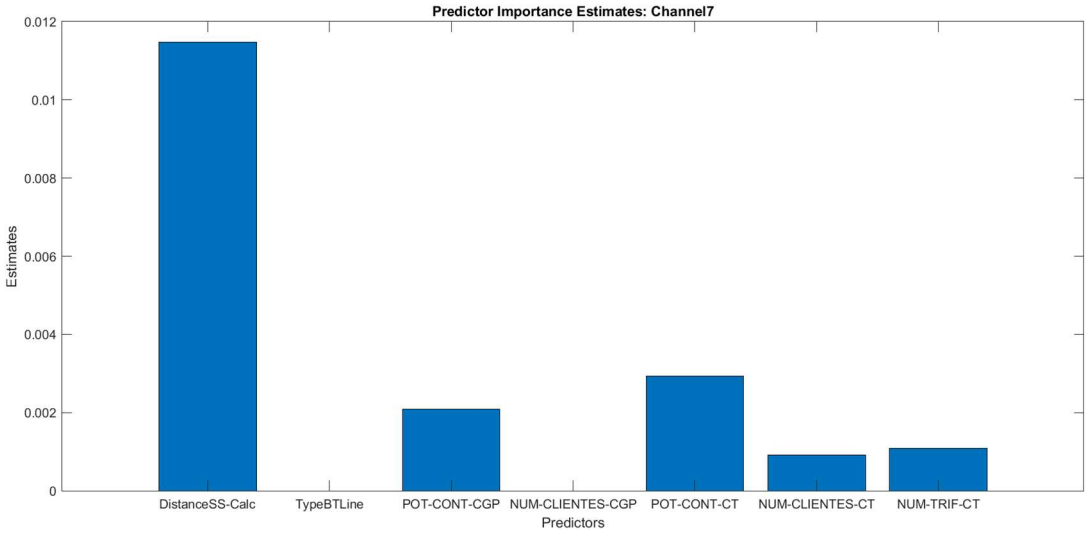


Figure 42: Channels Combination, predictors importance

The final set of variables are seven: Distance to the SS, the type of LV line ,the sum of contracted power and number of clients in the HCB and the contracted power, number of clients and number of three-phase clients in the SS. Although in Figure 42 shows no importance variables, that changes in other channels and so they have to be included.

The classification tree graph is showed in Figure 43. The different splits observed in the tree are more probable to run into a cluster two classification, since there are significantly more elements in this cluster respect to the others. This is a bias that tend to correct with the increment of the sample set, that for now is reduced.

The reduce number of nodes in some clusters tend to overfit the tree to the samples contained in the training set, not being proper for a general classification. Figure 44 shows the confusion matrices that show the problem previously mentioned, concentration of nodes in cluster two and no samples of cluster one or three in the testing set.

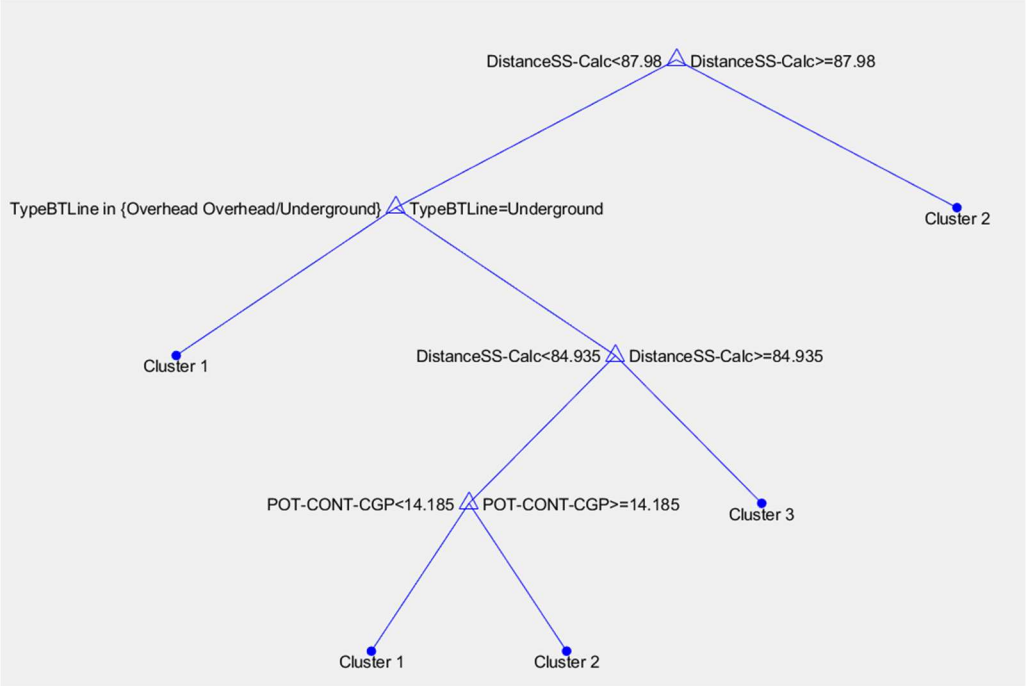


Figure 43: Channels Combination, classification tree

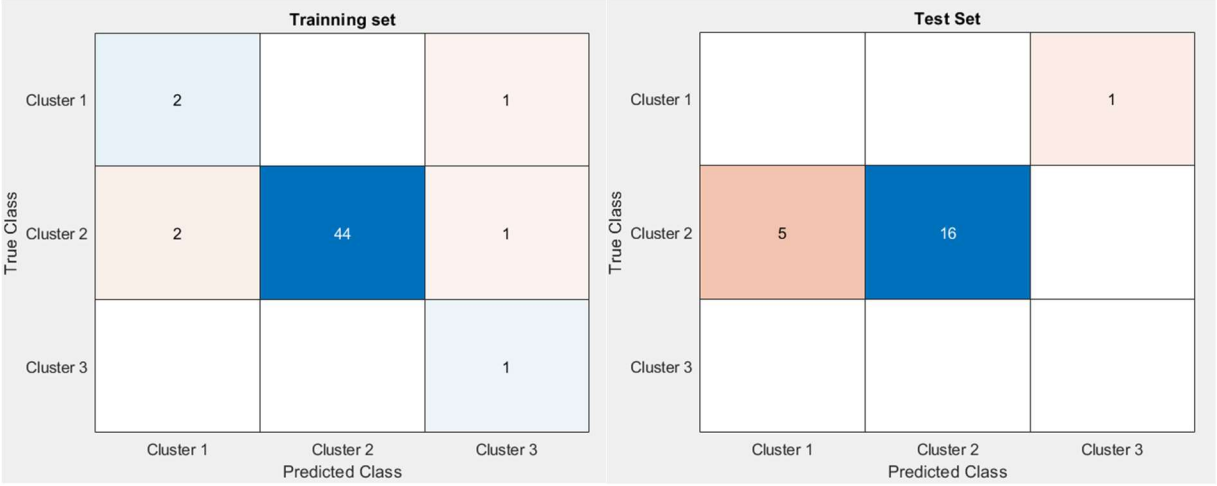


Figure 44: Channels Combination, Confusion Matrices

From the tree, it can be identified that high contracted power with long lines tend to be classified in cluster two, the one with better performance in all the channel, although the sample bias can affect since it does not classify well other clusters.

5.4.2 CHANNEL BY CHANNEL CLUSTERING

In the classification by channel, although the method is the same as the previous one, and the difficulties of number of representative variables and the reduced amount of samples continues being the main issues. Since in this section the behaviour of the channel 6 and 7 is very similar, this last channel will be the only one commented.

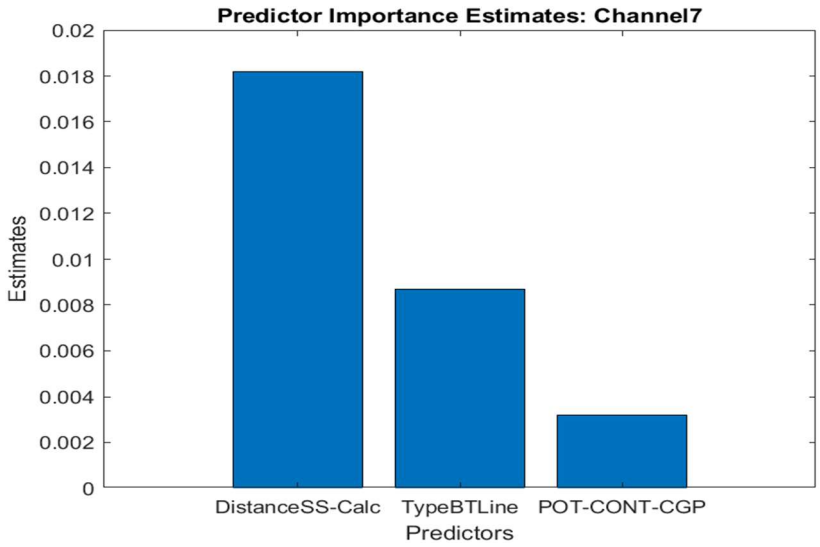


Figure 45: Channel by channel, channel 7 predictors importance

For the classification of channel 7 the Distance to the SS is more relevant, losing importance the contracted power specially. The existence of two only groups makes simpler the categorization, although it enhances the group number disproportion. However, the tree is correctly classifying most of the samples.

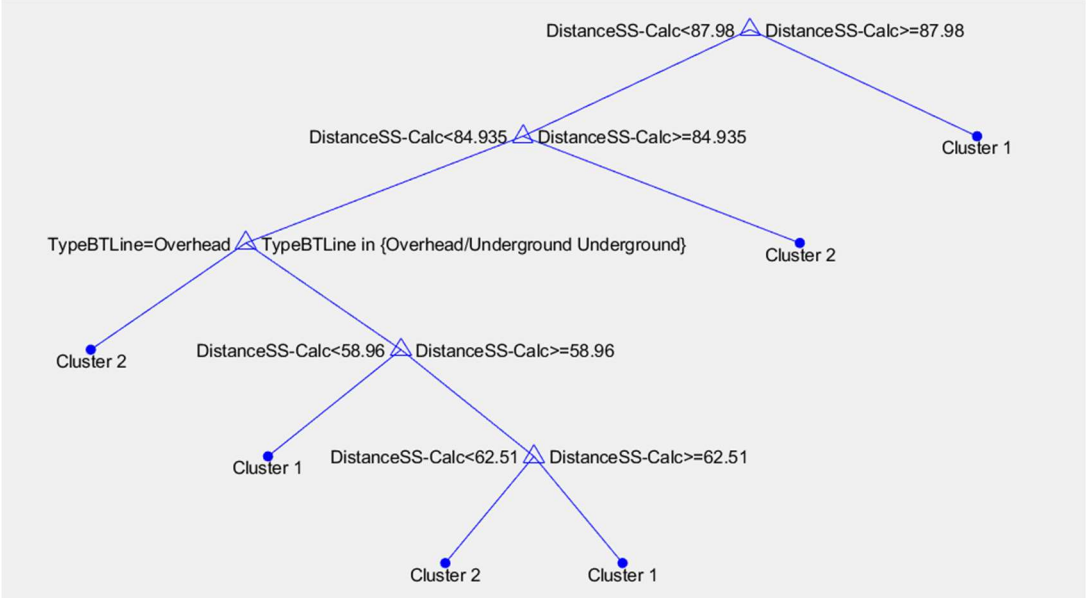


Figure 46: Channel by channel, channel 7 classification tree

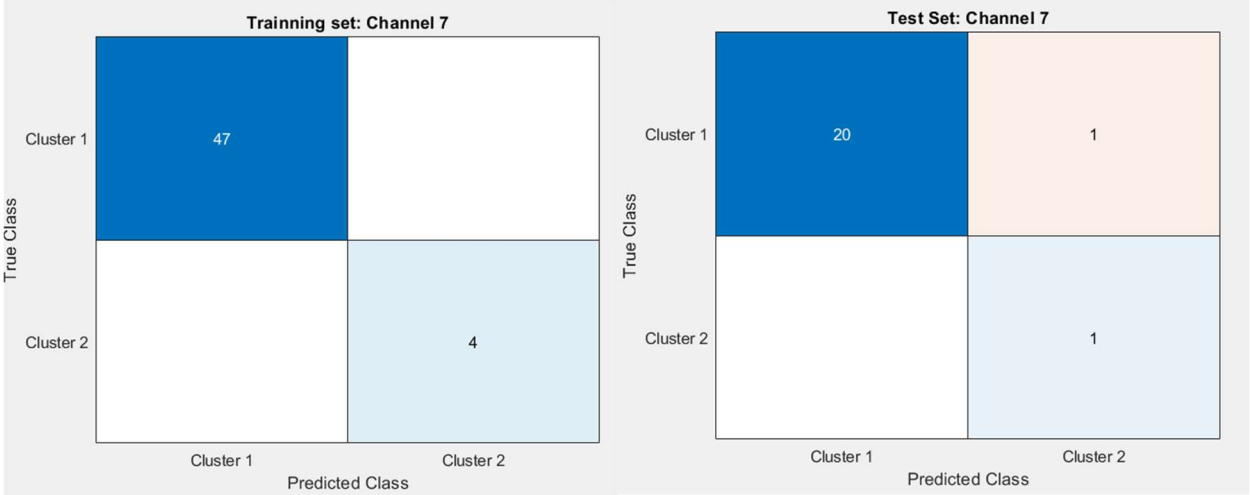


Figure 47: Channel by channel, channel 7 confusion matrices

The classification of these channels can be safer since it classifies correctly both clusters. The information given from the tree is that the worse performance for upper channels appears in long distances from the SS and in overhead lines. For lower channels, the classification does not work so well, and the line properties are not clear.

Chapter 6. CONCLUSIONS AND FUTURE WORK

Nowadays there is a reduced amount of communication performance measures. In addition to this, it is complex to extract any conclusion from them without a deep knowledge of the PLC communication in LV grids and all the particularities of the grid where they are obtained.

A promising process of solving partially this situation is the PCA method combined with clustering. Through it, those measures can be combined and represented for then be divided depending on variables that consider all the information. This division, after a detailed adjust and iteration, can contribute to the analysis process of the practitioners finding tendencies and confirm or reject hypothesis. The attention to this method has to be maximum since from it is possible to do holistic analysis if it is adjusted properly, and all the future processes will be biased from the scores obtained from it.

From a stand-alone data analysis carried out, some of the conclusions have been extracted that could be generalized. First, the increased noise power in the lower channels and the higher signal attenuation in the upper ones seen in papers could be reflected in the SNR values of the measures. SNR is lower in third and eight channels while the middle ones have typically better values. These values are more differenced if the median is taken instead of the mean, reaching 3-5% bigger numbers. SNR is affected by the noise power and the signal strength, making low frequencies suffer from the presence of noise and high frequencies from the attenuation. Second, availability is not as much affected by the attenuation, this conclusion is due to the increasing availability with frequency, not reflecting the decrease seen in the SNR. An important characteristic to mention about the dataset is the decreasing availability-SNR correlation with frequency, since some of the worst availability values in the upper channels happens for the best SNR values, while in the lower ones both are completely correlated.

The clustering process has been predominantly successful and from it could be concluded two different aspects. The first one, from the clustering done combining channels, is the presence of at least three groups of nodes with common characteristics based on the performance along the six channels studied. It has been selected three clusters because they have enough nodes to be generalized with the existent data and gives useful information about the nodes included in each group. These three groups would be one with a good performance in every channel, a second one with better performance in the lower channels and a third one average on lower channels and good in the upper ones. The most populated one will be the first cluster.

From characterization done channel by channel, upper channels have a clearer distinction between good and bad performance. However, the total correlation between availability and SNR in lower frequency channels previously mentioned reduces the nodes variance and complicates the grouping. The measures obtained from the grid seems enough for a proper clustering of the upper channels, while others need to be found for the lower ones.

From the PCA analysis and the following clustering, a method for the identification of specific characteristics of the worst performers was consider useful. This would be a classification method, which performance is linked to the clustering. The classification data is fed from the clustering groups and topology data of those devices. With that knowledge, it tries to figure out the characteristics that divides each group. Lower channels are more inaccurately classified, also due to the imprecision of the previous clustering. Upper ones' classification is more accurate and seem to perform worse in longer lines, what is in accordance with the state-of-the-art conclusions.

Summing up the needs for improving the results observed in this work, the number of nodes analysed is crucial for both clustering and classification since both uses data driven methods, especially for precision in the classification. Other improvements for each section would be: a bigger research on the set of variables introduced in the PCA, particularizing more the variables for the cases with a worse clustering result; adjust the number of clusters to the new data; the identification of variables that reflects indirectly better the noise or the

common location can improve heavily the classification in the lower channels. Another future improvement in the process is the close work with technicians, to also work in the verification or rejection of stablished hypothesis instead of working from a blank page.

Chapter 7. BIBLIOGRAPHY

[1] PRIME Alliance TWG, "PRIME v1.4 White Paper." Accessed: May 06, 2020. [Online]. Available: https://www.prime-alliance.org/wp-content/uploads/2014/10/whitePaperPrimeV1p4_final.pdf.

[2] I. Fernandez, A. Arrinda, I. Angulo, D. De La Vega, N. Uribe-Perez, and A. Llano, "Field Trials for the Empirical Characterization of the Low Voltage Grid Access Impedance From 35 kHz to 500 kHz," *IEEE Access*, vol. 7, pp. 85786–85795, 2019, doi: 10.1109/ACCESS.2019.2924253.

[3] I. Fernández *et al.*, "Characterization of the frequency-dependent transmission losses of the grid up to 500 kHz," Madrid, Spain, Jun. 2019, vol. 1146.

[4] M. R. Fliss, J. H. Fernandez, A. Omri, and G. Oligeri, "NB-PLC Successful Transmission Probability Analysis," in *2019 2nd International Conference on Smart Grid and Renewable Energy (SGRE)*, Doha, Qatar, Nov. 2019, pp. 1–6, doi: 10.1109/SGRE46976.2019.9020694.

[5] Guangbin Chu, Jianqi Li, and Weilin Liu, "Narrow band power line channel characteristics for low voltage access network in China," in *2013 IEEE 17th International Symposium on Power Line Communications and Its Applications*, Johannesburg, Mar. 2013, pp. 297–302, doi: 10.1109/ISPLC.2013.6525867.

[6] I. Fernandez *et al.*, "Characterization of non-intentional emissions from distributed energy resources up to 500 kHz: A case study in Spain," *Int. J. Electr. Power Energy Syst.*, vol. 105, pp. 549–563, Feb. 2019, doi: 10.1016/j.ijepes.2018.08.048.

[7] I. Arechalde, M. Castro, I. Garcia-Borreguero, A. Sendin, I. Urrutia, and A. Fernandez, "Performance of PLC communications in frequency bands from 150 kHz to 500 kHz," in *2017 IEEE International Symposium on Power Line Communications and its Applications (ISPLC)*, Madrid, Spain, 2017, pp. 1–5, doi: 10.1109/ISPLC.2017.7897123.

[8] M. Wolkerstorfer, B. Schweighofer, H. Wegleiter, D. Statovci, H. Schwaiger, and W. Lackner, "Measurement and simulation framework for throughput evaluation of narrowband power line communication links in low-voltage grids," *J. Netw. Comput. Appl.*, vol. 59, pp. 285–300, Jan. 2016, doi: 10.1016/j.jnca.2015.05.022.

[9] E. Standards, "UNE EN 50065-1:2012 Signalling on low-voltage electrical installations in the frequency range 3 kHz to 148,5 kHz - Part 1: General requirements, frequency bands and electromagnetic disturbances," *https://www.en-standard.eu*. https://www.en-standard.eu/une-en-50065-1-2012-signalling-on-low-voltage-electrical-installations-in-the-frequency-range-3-khz-to-148-5-khz-part-1-general-requirements-frequency-bands-and-electromagnetic-disturbances/ (accessed Aug. 29, 2020).

[10] i-DE Grupo Iberdrola, "Cuadros de distribución en BT con embarrado aislado para Centros de Transformación Compactos." Apr. 2019.

[11] i-DE Grupo Iberdrola, "ESPECIFICACIONES PARTICULARES PARA SISTEMAS DE TELEGESTIÓN Y AUTOMATIZACIÓN DE RED." Jul. 2019.
[12] i-DE Grupo Iberdrola, "ESPECIFICACIONES PARTICULARES PARA INSTALACIONES DE ALTA TENSIÓN (HASTA 30 kV) Y BAJA TENSIÓN." May 2019.
[13] i-DE Grupo Iberdrola, "Líneas aéreas de Baja Tensión con cables aislados." Jan. 1998.
[14] i-DE Grupo Iberdrola, "Cajas generales de protección (CGP)." Jul. 2010.
[15] i-DE Grupo Iberdrola, "ESPECIFICACIONES PARTICULARES PARA INSTALACIONES DE ENLACE." Sep. 2013.

Chapter 8. ANNEX I: SUSTAINABLE DEVELOPMENT GOALS (SDGS)

The project, as part of two institutions concern about the SDGs, Iberdrola and Universidad Pontificia Comillas are involved in most of the aspects to achieve them, although some of them are more relevant in this specific topic.

In particular, this project has a direct relation with goal 9, the one named industry, innovation and infrastructure; and goal 17, about partnerships for the goals. As part of an investigation project for improving communication of DSOs with meters, the thesis is part of the innovation and the industry. On the other hand, it exists due to a partnership between Iberdrola and ICAI and works for developing this relationship.

There are also other two goals that will be indirectly affected by the consequences of the achievements of this thesis:

- Goal 7: affordable and clean energy
- Goal 13: climate action

Achieving a better communication in the low voltage grid will allow a more digitalized grid, which is more efficient one dealing with renewable and distributed generation. That improvement will help achieving both of the previous mentioned goals at the same time.

Chapter 9. ANNEX II: SCRIPTS

9.1 FORMATTING PYTHON SCRIPT

9.1.1 MAIN

```
import copy
import time

import openpyxl
import os
import win32com.client as win32
from classes import CT


saveDir = 'Archivos Creados'
saveFileCTRef = '_v0-200517'
saveFileSum = 'PRIME_canalesPor10Min_DEF200422_v2'

redColor='FF0000'
yellowColor='FFFF00'
greenColor='00FF00'
blackColor='000000'

# obtain the above directory
programDir = os.getcwd()
baseDir = os.path.dirname(os.getcwd())
print('Base Folder: '+baseDir)

if os.path.isdir(os.path.join(baseDir, saveDir)) == 0:
    os.mkdir(os.path.join(baseDir, saveDir))

allCT = {}
titlesCT = ['Time', 'AVAIL 3', 'AVAIL 4', 'AVAIL 5', 'AVAIL 6', 'AVAIL 7', 'AVAIL
8', 'RSSI 3', 'RSSI 4', 'RSSI 5',
            'RSSI 6', 'RSSI 7', 'RSSI 8', 'SNR 3', 'SNR 4', 'SNR 5', 'SNR 6',
'SNR 7', 'SNR 8']
titlesSumCT = ['Name', 'AVAIL 3', 'AVAIL 4', 'AVAIL 5', 'AVAIL 6', 'AVAIL 7',
'AVAIL 8', 'RSSI 3', 'RSSI 4', 'RSSI 5',
               'RSSI 6', 'RSSI 7', 'RSSI 8', 'SNR 3', 'SNR 4', 'SNR 5', 'SNR 6',
'SNR 7', 'SNR 8', 'MAC', 'Address_CGP',
               'Distance SS', 'Type BT', 'COD_CT', 'COD_CGP']

Excel = win32.gencache.EnsureDispatch('Excel.Application')
Excel.Visible = 0
win32c = win32.constants
```

```
# Obtain all the subdirectories of the CTs
os.chdir(baseDir)
with os.scandir(baseDir) as entries:
    for entry in entries:
        # Get data form original files and create class
        if entry.is_dir() and entry.name != os.path.basename(programDir) and
entry.name != saveDir:
            ct = CT(entry.name, entry.path, entry.stat())
            allCT[entry.name] = ct

            os.chdir(os.path.join(baseDir, saveDir))

            ####################################
            # Creation of data files of each CT #
            ####################################

            filePath = os.path.join(baseDir, saveDir, ct.name.replace(' ',
'_').lower() + saveFileCTRef+'.xlsx')
            wb = openpyxl.Workbook()
            print('Creating {} Excel file...'.format(ct.name.replace(' ',
'_').lower() + saveFileCTRef))

            if 'Summary' not in wb.sheetnames:
                wb.create_sheet('Summary', 0)
            ws = wb['Summary']
            ctLastRowSum = 0
            ws.append(titlesSumCT)
            ctLastRowSum += 1
            for cell in ws[ctLastRowSum]:
                cell.font = openpyxl.styles.Font(bold=True)
                cell.alignment = openpyxl.styles.Alignment(horizontal='center')
                cell.border =
openpyxl.styles.Border(bottom=openpyxl.styles.Side(border_style='double',
color="000000"))
            for cell in ws['A:A']:
                cell.font = openpyxl.styles.Font(bold=True)
                cell.alignment = openpyxl.styles.Alignment(horizontal='center')

            for device in ct.devices:
                if device not in wb.sheetnames:
                    wb.create_sheet(device)

                subdata = copy.deepcopy(ct.devices[device]['AVL'][:])
                for i, row in enumerate(subdata):
                    subdata[i]+= ct.devices[device]['RSSI'][i][1:]
                    subdata[i]+= ct.devices[device]['SNR'][i][1:]

                data = [titlesCT, subdata]
                for row in data:
                    if row[0].__class__ == str:
                        wb[device].append(row)
                    else:
                        for r in row:
```

```
                            wb[device].append(r)

                formulas = []

                for i in range(2, len(subdata[0][1:]) + 2, 1):
                    ref = device + '!' + chr(i + 96).upper() + ':' + chr(i +
96).upper()
                    formulas += ['=IFERROR(AVERAGE({}),0)'.format(ref)]

                data = [ct.devices[device]['name']] + formulas +
[ct.devices[device]['MAC'],

ct.devices[device]['ADDRESS_CGP'],

ct.devices[device]['DistanceSS'],

ct.devices[device]['TypeBTLine'],

ct.devices[device]['COD_CT'],

ct.devices[device]['COD_CGP']]
                ws.append(data)
                ctLastRowSum += 1

            for row in ws['B{}:S{}'.format(ctLastRowSum - len(ct.devices),
ctLastRowSum)]:
                for cell in row:
                    cell.number_format = '0.00'
            ws.conditional_formatting.add('B{}:G{}'.format(ctLastRowSum -
len(ct.devices), ctLastRowSum),

openpyxl.formatting.rule.ColorScaleRule(start_type='num',
start_value=0,start_color=redColor,

mid_type='percentile', mid_value=50,mid_color=yellowColor,

end_type='num', end_value=100,end_color=greenColor))
            ws.conditional_formatting.add('H{}:M{}'.format(ctLastRowSum -
len(ct.devices), ctLastRowSum),

openpyxl.formatting.rule.ColorScaleRule(start_type='min',start_color=redColor,

mid_type='percentile', mid_value=50,mid_color=yellowColor,

end_type='max',end_color=greenColor))
            ws.conditional_formatting.add('N{}:S{}'.format(ctLastRowSum -
len(ct.devices), ctLastRowSum),

openpyxl.formatting.rule.ColorScaleRule(start_type='min',start_color=redColor,

mid_type='percentile', mid_value=50, mid_color=yellowColor,

end_type='max',end_color=greenColor))
```

```
            for column in
range(openpyxl.utils.cell.column_index_from_string('T'),openpyxl.utils.cell.colum
n_index_from_string('Y')+1,1):

ws.column_dimensions[openpyxl.utils.get_column_letter(column)].width = 20
                ws.column_dimensions[openpyxl.utils.get_column_letter(column)]
            for row in ws['T:Y']:
                for cell in row:
                    if cell.column_letter == 'U':
                        ws.column_dimensions[cell.column_letter].width = 30
                    else:
                        ws.column_dimensions[cell.column_letter].width = 20
                    cell.number_format = 'General'
                    cell.alignment =
openpyxl.styles.Alignment(horizontal="center")

            # Creating the bar chart of availability for devices and channels
            chart = openpyxl.chart.BarChart()
            for i,device in enumerate(ct.devices, start=2):
                values = openpyxl.chart.Reference(ws,min_col=2, min_row=i,
max_col=7)
                series = openpyxl.chart.Series(values=values, title=device)
                chart.series.append(series)
            categories = openpyxl.chart.Reference(ws, min_col=2, min_row=1,
max_col=7)
            chart.set_categories(categories)
            chart.title = ' Availability '
            ws.add_chart(chart,'B10')
            if 'Sheet' in wb.sheetnames:
                wb.remove(wb['Sheet'])
            wb.save(ct.name.replace(' ', '_').lower() + saveFileCTRef + '.xlsx')

############################
# Create summary file
############################

wb = openpyxl.Workbook()
print('Creating {} Excel file...'.format(saveFileSum))

if 'Datos' not in wb.sheetnames:
    wb.create_sheet('Datos', 0)
ws = wb['Datos']
ws.append(['Time', 'Disponibilidad', 'Canal', 'CT'])

for ct in allCT:
    for canal in range(3,len(allCT[ct].channelsAVL[0])+2,1):
        for ntiempo in range(len(allCT[ct].channelsAVL)):
            ws.append([allCT[ct].channelsAVL[ntiempo][0],
allCT[ct].channelsAVL[ntiempo][canal-2], canal, allCT[ct].name.replace(' ',
'_').lower()])

ws.auto_filter.ref=ws.dimensions
if 'Sheet' in wb.sheetnames:
```

```
    wb.remove(wb['Sheet'])
wb.save (saveFileSum + '.xlsx')

Excel = win32.gencache.EnsureDispatch('Excel.Application')
Excel.Visible = 0
win32c = win32.constants
wb = Excel.Workbooks.Open(os.path.join(baseDir, 'Archivos Creados',saveFileSum+
'.xlsx'))
pvt_sheet = wb.Sheets.Add(Before=wb.Worksheets('Datos'))
pvt_rng_beg = pvt_sheet.Cells(1,1)
pvt_rng_end = pvt_sheet.Cells(1,1)
pvt_dest_rng = pvt_sheet.Range(pvt_rng_beg, pvt_rng_end)
pvt_sheet.Name = 'TablaDinamica'
pivotTableName = 'TablaDisponibilidad'

src_sheet = wb.Worksheets('Datos')
pvt_src = src_sheet.UsedRange

PivotCache = wb.PivotCaches().Create(SourceType=win32c.xlDatabase,
SourceData=pvt_src, Version=win32c.xlPivotTableVersion14)
PivotTable =PivotCache.CreatePivotTable(TableDestination=pvt_dest_rng,
TableName=pivotTableName, DefaultVersion=win32c.xlPivotTableVersion14)

PivotTable.PivotFields('Time').Orientation = win32c.xlRowField
PivotTable.PivotFields('Time').Position = 1
PivotTable.PivotFields('CT').Orientation = win32c.xlColumnField
PivotTable.PivotFields('CT').Position = 1
PivotTable.PivotFields('Canal').Orientation = win32c.xlColumnField
PivotTable.PivotFields('Canal').Position = 2
DataField =
PivotTable.AddDataField(PivotTable.PivotFields('Disponibilidad'),'Promedio
Disponibilidad',win32c.xlAverage)

chrt_sheet = wb.Charts.Add(Before=wb.Worksheets('TablaDinamica'))
chrt_sheet.Name = 'Grafica Detalle'
chrt_sheet.ChartType =  win32c.xlLine
chrt_sheet.HasTitle = True
chrt_sheet.ChartTitle.Text = "Disponibilidad"

sum_sheet = wb.Sheets.Add(Before=wb.Charts('Grafica Detalle'))
sum_sheet.Name = 'Resumen'

Excel.Application.DisplayAlerts = False
wb.SaveAs(os.path.join(baseDir,saveDir,saveFileSum+ '.xlsx'))
Excel.Application.DisplayAlerts = True
Excel.Application.Quit()

# Average Avalitability table by channel and ct
wb = openpyxl.load_workbook(saveFileSum + '.xlsx')
if 'Resumen' not in wb.sheetnames:
    wb.create_sheet('Resumen')
ws = wb['Resumen']
LastRowResumen = 0
```

```
ws.append(['Canales','3','4','5','6','7','8','Prom.','Disper.'])
LastRowResumen+=1
for cell in ws[1]:
    cell.font = openpyxl.styles.Font(bold=True)
    cell.alignment = openpyxl.styles.Alignment(horizontal='center')
    cell.border =
openpyxl.styles.Border(bottom=openpyxl.styles.Side(border_style='double',color="0
00000"))

for row,ct in enumerate(allCT, start=2):
    ct_data = []
    for canal in range(2,len(allCT[ct].channelsAVL[0])+4,1):
        if canal == 2:
            ct_data += [ct]
        elif canal == 9:
            ct_data += ['=AVERAGE(B{0}:G{0})'.format(row)]
        elif canal == 10:
            ct_data += ['=MAX(B{0}:G{0})-MIN(B{0}:G{0})'.format(row)]
        else:
            ct_data += ['=GETPIVOTDATA("Promedio
Disponibilidad",{}TablaDinamica!$A$1,"Canal",{},"CT","{}")'.format('['+saveFileSu
m+ '.xlsx]',canal,ct.replace(' ', '_').lower())]

    ws.append(ct_data)
    LastRowResumen += 1

for row in ws['B{}:I{}'.format(LastRowResumen-len(allCT),LastRowResumen)]:
    for cell in row:
        cell.number_format = '0.00'
ws.conditional_formatting.add('B{}:G{}'.format(LastRowResumen-
len(allCT),LastRowResumen),openpyxl.formatting.rule.ColorScaleRule(start_type='nu
m',start_value=0,start_color=redColor,

mid_type='percentile',mid_value=50,mid_color=yellowColor,

end_type='num',end_value=100,end_color=greenColor))
ws.conditional_formatting.add('H{}:H{}'.format(LastRowResumen-
len(allCT),LastRowResumen),openpyxl.formatting.rule.ColorScaleRule(start_type='mi
n',start_color=redColor,

mid_type='percentile',mid_value=50,mid_color=yellowColor,

end_type='max',end_color=greenColor))
ws.conditional_formatting.add('I{}:I{}'.format(LastRowResumen-
len(allCT),LastRowResumen),openpyxl.formatting.rule.ColorScaleRule(start_type='mi
n',start_color=redColor,

mid_type='percentile',mid_value=50,mid_color=yellowColor,

end_type='max',end_color=greenColor))

# Average Availability by channel
```

```
ct_data = ['All']
for canal in range(2,len(allCT[ct].channelsAVL[0])+1,1):
    letra = chr(canal+96).upper()
    ct_data += ['=AVERAGE({0}2:{0}{1})'.format(letra,LastRowResumen)]
ws.append(ct_data)
LastRowResumen+=1

for row in ws['B{}:I{}'.format(LastRowResumen-len(allCT),LastRowResumen)]:
    for cell in row:
        cell.number_format = '0.00'
ws.conditional_formatting.add('B{0}:G{0}'.format(LastRowResumen),openpyxl.formatt
ing.rule.ColorScaleRule(start_type='min',start_color=redColor,

mid_type='percentile',mid_value=50,mid_color=yellowColor,

end_type='max',end_color=greenColor))

# Average Availabillity by channel chart
chart = openpyxl.chart.BarChart()
values = openpyxl.chart.Reference(ws,min_row=LastRowResumen, min_col=2,
max_col=7)
series = openpyxl.chart.Series(values=values, title='Todos')
chart.series.append(series)
categories = range(3,8,1)
chart.title = ' Disponibilidad Media por canal '
chart.height = 6
chart.width = 12
ws.add_chart(chart,'B{}'.format(LastRowResumen+2))

# Availability chart by channel and ct
chart = openpyxl.chart.BarChart3D()
for canal in range(3,len(allCT[ct].channelsAVL[0])+2,1):
    values = openpyxl.chart.Reference(ws,min_row=2, min_col=canal-1,
max_row=LastRowResumen)
    series = openpyxl.chart.Series(values=values, title='Ch.{}'.format(canal))
    chart.series.append(series)
categories = openpyxl.chart.Reference(ws, min_row=2, min_col=1,
max_row=LastRowResumen)
chart.set_categories(categories)
chart.Name = 'Availability'
chart.title = ' Disponibilidad Canal y CT '
chart.height = 15
chart.width = 23
ws.add_chart(chart,'K2')

# Blank rows
blankRows1 = 13
for i in range(blankRows1):
    ws.append([''])
    LastRowResumen += 1

# Ranking channel Table
ws.append(['Posición (el 1 el mejor de los canales):'])
```

```
LastRowResumen += 1
for cell in ws[LastRowResumen]:
    cell.font = openpyxl.styles.Font(bold=True)
    cell.alignment = openpyxl.styles.Alignment(horizontal='center')
    cell.border =
openpyxl.styles.Border(bottom=openpyxl.styles.Side(border_style='double',color="0
00000"))

for row,ct in enumerate(allCT, start=2):
    ct_data = []
    for canal in range(2,len(allCT[ct].channelsAVL[0])+2,1):
        if canal == 2:
            ct_data += [ct]
        else:
            ct_data += ['=RANK({0}{1},$B{1}:$G{1})'.format(chr(canal+96-
1).upper(),row)]

    ws.append(ct_data)
    LastRowResumen += 1
ct_data =['All']
for canal in range(2,len(allCT[ct].channelsAVL[0])+1,1):
    letra = chr(canal+96).upper()
    ct_data += ['=RANK({0}{1},$B{1}:$G{1})'.format(letra,row+1)]
ws.append(ct_data)
LastRowResumen += 1
for row in ws['B{}:G{}'.format(LastRowResumen-(len(allCT)+1),LastRowResumen)]:
    for cell in row:
        cell.alignment = openpyxl.styles.Alignment(horizontal='center')
ws.conditional_formatting.add('B{}:G{}'.format(LastRowResumen-
(len(allCT)+1),LastRowResumen),openpyxl.formatting.rule.ColorScaleRule(start_type
='min',start_color=greenColor,

mid_type='percentile',mid_value=50,mid_color=yellowColor,

end_type='max',end_color=redColor))

# Blank rows 2
blankRows2=2
for i in range(blankRows2):
    ws.append([''])
    LastRowResumen += 1

# Ranking Count Table
ws.append(['Resumen: Orden vs Numero de veces'])
LastRowResumen += 1
for cell in ws[LastRowResumen]:
    cell.font = openpyxl.styles.Font(bold=True)
    cell.alignment = openpyxl.styles.Alignment(horizontal='center')
    cell.border =
openpyxl.styles.Border(bottom=openpyxl.styles.Side(border_style='double',color="0
00000"))

for row in range(1,7,1):
```

```
    ct_data = []
    for canal in range(2,len(allCT[ct].channelsAVL[0])+2,1):
        if canal == 2:
            ct_data += [row]
        else:
            ct_data +=
['=IF(COUNTIF({0}${2}:{0}${3},{1})=0,"",COUNTIF({0}${2}:{0}${3},{1}))'.format(chr
(canal+96-1).upper(),row,LastRowResumen-row-blankRows2-len(allCT),LastRowResumen-
row-blankRows2-1)]
    ws.append(ct_data)
    LastRowResumen += 1

for row in ws['B{}:G{}'.format(LastRowResumen-7,LastRowResumen)]:
    for cell in row:
        cell.alignment = openpyxl.styles.Alignment(horizontal='center')
ws.conditional_formatting.add('B{}:G{}'.format(LastRowResumen-
7,LastRowResumen),openpyxl.formatting.rule.ColorScaleRule(start_type='min',start_
color=redColor,

mid_type='percentile',mid_value=50,mid_color=yellowColor,

end_type='max',end_color=greenColor))
for cell in ws['A']:
    cell.font = openpyxl.styles.Font(bold=True)
    cell.alignment = openpyxl.styles.Alignment(horizontal='center')

wb.save (saveFileSum + '.xlsx')

#time.sleep(0.05)
try:
    wb = Excel.Workbooks.Open(os.path.join(baseDir, 'Archivos
Creados',saveFileSum+ '.xlsx'))
except:
    time.sleep(0.05)
    wb = Excel.Workbooks.Open(os.path.join(baseDir, 'Archivos Creados',
saveFileSum + '.xlsx'))
ws = wb.Sheets('Resumen')
#chart = ws.ChartObjects(2)
#chart.Chart.ChartType = win32c.xl3DColumn
wb.RefreshAll()
ws.Columns(1).AutoFit()
Excel.Application.DisplayAlerts = False
wb.SaveAs(os.path.join(baseDir, 'Archivos Creados',saveFileSum+ '.xlsx'))

# Average Availability table i ct files from average channel availability file
for ct in allCT:
    wb = openpyxl.load_workbook(ct.replace(' ', '_').lower() + saveFileCTRef +
'.xlsx')
    data =['All']
    for i in range(3,len(allCT[ct].channelsAVL[0])+2,1):
        data += ['=GETPIVOTDATA("Promedio
Disponibilidad",{}TablaDinamica!$A$1,"Canal",{},"CT","{}")'.format('['+saveFileSu
m+ '.xlsx]',i,ct.replace(' ', '_').lower())]
```

```
    wb['Summary'].append(data)
    ctLastRowSum += 1
    for cell in wb['Summary'][ctLastRowSum]:
        cell.number_format = '0.00'

wb['Summary'].conditional_formatting.add('B{0}:G{0}'.format(ctLastRowSum),openpyx
l.formatting.rule.ColorScaleRule(start_type='min',start_color=redColor,

mid_type='percentile', mid_value=50,mid_color=yellowColor,

end_type='max',end_color=greenColor))
    wb.save(ct.replace(' ', '_').lower() + saveFileCTRef + '.xlsx')
    ctLastRowSum -= 1

    wb = Excel.Workbooks.Open(
        os.path.join(baseDir, 'Archivos Creados', ct.replace(' ', '_').lower() +
saveFileCTRef + '.xlsx'))
    wb.RefreshAll()
    wb.Close(SaveChanges=True)
    Excel.Application.DisplayAlerts = True
Excel.Application.Quit()
```

9.1.2 CLASSES

```
import os, fnmatch, csv

import openpyxl

topoFile='CTs CAMPO.xlsx'

class CT:
    def __init__(self, name, path, info):
        self.name = name
        self.path = path
        self.info = info
        self.devices = self.GetDevices()
        self.channelsAVL= self.GetChannelsAVL()
    def __str__(self):
        return str(self.__dict__)

    def GetDevices(self):

        genPath = os.path.dirname(os.path.dirname(self.path))
        for (root, dirs, files) in os.walk(genPath):
            if topoFile in files:
                topoPath = os.path.join(root, topoFile)
                break
        os.chdir(os.path.dirname(topoPath))

        wb = openpyxl.load_workbook(topoFile)
        ws = wb[wb.sheetnames[0]]
        topoTitles = ws[1]
```

```
        for title in topoTitles:
            if 'SS_NAME' in title.value:
                SS_NAME = ws['{0}:{0}'.format(title.column_letter)]
            if 'MAC' in title.value:
                MAC_Add = ws['{0}:{0}'.format(title.column_letter)]
            if 'COD_CT_SIC' in title.value:
                COD_CT = ws['{0}:{0}'.format(title.column_letter)]
            if 'COD_CGP' in title.value:
                COD_CGP = ws['{0}:{0}'.format(title.column_letter)]
            if 'ADDRESS_CGP' in title.value:
                Add_CGP = ws['{0}:{0}'.format(title.column_letter)]
            if 'Distance' in title.value:
                DistanceSS = ws['{0}:{0}'.format(title.column_letter)]
            if 'Type' in title.value:
                TypeBTLine = ws['{0}:{0}'.format(title.column_letter)]

        devices = {}
        for fileName in os.listdir(os.path.join(self.path, "CC")):
            os.chdir(os.path.join(self.path, "CC"))
            deviceName = fileName[0:5]
            if fnmatch.fnmatch(fileName, '*_10m_AVL_Network_Avg_Parameters.csv'):
                devices[deviceName]= ({'name': deviceName, 'AVL': None, 'RSSI':
None, 'SNR': None, 'MAC': None, 'ADDRESS_CGP': None, 'DistanceSS': None,
'TypeBTLine': None, 'COD_CT': None, 'COD_CGP': None})

            if fnmatch.fnmatch(fileName,
deviceName+'_10m_AVL_Network_Avg_Parameters.csv'):
                AVLdata=[]
                with open(deviceName+'_10m_AVL_Network_Avg_Parameters.csv', 'r')
as file:
                    reader = csv.reader(file)
                    for i, row in enumerate(reader):
                        formRow = [row[7]]
                        if i!=0:
                            for number in row[0:6]:
                                if number != '':
                                    number = [float(number)*100]
                                else:
                                    number = ['']
                                formRow.extend(number)
                        else:
                            formRow.extend(row[0:6])
                        AVLdata.append(formRow)
                AVLdata.remove(AVLdata[0])
                devices[deviceName]['AVL'] = AVLdata

            if fnmatch.fnmatch(fileName,
deviceName+'_10m_RSSI_Network_Avg_Parameters.csv'):
                RSSIdata=[]
                with open(deviceName+'_10m_RSSI_Network_Avg_Parameters.csv', 'r')
as file:
                    reader = csv.reader(file)
```

```
                    for i,row in enumerate(reader):
                        formRow=[row[7]]
                        if i!=0:
                            for number in row[0:6]:
                                if number != '':
                                    number = [float(number)]
                                else:
                                    number = ['']
                                formRow.extend(number)
                        else:
                            formRow.extend(row[0:6])

                        RSSIdata.append(formRow)
                RSSIdata.remove(RSSIdata[0])
                devices[deviceName]['RSSI'] = RSSIdata

            if fnmatch.fnmatch(fileName,
deviceName+'_10m_SNR_Network_Avg_Parameters.csv'):
                SNRdata = []
                with open(deviceName + '_10m_SNR_Network_Avg_Parameters.csv',
'r') as file:
                 reader = csv.reader(file)
                 for i,row in enumerate(reader):
                     formRow = [row[7]]
                     if i != 0:
                         for number in row[0:6]:
                             if number != '':
                                 number = [float(number)]
                             else:
                                 number = ['']
                             formRow.extend(number)
                     else:
                         formRow.extend(row[0:6])

                     SNRdata.append(formRow)
                SNRdata.remove(SNRdata[0])
                devices[deviceName]['SNR'] = SNRdata

            for MAC in MAC_Add:
                if (deviceName.replace('_',':') in MAC.value or
deviceName.replace('_',':').lower() in MAC.value) and self.name in
SS_NAME[MAC.row-1].value:
                    devices[deviceName]['MAC'] = MAC.value
                    devices[deviceName]['ADDRESS_CGP'] = Add_CGP[MAC.row-1].value
                    devices[deviceName]['DistanceSS'] = DistanceSS[MAC.row-
1].value
                    devices[deviceName]['TypeBTLine'] = TypeBTLine[MAC.row-
1].value
                    devices[deviceName]['COD_CT'] = COD_CT[MAC.row-1].value
                    devices[deviceName]['COD_CGP'] = COD_CGP[MAC.row-1].value
        return devices

    def GetChannelsAVL(self):
```

```
        os.chdir(os.path.join(self.path, "CC"))
        for fileName in os.listdir(os.path.join(self.path, "CC")):
            if fnmatch.fnmatch(fileName,
'Channels_AVL_10m_Network_Avg_Parameters.csv'):
                ChannelAVL = []
                with open('Channels_AVL_10m_Network_Avg_Parameters.csv', 'r') as
file:
                    reader = csv.reader(file)
                    for i, row in enumerate(reader):
                        formRow = [row[7]]
                        if i != 0:
                            for number in row[0:6]:
                                if number != '':
                                    number = [float(number) * 100]
                                else:
                                    number = ['']
                                formRow.extend(number)
                        else:
                            formRow.extend(row[0:6])
                        ChannelAVL.append(formRow)
                ChannelAVL.remove(ChannelAVL[0])
        return ChannelAVL
```

9.1.3 PERFORMANCE INPUT DATA EXAMPLE

The input data is a .csv format file with the data titles in the first row and information in the rest. If there is no data available, there will be two commas together.

Canal 3,Canal 4,Canal 5,Canal 6,Canal 7,Canal 8,DayMinutes,Time
0.619,0.872,0.883,0.886,,,0,00:00
0.667,0.9,0.967,0.913,,,10,00:10
0.65,0.844,0.817,0.836,,,20,00:20
0.5,0.956,0.833,0.883,,,30,00:30
0.65,0.9,0.852,0.906,0.864,,40,00:40
0.684,1.0,0.722,0.867,0.867,,50,00:50
0.584,0.947,0.8,0.717,0.783,,60,01:00

0.583,0.949,0.815,0.878,0.872,,70,01:10
0.6,0.943,0.95,0.817,0.883,,80,01:20
0.7,0.955,0.817,0.9,0.856,,90,01:30
0.806,0.952,0.759,0.796,0.817,,100,01:40
,0.938,0.9,0.85,0.878,,110,01:50
,0.9,0.711,0.817,0.728,,120,02:00
,0.956,0.844,0.833,0.906,,130,02:10
,0.922,0.795,0.9,0.733,,140,02:20
,0.928,0.772,0.9,0.917,,150,02:30
,1.0,0.875,0.867,0.817,,160,02:40
,0.989,0.742,0.983,0.717,,170,02:50
,1.0,0.8,0.95,0.784,,180,03:00
,0.922,0.883,0.914,0.648,,190,03:10
,0.95,0.85,0.875,0.65,,200,03:20
,0.967,0.87,0.886,0.817,,210,03:30
,0.983,0.867,0.837,0.817,,220,03:40
,0.989,0.883,0.886,0.8,,230,03:50
,0.947,0.833,0.828,0.8,,240,04:00
,1.0,0.683,0.917,0.766,,250,04:10
,0.944,0.704,0.883,0.797,,260,04:20

,0.948,0.806,0.967,0.828,,270,04:30
,0.972,0.817,0.906,0.784,,280,04:40
,0.974,0.917,0.967,0.636,,290,04:50
,0.954,0.9,0.95,0.537,,300,05:00
,0.942,0.95,0.867,0.667,,310,05:10
,0.993,0.917,0.856,0.839,,320,05:20
,0.967,0.906,0.833,0.767,,330,05:30
,0.95,0.93,0.833,0.606,,340,05:40
,0.895,0.778,0.933,0.867,,350,05:50
,0.95,0.783,0.8,0.648,,360,06:00
,0.95,0.822,0.9,0.834,,370,06:10
0.75,1.0,0.847,0.839,0.733,,380,06:20
0.8,0.95,0.855,0.967,0.889,,390,06:30
0.87,0.928,0.739,0.881,0.767,,400,06:40
0.717,0.933,0.878,0.913,0.6,,410,06:50
0.717,0.95,0.778,0.902,0.767,,420,07:00
0.8,0.944,0.85,0.889,0.733,,430,07:10
0.771,0.967,0.861,0.971,0.667,,440,07:20
0.833,0.833,0.771,0.936,0.717,,450,07:30
0.867,0.922,0.983,0.917,0.765,,460,07:40

0.815,0.95,0.722,0.922,0.684,,470,07:50
0.933,0.961,,0.989,0.739,,480,08:00
0.783,0.932,,0.883,0.706,,490,08:10
0.889,0.933,,0.867,0.695,,500,08:20
0.933,0.917,,0.867,0.747,,510,08:30
0.611,0.917,,0.778,0.817,,520,08:40
0.556,0.989,,0.833,0.783,,530,08:50
0.7,1.0,,0.883,0.75,,540,09:00
0.759,,0.762,0.883,0.784,,550,09:10
0.778,,0.675,0.911,0.85,,560,09:20
0.642,,0.792,0.883,0.883,,570,09:30
0.611,,0.719,0.733,0.867,,580,09:40
0.669,,0.772,0.883,0.878,,590,09:50
0.673,,0.8,0.783,0.92,,600,10:00
0.656,,0.822,0.96,0.883,,610,10:10
0.722,,0.817,0.923,0.883,,620,10:20
0.462,,0.767,0.937,0.839,,630,10:30
0.611,,0.722,0.922,0.833,,640,10:40
0.5,0.861,0.917,0.943,0.7,,650,10:50
0.426,0.917,0.9,0.97,0.75,,660,11:00

0.574,0.9,0.783,0.911,0.646,,670,11:10
0.389,0.872,0.741,0.911,0.628,,680,11:20
0.396,1.0,0.784,0.928,0.703,,690,11:30
0.639,0.972,1.0,0.983,0.764,,700,11:40
0.5,0.889,0.967,0.983,0.483,,710,11:50
0.433,0.9,0.8,0.944,0.561,,720,12:00
0.416,0.892,0.792,0.917,0.856,,730,12:10
0.381,0.933,0.741,0.947,0.633,,740,12:20
0.416,0.95,0.772,0.883,0.63,,750,12:30
0.611,0.933,0.767,0.895,0.75,,760,12:40
0.54,0.883,0.833,0.917,0.667,,770,12:50
0.542,0.562,0.8,0.867,0.722,,780,13:00
0.537,0.911,0.833,0.867,0.784,,790,13:10
0.722,0.922,0.889,0.85,0.783,0.852,800,13:20
0.7,0.939,0.938,0.629,0.75,0.85,810,13:30
0.639,0.889,,0.861,0.722,0.759,820,13:40
0.656,0.9,,0.904,0.706,0.889,830,13:50
0.722,0.95,,0.949,0.683,0.933,840,14:00
0.744,0.833,,0.863,0.772,0.926,850,14:10
0.75,0.942,,0.944,0.817,0.834,860,14:20

0.917,0.961,,0.911,0.8,0.8,870,14:30
0.767,0.883,,0.989,0.833,0.784,880,14:40
0.717,0.989,,0.917,0.728,0.75,890,14:50
0.833,0.942,,0.95,0.802,0.833,900,15:00
0.759,0.877,,0.972,0.797,0.817,910,15:10
0.8,0.9,,0.922,0.745,0.907,920,15:20
0.717,0.933,0.833,0.867,0.828,0.815,930,15:30
0.783,0.833,0.85,0.956,0.703,0.852,940,15:40
0.811,0.922,0.85,0.822,0.567,0.685,950,15:50
0.8,0.933,0.85,0.9,0.688,0.772,960,16:00
0.87,0.928,0.75,0.8,0.542,0.7,970,16:10
0.783,0.967,0.9,0.784,0.5,0.75,980,16:20
0.883,0.9,0.854,0.917,0.683,0.703,990,16:30
0.867,0.917,0.65,0.883,0.634,0.744,1000,16:40
0.789,0.85,0.667,0.883,0.65,0.883,1010,16:50
0.694,0.767,0.833,0.928,0.571,0.75,1020,17:00
0.79,1.0,0.8,0.917,0.738,0.717,1030,17:10
0.661,1.0,0.571,0.722,0.634,0.63,1040,17:20
0.825,0.9,0.667,0.87,0.519,0.709,1050,17:30
0.658,0.967,0.759,0.852,0.567,0.6,1060,17:40

0.75,0.928,0.65,0.867,0.667,0.625,1070,17:50
0.833,0.941,0.75,0.741,0.562,0.684,1080,18:00
0.63,0.917,0.9,0.889,0.574,0.704,1090,18:10
0.803,0.945,0.822,,0.628,0.767,1100,18:20
0.783,0.954,0.907,,0.636,0.722,1110,18:30
0.815,0.908,0.85,,0.531,0.617,1120,18:40
0.85,0.889,0.6,,0.506,0.593,1130,18:50
0.639,0.967,0.683,,0.589,0.811,1140,19:00
0.889,0.889,0.767,,0.606,0.733,1150,19:10
0.967,0.95,0.817,,0.567,0.634,1160,19:20
1.0,0.95,0.794,,0.667,0.717,1170,19:30
0.939,0.906,0.795,,0.567,0.722,1180,19:40
0.926,0.978,0.878,,0.717,0.817,1190,19:50
0.817,0.983,0.85,0.767,0.695,0.783,1200,20:00
0.84,0.983,0.846,0.8,0.517,0.796,1210,20:10
0.87,0.95,0.783,0.883,0.583,,1220,20:20
0.933,0.967,0.85,0.844,0.5,,1230,20:30
0.906,0.944,0.833,0.933,0.7,,1240,20:40
0.861,0.983,0.815,0.883,0.667,,1250,20:50
0.856,0.967,0.593,0.833,0.5,,1260,21:00

0.767,0.983,0.933,0.767,0.634,,1270,21:10
1.0,0.972,0.733,0.833,0.625,,1280,21:20
,0.928,0.7,0.8,0.633,,1290,21:30
,0.978,0.767,0.8,0.783,,1300,21:40
,0.949,0.939,0.833,0.684,,1310,21:50
,0.969,0.95,0.806,0.667,,1320,22:00
,0.948,0.883,0.817,0.733,,1330,22:10
,0.967,0.878,0.917,0.817,,1340,22:20
,0.961,0.867,0.917,0.583,,1350,22:30
,0.967,0.783,0.889,0.667,,1360,22:40
,0.989,0.742,0.933,,,1370,22:50
,0.95,0.831,0.939,,,1380,23:00
,0.85,0.859,0.85,,,1390,23:10
,0.967,0.756,0.917,,,1400,23:20
,1.0,0.906,0.967,,,1410,23:30
,0.978,0.922,0.906,,,1420,23:40
,0.95,0.983,0.888,,,1430,23:50

9.2 TOPOLOGICAL VARIABLES INPUT DATA

DistanceSS	TypeBTLin	COD_CT	COD_CALL	NUM_CLIE	NUM_CLIE	POT_CON	DistanceSS	COORD_X	COORD_Y	COD_LINE	COD_BRIG	NUM_CLIE	NUM_CLIE	POT_CONT
96	Undergrou	2E+08	4,8E+13	25	0	90,3	96,01	4788501	4788501	24688	315	521	35	2037,12
14	Undergrou	2E+08	4,8E+13	31	1	114,85	12,36	4788590	4788590	24735	315	521	35	2037,12
95	Undergrou	2E+08	4,8E+13	16	0	54,7	94,43	4788551	4788551	24714	315	521	35	2037,12
45	Undergrou	2E+08	4,8E+13	34	2	115,8	49,2	4788594	4788594	24779	315	521	35	2037,12
112	Undergrou	2E+08	4,8E+13	17	2	61,25	104,49	4788515	4788515	24745	315	521	35	2037,12
25	Undergrou	2E+08	4,8E+13	31	1	111,75	30,28	4788592	4788592	24762	315	521	35	2037,12
82	Undergrou	2E+08	4,8E+13	37	2	138,46	85,85	4788597	4788597	24771	315	521	35	2037,12
70	Undergrou	2E+08	4,8E+13	12	0	44,95	61,31	4788560	4788560	24674	315	521	35	2037,12
67	Undergrou	2E+08	4,8E+13	17	1	57,5	58,03	4788537	4788537	24705	315	521	35	2037,12
162	Undergrou	21110030	2,81E+13	12	1	50,15	148,64	4481582	4481582	1824658	491	547	44	2790,83
180	Undergrou	21110030	2,81E+13	23	1	97,93	179,55	4481562	4481562	1824653	491	547	44	2790,83
220	Undergrou	21110030	2,81E+13	50	2	215,95	214,58	4481529	4481529	1824635	491	547	44	2790,83
220	Undergrou	21110030	2,81E+13	50	2	215,95	214,58	4481529	4481529	1824635	491	547	44	2790,83
15	Undergrou	21110030	2,81E+13	52	2	221,58	8,62	4481481	4481481	1824648	491	547	44	2790,83
55	Undergrou	20100001	2,81E+13	40	0	153,05	56,61	4480698	4480698	1820647	491	431	9	1785,2
96	Undergrou	20100001	2,81E+13	23	0	86,95	93,8	4480725	4480725	1820656	491	431	9	1785,2
22	Undergrou	20100001	2,81E+13	72	3	295,45	19,32	4480672	4480672	1820631	491	431	9	1785,2
55	Undergrou	20100001	2,81E+13	32	1	156,7	56,58	4480700	4480700	1820642	491	431	9	1785,2
80	Undergrou	20100001	2,81E+13	18	2	98,75	77,47	4480670	4480670	1820675	491	431	9	1785,2
96	Undergrou	20100001	2,81E+13	48	1	200,95	93,78	4480727	4480727	1820651	491	431	9	1785,2
106	Undergrou	20100001	2,81E+13	71	1	296,85	105,73	4480722	4480722	1820661	491	431	9	1785,2
70	Undergrou	20100001	2,81E+13	72	1	299,25	67,82	4480695	4480695	1820671	491	431	9	1785,2
11	Undergrou	46160039	2,81E+13	26	1	103,9	7,1	4472848	4472848	1779629	972	634	46	2594,59
140	Undergrou	46160039	2,81E+13	22	2	95,63	137,09	4472710	4472710	2109232	972	634	46	2594,59
20	Undergrou	46160039	2,81E+13	16	0	51,6	90,11	4472807	4472807	1779616	972	634	46	2594,59
76	Undergrou	46160039	2,81E+13	30	0	118,55	72,43	4472782	4472782	1779605	972	634	46	2594,59
78	Overhead	46160039	2,81E+13	23	1	93,5	75,27	4472806	4472806	1779623	972	634	46	2594,59
60	Undergrou	5050499	4,63E+13	50	3	194,65	108,48	4372117	4372117	1281322	577	408	27	1800
103	Overhead	5050499	4,63E+13	23	2	102,39	104,72	4372106	4372106	1282069	577	408	27	1800
115	Overhead	5050499	4,63E+13	31	1	120,4	118,71	4372111	4372111	1165177	577	408	27	1800
78	Overhead/	5050499	4,63E+13	30	3	108,85	76,79	4372176	4372176	1165186	577	408	27	1800
143	Undergrou	5050499	4,63E+13	28	3	155,13	136,21	4372105	4372105	1169356	577	408	27	1800
84	Undergrou	15150017	2,81E+13	32	1	135,65	84,37	4470867	4470867	1756052	478	378	28	1777,88
100	Undergrou	15150017	2,81E+13	1	0	1,72	98,57			1756047	478	378	28	1777,88
82	Undergrou	15150017	2,81E+13	13	1	64,2	84,02	4470766	4470766	1840121	478	378	28	1777,88
82	Undergrou	15150017	2,81E+13	16	1	75,35	96,78	4470761	4470761	1840116	478	378	28	1777,88

9.3 CHANNEL COMBINATION MATLAB SCRIPT

```
%% Load Variables
clc
clear
close all
format long

formulas = ["Mean","Std","Median"];
PerformanceData = featureCalc(formulas);
allCT = PerformanceData.CT;
allDevices = PerformanceData.Device;
Results = removevars(PerformanceData,["CT" "Device"]);
```

```
Other = readtable("DatosPCA.xlsx",'Sheet','Other');
OldParameters =
readtable("DatosPCA.xlsx",'Sheet','Parameters','PreserveVariableNames',true);
Parameters = [];

for i=1:length(allDevices)
     ctIdx = contains(Other.SSName,replace(allCT(i),'_',' 
'))|contains(Other.SSName,upper(allCT(i)));
     macIdx =
contains(Other.MAC,replace(allDevices(i),'_',':'))|contains(Other.MAC,replace(low
er(allDevices(i)),'_',':'));
     index = ctIdx & macIdx;
     Parameters = [Parameters; OldParameters(index,:)];
end

TypeLineBin = string(Parameters.TypeBTLine(:,1))=="Underground";
ParamMAT = [Parameters.DistanceSS_Calc TypeLineBin Parameters.COD_CT
Parameters.POT_CONTRATADA Parameters.POT_TPM45 Parameters.NUM_CLIENTES_TRIF
Parameters.NUM_CLIENTES_TPM45 Parameters.COD_LINEA Parameters.COD_BRIGADA];
ParamMATName = ["DistanceSS-Calc" "TypeBTLine" "POT-CONT-CGP"];

for i = formulas
    eval(['Res' char(i) '=
splitvars(removevars(Results,formulas(formulas~=i)),i);'])
    eval(['Avail' char(i) '= Res' char(i) '{:,1:6};'])
    eval(['RSSI' char(i) '= Res' char(i) '{:,7:12};'])
    eval(['SNR' char(i) '= Res' char(i) '{:,13:18};'])
end

% AVAILMatrix = [Results.AVAIL3 Results.AVAIL4 Results.AVAIL5 Results.AVAIL6
Results.AVAIL7 Results.AVAIL8];
% RSSIMatrix = [Results.RSSI3 Results.RSSI4 Results.RSSI5 Results.RSSI6
Results.RSSI7 Results.RSSI8];
% SNRMatrix = [Results.SNR3 Results.SNR4 Results.SNR5 Results.SNR6 Results.SNR7
Results.SNR8];

DistanceSS_Calc = ParamMAT(:,1);
TypeBTLine = categorical(string(Parameters.TypeBTLine(:,1)));
COD_CT = categorical(ParamMAT(:,3));
POT_CONT_CGP = ParamMAT(:,4);
COD_LINEA = categorical(ParamMAT(:,8));
COD_BRIG = categorical(ParamMAT(:,9));


%% PCA
close all

PC1 = [];
PC2 = [];
PC1Coeff = [];
PC2Coeff = [];
muMat = [];
```

```
revData = [];
for i = 1:6
    PCAMat = [AvailMean(:,i), SNRMean(:,i), AvailMedian(:,i), SNRMedian(:,i),
AvailStd(:,i),SNRStd(:,i)];
    %plot3(PCAMat(:,1),PCAMat(:,2),PCAMat(:,3),'o','LineWidth',2)
    [coeff,score,latent,tsquared,explained,mu]=pca(PCAMat);
    PC1 = [PC1 score(:,1)];
    PC2 = [PC2 score(:,2)];
    PC1Coeff = [PC1Coeff coeff(:,1)];
    PC2Coeff = [PC2Coeff coeff(:,2)];
    muMat = [muMat;mu];
    revData = [revData [PC1(:,i) PC2(:,i)]*[PC1Coeff(:,i)
PC2Coeff(:,i)]'+muMat(i,:)];
    figure;
    %bar(coeff(:,1:2)); title(strcat("Channel ",num2str(i))); legend("PC1","PC2")
end

ChannelsV = [3:8].*ones(1,length(PC1))';

pareto(explained)
title("% Explained by each component")

% SELECT 2 DATA FOR CLUSTERING (AVAIL-SNR  PC1-PC2)
% Data1 = SNRMedian;%normalize(SNRMatrix,'range',[0,1]);
% Data2 = AvailMedian;%normalize(AVAILMatrix,'range',[0,1]);
% NameData1 = 'SNR';
% NameData2 = 'AVAIL';

Data1 = PC2;
Data2 = PC1;
NameData1 = 'PC2';
NameData2 = 'PC1';

figure;
p = plot3(ChannelsV,Data1,Data2,'o','LineWidth',2);
xlabel('Channel'); ylabel (NameData1); zlabel (NameData2); title (strcat('All
Nodes ',NameData1,'-',NameData2));
view(3);
colors = [p(1).Color;p(2).Color;p(3).Color;p(4).Color;p(5).Color;p(6).Color];
% hold off;

%% Clustering
qe=[];
for k = 1:15
    [idx, ctrs, sumd] = kmeans([Data1,Data2],k, 'distance','sqeuclidean');
    qe(k)=sum(sumd);
end
figure; bar(qe);axis tight;
xlabel('K'); ylabel('QE');

%Run kmeans
K=3;
[idx, ctrs, sumd] = kmeans([Data1,Data2],K, 'distance','sqeuclidean');
```

```
% Plot the prototypes (centers) of the clusters
figure;
plot3(3:8,ctrs(:,1:6)',ctrs(:,7:12)','-o','linewidth',4);
title(strcat("Centers of clusters ",NameData1,'-',NameData2))
legend(strcat('Cluster ',num2str([1:K]')))
xlabel('Channel'); ylabel (NameData1); zlabel (NameData2);

% Plot the prototypes (centers) of the clusters by variable
%Reverse from PC to var
revCtrs = [];
for i = 1 : 6
    revCtrs = [revCtrs [ctrs(:,i) ctrs(:,i+6)]*[PC1Coeff(:,i)
PC2Coeff(:,i)]'+muMat(i,:)];
end

if NameData1 == 'PC2'
    figure;

plot3(3:8,revCtrs(:,1:size(PCAMat,2):size(PCAMat,2)*6),revCtrs(:,2:size(PCAMat,2)
:size(PCAMat,2)*6),'-o','linewidth',4);
    xlabel('Channel'); ylabel ('AVAIL'); zlabel ('SNR');
    title("Centers of clusters Mean");
    legend(strcat('Cluster ',num2str([1:K]')))
    if size(PCAMat,2)>2
        figure;

plot3(3:8,revCtrs(:,3:size(PCAMat,2):size(PCAMat,2)*6),revCtrs(:,4:size(PCAMat,2)
:size(PCAMat,2)*6),'-o','linewidth',4);
        xlabel('Channel'); ylabel ('AVAIL'); zlabel ('SNR');
        title("Centers of clusters Median");
        legend(strcat('Cluster ',num2str([1:K]')))
    end
    if size(PCAMat,2)>4
        figure

plot3(3:8,revCtrs(:,5:size(PCAMat,2):size(PCAMat,2)*6),revCtrs(:,6:size(PCAMat,2)
:size(PCAMat,2)*6),'-o','linewidth',4);
        xlabel('Channel'); ylabel ('AVAIL'); zlabel ('SNR');
        title("Centers of clusters STD");
        legend(strcat('Cluster ',num2str([1:K]')))
    end

end
%% Nodes in each cluster
newcolors = {'#F00','#F80','#FF0','#0B0','#00F','#80F'};
markersL = {'-o','-x','-s','-^','-p','-d'};
markers = {'o','x','s','^','p','d'};

for i = 1:K
    colororder(colors);
    index = idx==i;
    figure;
```

```
p=plot3(ChannelsV(index,:)',Data1(index,:)',Data2(index,:)',markersL{1},'LineWidt
h',2);
    view(3)
    title(strcat('Cluster ',num2str(i)));
    xlabel('Channel'); ylabel (NameData1); zlabel (NameData2);
    set(gca,'Xtick',3:8)

    figure;

h=plot3(ChannelsV(index,:)',revData(index,4:6:36)',revData(index,3:6:36)',markers
L{1},'LineWidth',2);
    view(3)
    title(strcat('Cluster ',num2str(i)));
    xlabel('Channel'); ylabel ("SNRMedian"); zlabel ("AVAILMedian");
    set(gca,'Xtick',3:8)
end

%% Classification
%close all

idx_str = [];
for i = idx
    idx_str = strcat("Cluster ",num2str(i));
end

classData = table(DistanceSS_Calc,TypeBTLine,POT_CONT_CGP,'VariableNames',
ParamMATName');

%Get the training and Test sets
[X_TR, X_TS, y_TR, y_TS] = trainingAndTestSets(classData, idx_str, 0.5,'Random');
tree = fitctree(X_TR,y_TR,'CategoricalPredictors',logical([0 1 0])
,'PredictorNames',ParamMATName','AlgorithmForCategorical','Exact');

classData.Labels = idx_str;
writetable(classData,'ClassificationData.csv');

Ctree = crossval(tree);
imp = [];
best = 0;
for i = 1:10
    %Calculate importance of predictors
    imp = [imp predictorImportance(Ctree.Trained{i})'];
    %Calculate Ccuracy of each tree Da problema el Ctree.Trained con
    %TypeBTLine
    acc = 100*(1-loss(tree,X_TR,y_TR));
    if acc > best
        best = acc;
        best_ind = i;
    end
end
```

```
figure; bar(mean(imp')); title('Predictor Importance Estimates');
ylabel('Estimates'); xlabel('Predictors'); h = gca; h.XTickLabel =
tree.PredictorNames;

tree_def = tree;%Ctree.Trained{best_ind,1};
view(tree_def,'mode','graph');
Nodes = [tree_def.Parent tree_def.NodeSize]
%Now compute the errors of the tree
y_est = predict(tree_def,X_TR);

figure;
confusionchart(string(y_est(:,1)),y_TR);
title("Trainning set")

%compute classification error
disp(['Classification tree: TRAINING Set Folded ERROR: ',
num2str(100*kfoldLoss(Ctree)),'%']);
disp(['Classification tree: TRAINING Set Folded ACCURACY: ', num2str(100*(1-
kfoldLoss(Ctree))),'%']);
disp(['Classification best tree: TRAINING Set  ACCURACY: ', num2str(best),'%']);

% Test the classifier in the TEST SET
y_est_TS = predict(tree_def,X_TS);

figure;
confusionchart(string(y_est_TS(:,1)),y_TS);
title("Test Set")

%compute classification error
disp(['Classification tree: TEST Set ERROR: ',
num2str(100*loss(tree_def,X_TS,y_TS)),'%)']);
disp(['Classification tree: TEST Set ACCURACY: ', num2str(100*(1-
loss(tree_def,X_TS,y_TS))),'%)']);
```

9.4 CHANNEL BY CHANNEL MATLAB SCRIPT

```
%% Load Variables
clc
clear
close all
format long

formulas = ["Mean","Std","Median"];
PerformanceData = featureCalc(formulas);
allCT = PerformanceData.CT;
allDevices = PerformanceData.Device;
Results = removevars(PerformanceData,["CT" "Device"]);

Other = readtable("DatosPCA.xlsx",'Sheet','Other');
```

```
OldParameters =
readtable("DatosPCA.xlsx",'Sheet','Parameters','PreserveVariableNames',true);
Parameters = [];

for i=1:length(allDevices)
     ctIdx = contains(Other.SSName,replace(allCT(i),'_',' 
'))|contains(Other.SSName,upper(allCT(i)));
     macIdx =
contains(Other.MAC,replace(allDevices(i),'_',':'))|contains(Other.MAC,replace(low
er(allDevices(i)),'_',':'));
     index = ctIdx & macIdx;
     Parameters = [Parameters; OldParameters(index,:)];
end


TypeLineBin = string(Parameters.TypeBTLine(:,1))=="Underground";
ParamMAT = [Parameters.DistanceSS_Calc TypeLineBin Parameters.COD_CT
Parameters.POT_CONTRATADA Parameters.POT_TPM45 Parameters.NUM_CLIENTES_TRIF
Parameters.NUM_CLIENTES_TPM45 Parameters.COD_LINEA Parameters.COD_BRIGADA];
ParamMATName = ["DistanceSS-Calc" "TypeBTLine" "POT-CONT-CGP"];

for i = formulas
    eval(['Res' char(i) '=
splitvars(removevars(Results,formulas(formulas~=i)),i);'])
    eval(['Avail' char(i) '= Res' char(i) '{:,1:6};'])
    eval(['RSSI' char(i) '= Res' char(i) '{:,7:12};'])
    eval(['SNR' char(i) '= Res' char(i) '{:,13:18};'])
end


TypeLineBin = string(Parameters.TypeBTLine(:,1))=="Underground";
ParamMAT = [Parameters.DistanceSS_Calc TypeLineBin Parameters.COD_CT
Parameters.POT_CONTRATADA Parameters.POT_TPM45 Parameters.NUM_CLIENTES_TRIF
Parameters.NUM_CLIENTES_TPM45 Parameters.COD_LINEA Parameters.COD_BRIGADA];
ParamMATName = ["DistanceSS-Calc" "TypeBTLine" "POT-CONT-CGP"];

DistanceSS_Calc = ParamMAT(:,1);
TypeBTLine = categorical(string(Parameters.TypeBTLine(:,1)));
COD_CT = categorical(ParamMAT(:,3));
POT_CONT = ParamMAT(:,4);

colors=[0,0.447,0.741;0.85,0.325,0.098;0.929,0.694,0.125;0.494,0.184,0.556];
%% Bucle para agrupar y clasifiar cada vez por un canal

for channel = 1:6

    % Normalize the variables introduced in PCA
    PCAMat = [AvailMean(:,channel), SNRMean(:,channel), AvailMedian(:,channel),
SNRMedian(:,channel), AvailStd(:,channel),SNRStd(:,channel)];
    %plot3(PCAMat(:,1),PCAMat(:,2),PCAMat(:,3),'o','LineWidth',2)
    [coeff,score,latent,tsquared,explained,mu]=pca(PCAMat);
    % Select PC1 scores
    PC1 = score(:,1);
```

```
    PC2 = score(:,2);
    PC1Coeff = coeff(:,1);
    PC2Coeff = coeff(:,2);
    revData = [PC1 PC2]*[PC1Coeff PC2Coeff]'+mu;

    %Check xomponents explanation
    figure(1);pareto(explained);
    title("% Explained by each component")

%     SELECT 2 DATA FOR CLUSTERING (AVAIL-SNR  PC1-PC2)
%      Data1 = normalize(SNRMean(:,channel),'range',[0,1]);
%      Data2 = normalize(AvailMean(:,channel),'range',[0,1]);
%      NameData1 = 'SNR';
%      NameData2 = 'AVAIL';

    Data1 = PC2;
    Data2 = PC1;
    NameData1 = 'PC2';
    NameData2 = 'PC1';

    % Plot all nodes
    figure(2);
    plot(Data1',Data2','o','LineWidth',2);
    xlabel(NameData1);ylabel (NameData2); title (strcat("All Nodes ",NameData1,'-
',NameData2," Channel ",num2str(channel+2)));
    %axis([-0.4 0.4 -1.5 1])

    % Check quantization error (get K)
    qe=[];
    for k = 1:15
        [idx, ctrs, sumd] = kmeans([Data1,Data2],k, 'distance','sqeuclidean');
        qe(k)=sum(sumd);
    end
    figure(3); bar(qe);axis tight;
    xlabel('K'); ylabel('QE');

    %Run kmeans with K number of clusters
    K=[2 2 2 2 2 2];
    [idx, ctrs, sumd] = kmeans([Data1,Data2],K(channel),
'distance','sqeuclidean');

    % Plot the prototypes (centers) of the clusters
%     for i = 1:K(channel)
%         figure(channel*10+1);hold on;
%         plot(ctrs(i,2),ctrs(i,1),'o','LineWidth',4);
%         title("Centers of clusters PC1")
%         legend(strcat('Cluster ',num2str([1:K(channel)]')));
%         axis([-0.4 0.4 -1.5 1]);
%     end
    % Nodes in each cluster
    figure(channel*10+2); hold on;
    for i = 1:K(channel)
        index = idx==i;
```

```
        p = plot(Data1(index==1,:)',Data2(index==1,:)','o','LineWidth',2);
        legend(strcat('Cluster ',num2str([1:K(channel)]')));
        %axis([-1.5 1 -1.5 1]);
        title(strcat("Nodes in Clusters Channel ",num2str(channel+2)));
        xlabel(NameData1);ylabel (NameData2);
        p.Color=colors(i,:);
    end

    % nodes in each cluster but median in axes
    figure(channel*10+3); hold on;
    for i = 1:K(channel)
        index = idx==i;
        p = plot(revData(index==1,4)',revData(index==1,3)','o','LineWidth',2);
        legend(strcat('Cluster ',num2str([1:K(channel)]')));
        %axis([-1.5 1 -1.5 1]);
        title(strcat("Nodes in each cluster: Channel ",num2str(channel+2)));
        xlabel("SNRMedian");ylabel ("AVAILMedian");
        p.Color=colors(i,:);
    end

    % More understable label instead of number
    idx_str = [];
    for i = idx
        idx_str = strcat("Cluster ",num2str(i));
    end

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %% %%%%%%%%%%%%%%%% Clasifier %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    FileName = strcat("ClassdataChannel",num2str(channel+2));
    %Set classification variables in a table
    FileTable= table(DistanceSS_Calc,TypeBTLine,POT_CONT,'VariableNames',
ParamMATName');

    %Get the training and Test sets
    [X_TR, X_TS, y_TR, y_TS] = trainingAndTestSets(FileTable, idx_str,
0.5,'Random');
    tree = fitctree(X_TR,y_TR,'CategoricalPredictors',logical([0 1 0])
,'PredictorNames',ParamMATName','AlgorithmForCategorical','Exact');

    % Write Labels in table and save in csv file
    FileTable.Labels = idx_str;
    writetable(FileTable,strcat(FileName,".csv"));

    Ctree = crossval(tree);
    imp = [];
    for i = 1:10
        %Calculate mean importance of predictors
        imp = [imp predictorImportance(Ctree.Trained{i})'];
    end
    figure(channel*10+4); bar(mean(imp')); title(strcat('Predictor Importance
Estimates: Channel ',num2str(channel+2))); ylabel('Estimates');
xlabel('Predictors'); h = gca; h.XTickLabel = tree.PredictorNames;
```

```
    % Get first tree visualization
    tree_def = tree;
    %view(tree_def,'mode','graph');
    Nodes = [tree_def.Parent tree_def.NodeSize]
    % Predictions with the TRAINED SET
    y_est = predict(tree_def,X_TR);

    % Prediction errors in confusion chart
    figure(channel*10+5);
    confusionchart(string(y_est(:,1)),y_TR);
    title(strcat("Trainning set: Channel ",num2str(channel+2)))

    % Compute classification error
    disp(['Classification tree Channel ',num2str(channel+2),': TRAINING Set
Folded ERROR: ', num2str(100*kfoldLoss(Ctree)),'%']);
    disp(['Classification tree Channel ',num2str(channel+2),': TRAINING Set
Folded ACCURACY: ', num2str(100*(1-kfoldLoss(Ctree))),'%']);
    disp(['Classification Visulized tree Channel ',num2str(channel+2),': TRAINING
Set ACCURACY: ', num2str(100*(1-loss(tree_def,X_TR,y_TR))),'%']);

    % Predictions with TEST SET
    y_est_TS = predict(tree_def,X_TS);

    % See Prediction errors in confusion chart
    figure(channel*10+6);
    confusionchart(string(y_est_TS(:,1)),y_TS);
    title(strcat("Test Set: Channel ",num2str(channel+2)))

    % Compute classification error
    disp(['Classification tree Channel ',num2str(channel+2),': TEST Set ERROR: ',
num2str(100*loss(tree_def,X_TS,y_TS)),'%)']);
    disp(['Classification tree Channel ',num2str(channel+2),': TEST Set ACCURACY:
', num2str(100*(1-loss(tree_def,X_TS,y_TS))),'%)']);
end
```