ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

PREPROCESAMIENTO DE LOS DATOS DE LAS UNIDADES DE MEDICIÓN SINCROFASORIAL (PMUS) UTILIZANDO LA TÉCNICA LIMPIEZA DE DATOS - APLICACIÓN AL SISTEMA NACIONAL INTERCONECTADO ECUATORIANO

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO ELÉCTRICO

DIANA FERNANDA GUEVARA ESTACIO

diana.guevara@epn.edu.ec

DIRECTOR: DR.-ING. NELSON VICTORIANO GRANDA GUTIÉRREZ

nelson.granda@epn.edu.ec

Quito, Enero 2019

AVAL

Certifico que el presente trabajo fue desarrollado por Diana Fernanda Guevara Estacio, bajo mi supervisión.

DR.-ING. NELSON VICTORIANO GRANDA GUTIÉRREZ
DIRECTOR DEL TRABAJO DE TITULACIÓN

DECLARACIÓN DE AUTORÍA

Yo, Diana Fernanda Guevara Estacio, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

_____

DIANA FERNANDA GUEVARA ESTACIO

DEDICATORIA

A mi madre Clara por ser un rayo de luz ante cualquier circunstancia de mi vida, por ser el pilar fundamental de todos mis logros, por creer siempre fielmente en mí y convertirme en la mujer que soy ahora.

A mis hermanos María Elena, David y Sergio, por haberme apoyado incondicionalmente en cada etapa de mi vida, ya que sin ustedes nada de esto sería posible. Son la mejor inspiración para superarme cada día.

AGRADECIMIENTO

Agradezco a mi madre por ser el motor principal de cada paso que doy en mi vida, por todo su amor y apoyo incondicional que solo ella sabe como entregarlo, por estar siempre en mis éxitos y fracasos, por hacer de mí una mejor persona cada día, por eso y por más te viviré agradecida toda una vida.

A toda mi familia, en especial a mis hermanos María Elena, David y Sergio, quienes siempre supieron como ser un soporte en mi vida, gracias porque más que ser mis hermanos cada uno de ustedes fue un padre para mí, a mis cuñados Luis y Jenny, por impartirme siempre su sabiduría para actuar ante las adversidades, mis sobrinos Katherine, Leidy, Richard y Esteban, por ser mi motivación para culminar esta meta.

A mis profesores Franklin Quilumba, Nelson Granda y Diego Echeverría, por ser mis mentores e impartirme sus conocimientos y brindarme su amistad, a lo largo de este trayecto para culminar esta meta.

A mis amigos, Elvis por ser mi persona incondicional ante cualquier circunstancia y en especial gracias por aportarme tus conocimientos en mi vida universitaria, Andrea por todos tus consejos, los cuales me han permito salir adelante en la vida, Samy por ser una persona indispensable y especial, la cual me ha enseñado algo cada día, y a todos mis amigos con quienes he convivido y compartido tantas experiencias que marcaron mi vida a lo largo de esta carrera, por brindarme su amistad, enseñanzas y apoyo incondicional, espero que esta amistad perdure para siempre.

ÍNDICE DE CONTENIDO

RESUMEN

En el presente Proyecto de Titulación se aplican las técnicas de “Limpieza de datos” sobre la base de datos de las mediciones sincrofasoriales de frecuencia, tasa de frecuencia, fasores de secuencia positiva de las ondas sinusoidales de voltaje y corriente de las PMUs del Sistema Nacional Interconectado. En primera instancia, se realiza una descripción detallada del conjunto de tareas o procedimientos que comprenden esta técnica, las cuales hacen uso de diversas funciones o metodologías para detectar, diagnosticar, e imputar datos anómalos, con la finalidad de crear una base de datos fidedigna e influir en la calidad de los resultados de posteriores análisis o estudios.

Por consiguiente, se implementa una aplicación para el desarrollo de la limpieza de datos de las diferentes mediciones sincrofasoriales en App Designer del software MATLAB. Dicha aplicación se encuentra divida en secciones o etapas que permiten el desarrollo de esta técnica, y cada una de ellas ejecutan una rutina que permiten: el manejo apropiado de la base de datos, descripción de las características de la señal temporal seleccionada, tratamiento de datos vacíos, filtrado de ruido, detección e imputación de datos atípicos, con el objetivo de resolver las inconsistencias de una forma interactiva para el usuario.

Finalmente, se obtienen los errores del conjunto de metodologías empleadas a lo largo del proceso de limpieza de datos, para así seleccionar y validar las más adecuadas dependiendo de la señal analizada. De este modo, se obtiene el preprocesamiento del conjunto de datos de las diferentes señales temporales.

PALABRAS CLAVE: base de datos, mediciones sincrofasoriales, PMU, datos anómalos, limpieza de datos, imputación de datos.

ABSTRACT

This work focuses on applying "data cleaning" techniques on a database of synchrophasor measurement units (PMUs) of the Ecuadorian National Interconnected System. Data include frequency, rate of frequency, phasor voltage and current positive sequence sinusoidal waves. In the first instance, a detailed description of the set of tasks or procedures that comprise the cleaning techniques is presented. Such diverse functions or methodologies help detecting, diagnosing, and imputing anomalous data, with the purpose of creating a reliable database and influencing on the quality of the results of subsequent analyzes or studies.

Consequently, an application is developed for data cleaning of different synchrophasor measurements in MATLAB's App Designer software. This application is divided into sections or stages that allow the implementation of these techniques, and each of them executes a routine that: appropriately manages the database, describes the characteristics of the selected time signal, missing data processing, noise filtering, detection and imputation of atypical data, with the aim to resolve the inconsistencies in an interactive way for the user.

Finally, the errors are obtained of the set of methodologies applied of the cleaning process. This helps to select and validate the most suitable methods depending on the analyzed signal. Thus, Pre-processed data set is obtained of the different time series signals.

KEYWORDS: database, synchrophasor measurements, PMU, anomalous data, data cleaning, data cleansing, data imputation.

1. INTRODUCCIÓN

En la actualidad, para una operación segura y confiable de la red eléctrica se ha desarrollado nuevas tecnologías, como son las Unidades de Medición Fasorial (PMUs), con la finalidad de mejorar el monitoreo, protección y control de los Sistemas Eléctricos de Potencia. La inclusión de las PMUs han permitido la estimación de mediciones instantáneas y precisas de frecuencia, variación de cambio de frecuencia, fasores de secuencia positiva de las ondas sinusoidales de voltaje y corriente de una red eléctrica en tiempo real, independientemente del estado operativo en el que esta se encuentre.

Estos medidores inteligentes PMUs, permiten la obtención y suministro de los datos fasoriales de las diferentes señales de acuerdo con una estampa de tiempo. El uso de estos datos es de suma importancia, ya que facilita el diseño de técnicas o mecanismos que permitirán evaluar la vulnerabilidad del sistema eléctrico de potencia, a través de análisis en tiempo real o después de ocurrida una contingencia.

La limpieza de datos es un término sobrecargado y a menudo se usa de manera general para referirse a una variedad de tareas destinadas a mejorar la calidad de los datos. Estas tareas se consiguen mediante la unión de varias operaciones y en el presente proyecto se analizan algunas técnicas comunes de limpieza de datos para comprender mejor las operaciones subyacentes.

Previo al uso de los datos fasoriales en cualquier análisis, surge la necesidad de realizar su preprocesamiento, a través de técnicas de limpieza de datos, debido a que en el proceso los datos son altamente susceptibles a la presencia de datos anómalos, lo que representa importantes dificultades para futuros estudios.

En este contexto, en el presente estudio técnico se analiza la base de datos de las diferentes mediciones sincrofasoriales de las PMUs del Sistema Nacional Interconectado, desarrollando una aplicación computacional mediante el software MATLAB, la cual tiene como propósito aplicar la técnica de limpieza de datos a las diferentes señales temporales. De esta forma se busca influir en la calidad de los resultados de cualquier estudio y así satisfacer los requisitos de su uso previsto.

1.1. Objetivos

El objetivo general de este Proyecto Integrador es:

- Realizar el preprocesamiento de datos de las mediciones obtenidas de las Unidades de Medición Sincrofasorial (PMUs) utilizando diferentes técnicas de limpieza de datos a través de la herramienta de software MATLAB aplicado al Sistema Nacional Interconectado.

Los objetivos específicos de este Proyecto Integrador son:

- Analizar las características y comportamiento de la base de datos extraída del PDC, tanto para las medidas de voltaje, corriente, frecuencia y ángulo de fase, mediante la aplicación de la estadística descriptiva.
- Identificar datos anómalos tales como datos faltantes, inconsistentes, atípicos, ruido en series de tiempo en la base de datos de las mediciones sincrofasoriales del SNI.
- Evaluar distintos algoritmos con enfoques estadísticos autorregresivos, probabilísticos paramétricos y no paramétricos, machine learning relacionado con el aprendizaje no supervisado y filtros aplicados a series de tiempo para la limpieza de las mediciones sincrofasoriales del SNI.
- Implementar una herramienta de software MATLAB, que permita el análisis y evaluación del preprocesamiento de datos de las mediciones sincrofasoriales del SNI.
- Realizar una comparación de las diferentes metodologías aplicadas a las mediciones sincrofasoriales registradas para de esta manera escoger la más adecuada para la limpieza de datos.

1.2. Alcance

En el presente estudio técnico en primera instancia se analizará la base de datos de mediciones estimadas de frecuencia, tasa de frecuencia, fasores de secuencia positiva de las ondas sinusoidales de voltaje y corriente del Sistema Nacional Interconectado extraídas de las Unidades de Medición Sincrofasorial (PMUs), a través de la estadística descriptiva la cual permitirá tener una descripción, interpretación y representación de los datos para observar las características y el comportamiento de los mismos, con la finalidad de encontrar anomalías en dichas series de tiempo.

Se estudiará distintos métodos o algoritmos con enfoques estadísticos autorregresivos, probabilísticos paramétricos y no paramétricos, machine learning relacionado con el aprendizaje no supervisado y filtros para series de tiempo que permitirán la imputación de datos, eliminación de ruido y suavizado de las señales para todas mediciones sincrofasoriales de las PMUs.

Se implementará y simulará un módulo con la herramienta de software MATLAB, para así preprocesar a los datos registrados de las unidades de mediciones sincrofasoriales de las PMUs instaladas en el SNI a través de la técnica de limpieza de datos.

2. MARCO TEÓRICO

En esta sección se presenta el sustento teórico del proyecto a través de la descripción de las características, funcionamiento, ubicación y aplicaciones de las Unidades de Medición Sincrofasorial (PMUs). De igual forma, se realiza una descripción del preprocesamiento de datos utilizando la técnica de limpieza de datos en series de tiempo, haciendo énfasis en los diferentes enfoques estadísticos, probabilísticos y machine learning que involucran el desarrollo de esta técnica.

2.1. Análisis Fasorial y Sincrofasorial

En la actualidad los sistemas eléctricos de potencia (SEPs), operan cada vez más cerca de sus límites máximos de transferencia, debido a un continuo crecimiento en el consumo de energía eléctrica e instalación de generación. Debido a estos factores se reducen los márgenes de estabilidad y surge la necesidad de una mejor supervisión de las redes eléctricas para garantizar una operación segura y confiable, independientemente de su estado operativo [1], [2]. La introducción de las Unidades de Medición Fasorial (PMU - Phasor Measurement Units), adquiere importancia estratégica para el monitoreo y control de los sistemas eléctricos de potencia en tiempo real [3]. Las PMUs realizan mediciones instantáneas y precisas, de tal forma que constituyen un sistema de medición, que suministra datos necesarios para el control íntegro de sistemas de potencia. Esto ha alentado su proliferación en sistemas eléctricos de potencia en todo el mundo, por lo que se enfatiza la necesidad de una mayor investigación, exploración e implementación en las áreas de medición, protección y control [4].

2.1.1. Análisis fasorial

Un fasor es una representación vectorial constante de una función sinusoidal, asumiendo que la amplitud, frecuencia y fase permanecen constantes [5]. El concepto de usar fasores es introducir una forma de describir sintéticamente una señal sinusoidal en redes AC, asumiendo una frecuencia nominal constante, por consiguiente se aplica directamente el concepto de fasor, esto se remonta a la investigación de Charles Proteus Steinmetz, en 1983 [6].

La definición matemática clásica de fasor se basa en una señal de corriente alterna genérica $x(t)$ que se representa por la Ecuación 2.1.

$$x(t) = X_m \cos(\omega t + \phi)$$

Ecuación 2.1. Representación de una onda sinusoidal de corriente alterna

Donde X_m es el valor pico de la señal, $\omega = 2\pi f$ es la frecuencia angular del sistema, f es la frecuencia instantánea y ϕ es el ángulo de fase de la señal. Empleando la identidad de Euler, la Ecuación 2.1. puede ser expresada como:

$$x(t) = Re\{X_m e^{j(\omega t+\phi)}\}$$

$$x(t) = Re\{X_m e^{j\phi}[e^{j\omega t}]\}$$

En condiciones de estado estable el término $e^{j\omega t}$ puede suprimirse, ya que la frecuencia se considera un parámetro constate. Bajo esta consideración dicha expresión puede representarse mediante un número complejo $\overline{X}$ que gira a la velocidad angular ω, conocido como su *representación fasorial* o simplemente *fasor* [7], como se muestra en la Ecuación 2.2.

$$x(t) \Leftrightarrow \overline{X} = \left(\frac{X_m}{\sqrt{2}}\right) e^{j\phi}$$

$$\overline{X} = \left(\frac{X_m}{\sqrt{2}}\right)(\cos\phi + j\sin\phi) = X_r + jX_i$$

Ecuación 2.2. Representación fasorial

Dónde: X_r y X_i son las componentes rectangulares real e imaginaria de la representación fasorial, con su respectivo módulo $X_m/\sqrt{2}$ que corresponde al valor RMS (Root Mean Square) de la señal sinusoidal. La Figura 2.1. muestra el instante de tiempo en el cual la longitud X_m va rotando a ω rad/s. A medida que la longitud X_m rota, su proyección instantánea a lo largo del eje horizontal es el valor instantáneo del fasor. Suponiendo que la longitud X_m está alineada con el eje horizontal en el tiempo "t = 0", entonces se asocia que el ángulo de fase es 0°. Por lo tanto, se puede expresar el fasor como $\overline{X_1} = (X_m/\sqrt{2})\angle 0$. Ahora suponiendo que en el tiempo "t = 0", al rotar la longitud X_m esta se encuentra en el ángulo ϕ° respecto al eje horizontal, entonces se asocia que el ángulo de fase es ϕ°, expresando el fasor como $\overline{X_2} = (X_m/\sqrt{2})\angle\phi$, de esta manera se muestra la representación fasorial de una onda sinusoidal dada por la Ecuación 2.2. [8].

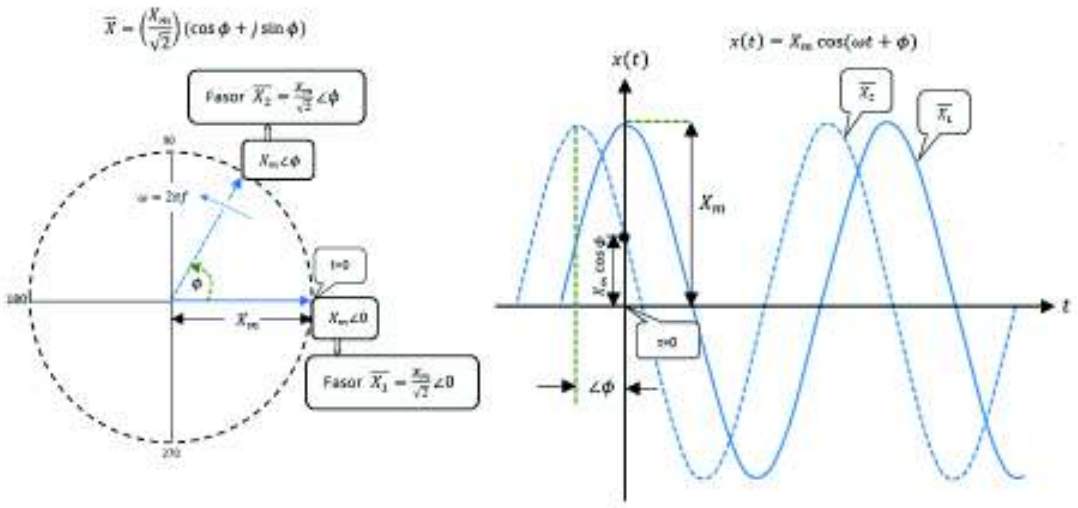


Figura 2.1. Representación fasorial de una onda sinusoidal

2.1.2. Análisis sincrofasorial

Un *sincrofasor* o *fasor sincronizado* se define como "un fasor calculado a partir de datos muestreados usando una señal de tiempo estándar como la referencia para la medición" [9]. Es decir, la estimación del sincrofasor se basa en la misma idea subyacente de fasor, con la diferencia principal que el ángulo de fase se calcula utilizando el Tiempo Universal Coordinado (UTC – Coordinated Universal Time) como una referencia de tiempo. Tal elección permite tener una referencia única para todas las señales sinusoidales a medir en cualquier parte de la red eléctrica, debido a que la difusión del tiempo depende de los sistemas de satélite [6].

La expresión de sincrofasor tiene la misma representación de la Ecuación 2.2. la diferencia es que en este caso el ángulo de fase ϕ es el desplazamiento entre la onda $x(t)$ con una forma de onda cosenoidal ficticia, la misma que tiene una frecuencia igual a la frecuencia nominal del sistema y es sincronizada al UTC. Dicha onda cosenoidal tiene su punto máximo en un instante de tiempo absoluto, que se toma como referencia "t = 0". Al comienzo de cada segundo (que se remonta a este instante "t = 0"), una señal conocida como "1 PPS" (1 pulso por segundo) se transmite desde el GPS (Sistema de Posicionamiento Global), entonces cuando el valor máximo de la función $x(t)$ coincida con la señal "1 PPS" el ángulo de fase ϕ tendrá un valor de 0°, y cuando el cruce por cero positivo de la función $x(t)$ coincida con la señal "1 PPS" tendrá un valor de -90°, tal como se muestra en la Figura 2.2. [8], [10].

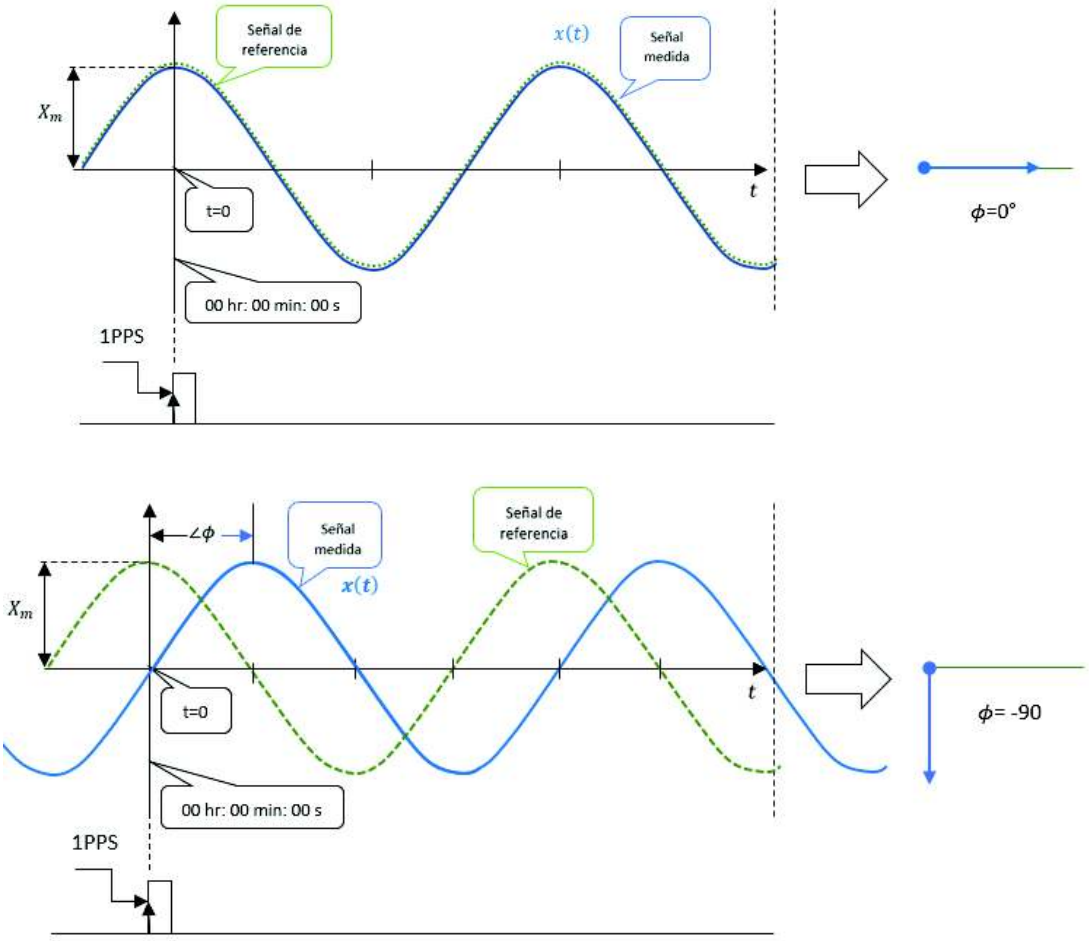


Figura 2.2. Convención para representación de sincrofasores

2.2. Tecnología de medición sincrofasorial

2.2.1. Unidades de medición sincrofasorial (PMU)

El concepto de una unidad de medición sincrofasorial (PMU), también conocida como sincrofasor, fue desarrollado a finales de los años 80 por el Dr. Arun G. Phadke y el Dr. James S. Thorp [6]. Una unidad de medición sincrofasorial es un dispositivo electrónico moderno de medición y protección que permite calcular o estimar medidas de frecuencia, tasa de frecuencia, fasores de secuencia positiva de las ondas sinusoidales de voltaje y corriente de la red eléctrica a una tasa de muestreo de 60 muestras por segundo, independientemente del estado operativo en el que se encuentre el sistema [10].

2.2.2. Elementos de una PMU

Algunos de los elementos principales de una PMU son:

2.2.2.1. Filtro Anti-aliasing

El filtro pasa-bajos anti-aliasing analógico se emplea comúnmente en todos los sistemas de estimación fasorial, el objetivo de este filtro es limitar las señales de entrada a menos de la mitad de la frecuencia de muestreo [11].

2.2.2.2. Conversor Analógico/Digital

Convierte los valores de las señales eléctricas analógicas de voltaje y corriente en valores digitales [11].

2.2.2.3. Microprocesador

El microprocesador calcula las estimaciones de todas las señales de corriente y voltaje de secuencia positiva, mediante el uso de un algoritmo de estimación de fasor. Estos algoritmos emplean N muestras en un período de tiempo específico para llevar a cabo la estimación fasorial. El algoritmo más comúnmente aplicado es la Transformada Discreta de Fourier (DFT) y/o algoritmos propietarios [12].

2.2.2.4. Oscilador de enganche de fase

Es un oscilador que mantiene constantes a la frecuencia de la referencia y la señal medida [9].

2.2.2.5. Receptor de señales GPS

El receptor de señales GPS permite la sincronización de las lecturas de las medidas a una misma referencia de tiempo tomadas en puntos distantes [12].

2.2.2.6. Elementos primarios de medición TC y TP

En el sistema de potencia los valores de voltaje y corriente son elevados, debido a esto su medida no se puede obtener en forma directa. Por tal razón se usan los transformadores de medida, los cuales reproducen un valor proporcional de la magnitud eléctrica del sistema de potencia, a su vez se utilizan para aislar los equipos de control, protección y medida de los altos voltajes de los circuitos primarios [13].

Los transformadores de medida son:

Transformadores de corriente (TC)

Reducen la alta corriente del circuito de potencia a una corriente baja, la cual se puede llevar sin peligro a los aparatos de protección y medida [14].

Transformadores de potencial (TP)

Reducen el alto voltaje del circuito de potencia a bajo voltaje [14].

2.2.2.7. Sistema de posicionamiento global (GPS)

El sistema de posicionamiento global es un sistema de navegación basado en satélites que proporciona información sobre la ubicación y tiempo de cualquier punto del planeta, también es capaz de proporcionar la señal de un pulso por segundo "1 PPS", la cual indica el inicio de cada segundo del Tiempo Universal Coordinado (UTC) con una exactitud de alrededor de 1μs, siendo esta muy importante para marcar la referencia de tiempo para la estimación de fasores [15].

2.2.2.8. Concentrador de datos fasoriales (PDC)

Un Concentrador de Datos de Fasores (PDC - Phasor Data Concentrator) funciona como un nodo en una red de comunicación donde los datos de sincrofasores de un número de PMU o PDC se procesan y transmiten como un único flujo a los PDC y / o aplicaciones de nivel superior. El PDC procesa los datos del sincrofasor por marca de tiempo para crear un conjunto de medidas para todo el sistema [16]. De acuerdo con la utilización y ubicación de los PDCs pueden ser locales, regionales y centrales o SuperPDC. En la Tabla 2.1. se describe la clasificación de los PDCs.

Tabla 2.1. Clasificación de los PDCs

CLASIFICACIÓN DE LOS PDCs	
NOMBRE	**CARACTERÍSTICA**
Locales	Los PDCs locales se ubican dentro de la misma subestación, receptando y sincronizando en el tiempo los datos de los sincrofasores de todas las PMUs de la subestación, estos son enviados a otros PDCs como los regionales o centrales. Una de las ventajas de este tipo de PDC es que si ocurre una falla en la comunicación los datos son guardados en el historial del PDC local, así no hay pérdida de información [5].
Regionales	Los PDCs regionales concentran y sincronizan en el tiempo los datos de sincrofasores de múltiples PDCs locales o PMUs de una determinada área eléctrica, y las envían al PDC central, por lo que poseen mayor capacidad que los locales [12].
Central o Super-PDC	El PDC central concentra los datos de sincrofasores de todas las PMUs instaladas en el sistema eléctrico, ya sea que los datos se receptan desde los PDC regionales, locales o directamente desde las PMUs. A su vez sincroniza los datos en el tiempo para su almacenamiento histórico, y conjuntamente se envía la información de todo el sistema a las aplicaciones de análisis [12].

2.2.3. Funcionamiento de una PMU

Las entradas analógicas son las ondas sinusoidales de corrientes y voltajes obtenidas de los devanados secundarios de los transformadores de corriente y de potencial. Las tres señales de corrientes y voltajes se utilizan para que se pueda llevar a cabo las mediciones de secuencia positiva a la salida de la PMU. Se procede a filtrar cada señal analógica mediante el uso del filtro anti-aliasing, y luego son enviadas al convertidor analógico/digital. Para mantener la frecuencia constante de las señales se utiliza el reloj de muestreo, el cual se encuentra sincronizado en fase con la señal de un pulso por segundo "1 PPS" proporcionada por un receptor del GPS a través del oscilador de enganche de fase. Las señales obtenidas del convertidor analógico/digital (típicamente se encuentran dentro del rango de ± 10 V) son enviadas al microprocesador junto con sus respectivas estampas de tiempo. El microprocesador mediante el algoritmo Transformada Discreta de Fourier, método más usado para la estimación de fasores, permite calcular los fasores de todas las señales de corriente y voltaje de secuencia positiva a través de N muestras en un período de tiempo específico. Finalmente se transmite está información mediante el puerto de comunicación hacia el concentrador de datos fasoriales, para el caso de CENACE el PDC es de tipo Super-PDC, en la Figura 2.3. se puede observar el esquema de funcionamiento de una PMU [8], [10], [12].

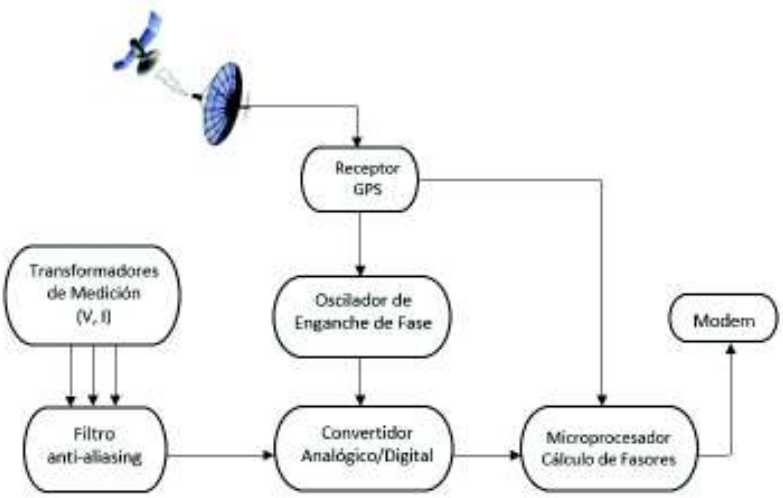


Figura 2.3. Estructura básica de una PMU

2.2.4. Aplicaciones de las PMUs

El sistema de potencia está distribuido en un área geográfica muy grande normalmente la mayoría de los relés de protección hacen uso de la información disponible localmente, con tiempos de respuesta de varios milisegundos, suficiente para la protección local.

Actualmente los sistemas SCADA permiten una visualización integral de los sistemas eléctricos de potencia y debido a las limitaciones en la frecuencia de muestreo proporcionan una visión relativamente estática. La introducción de las unidades de medición sincrofasorial permite la consideración del control basado en el valor medido de cantidades remotas, la información de diversos lugares puede estar disponible en cualquier parte del sistema mediante la integración de las PMUs y el sistema SCADA, permitiendo así una visión dinámica del sistema de potencia y a su vez una gran cantidad de aplicaciones novedosas que pueden proporcionar protección y control de área amplia [9], [12], [17]. Algunas de estas aplicaciones se detallan a continuación:

2.2.4.1. Estimador de estado

El estimador de estado se basa en la medición de los flujos de potencia activa y reactiva y de magnitudes de voltaje en los nodos más importantes de la red, para luego comunicarlas a un sitio central para su procesamiento, los datos se adquieren en una ventana de tiempo de segundos a minutos [17]. Por lo tanto, el estado calculado es, en el mejor de los casos, una aproximación a un estado promedio del sistema, y las estimaciones que se producen por el programa de estimación de estado se denominan estimaciones de estado estático.

Con la inclusión de las PMUs, el estimador de estado podría obtener mejores resultados, a través de las mediciones sincronizadas de los fasores de voltaje y corriente de secuencia positiva en cada barra del sistema, sin la necesidad de correr flujos de potencia y sin conocer otros parámetros de la red. Cada ubicación clave del sistema de potencia se equipa con mediciones de las PMUs y la información está disponible en una ubicación central. Por tanto no es necesario tener PMUs en cada ubicación para realizar la estimación del estado, debido a que la red se divide en "ubicaciones observables" donde se encuentran PMUs y "ubicaciones no observables" donde no se encuentran PMUs, pero las estimaciones pueden calcularse indirectamente, a partir de los valores estimados en ubicaciones observables, y así realizar estimaciones cercanas en ubicaciones no observables [8].

2.2.4.2. Estabilidad de voltaje a corto plazo

La estabilidad de voltaje se refiere a la capacidad de un sistema eléctrico de potencia para mantener voltajes constantes en todas las barras del sistema, después de estar sujeto a una perturbación a partir de una condición inicial de operación dada [1], [17]. La estabilidad de voltaje tiene un fuerte acoplamiento con el flujo de potencia reactiva Q, es decir, un sistema es estable en voltaje, si la sensibilidad V-Q es positiva para cada barra, y es inestable si la sensibilidad V-Q es negativa al menos en una barra del sistema.

El intervalo de tiempo de interés es del orden de varios segundos y las herramientas más utilizadas para analizar la estabilidad de voltaje en los sistemas eléctricos de potencia son las curvas de Potencia-Voltaje (PV) y la determinación del margen de cargabilidad, dado por la capacidad de transferencia disponible [18].

Con la inclusión de la tecnología sincrofasorial se han desarrollado nuevas metodologías que han permitido el monitoreo de la estabilidad de voltaje en sistemas eléctricos de potencia en tiempo real. Una de las nuevas técnicas desarrolladas es el método equivalente de Thevenin, el cual permite estimar la curva P-V de corredores de transmisión en barras de envío (B1) y de recepción (B2) en las que se encuentran instaladas PMUs como se muestra en la Figura 2.4. [17], [19].

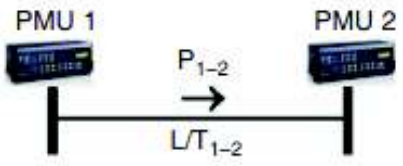


Figura 2.4. Corredor de transmisión monitoreado por PMUs [17]

2.2.4.3. Estabilidad de frecuencia

La estabilidad de frecuencia se refiere a la capacidad de un sistema eléctrico de potencia para conservar una frecuencia constante, es decir, dentro de los rangos aceptables de operación, luego de un disturbio severo el cual produce un desequilibrio significativo entre la generación y la carga [1].

La inestabilidad de la frecuencia a corto plazo se caracteriza por la aparición de un área con déficit de generación, en la cual la generación tiende a ser menor que la carga, porque no hay un deslastre suficiente de carga, por otro lado, cuando se desconecta una gran cantidad de carga, se puede formar un área con superávit de generación, por todo esto la frecuencia sube o decae rápidamente, causando la desconexión de carga y/o generación en pocos segundos [10]. En estos casos, los problemas de la inestabilidad de frecuencia podrían provocar un apagón en pocos segundos o la interrupción del funcionamiento de las unidades de generación. La desviación de frecuencia de su valor nominal es un claro indicador del efecto dinámico producido por la contingencia. Así cuanto mayor es la desviación de la frecuencia, mayor es el efecto producido por la contingencia. Por tal motivo, el uso de las unidades sincrofasoriales resulta indispensable para monitorear la frecuencia en tiempo real y así brindar al operador señales de alerta necesarias para la toma de decisiones [17].

2.2.4.4. Estabilidad estática de ángulo

La estabilidad de ángulo se refiere a la capacidad de las máquinas sincrónicas de un sistema de potencia para permanecer en sincronismo después de haber sido sometidas a una perturbación. Este tipo de estabilidad depende de la capacidad de mantener o restaurar el equilibrio entre el torque electromagnético y torque mecánico en cada generador del sistema [10].

Los cambios en el torque electromagnético de una máquina síncrona se expresan en función de dos componentes: el "torque sincronizante" (en fase con la desviación del ángulo del rotor $\Delta\delta$) y el "torque amortiguador" (en fase con la desviación de velocidad $\Delta\omega$), como se muestra en la Ecuación 2.3. [10].

$$\Delta T_e = T_s \,.\, \Delta\delta + T_D \,.\, \Delta\omega$$

Ecuación 2.3. Representación del torque electromagnético

Dónde: $T_s\ \Delta\delta$ representa el torque sincronizante y $T_D \Delta\omega$ representa el torque amortiguador.

El concepto de estabilidad de ángulo implica analizar la relación entre el intercambio de potencia y la posición angular de los rotores de las máquinas sincrónicas. La diferencia angular entre dos barras del sistema de potencia es una medida directa de la capacidad de transmisión entre estos nodos. Por lo cual las PMUs se ubican en barras y subestaciones para garantizar una adecuada supervisión para la estabilidad estática de ángulo [1], [17].

2.2.4.5. Estabilidad transitoria

La estabilidad transitoria es la capacidad de un sistema eléctrico de potencia para mantener el sincronismo cuando es sometido a una gran perturbación como un cortocircuito, pérdida de la línea o unidad de generación [10]. Esta estabilidad se manifiesta usualmente en forma de separación angular aperiódica debido a la falta de torque sincronizante, sin embargo, en grandes sistemas eléctricos de potencia, se presenta debido a la falta de torque sincronizante y/o amortiguador, después de la ocurrencia de varias oscilaciones. El intervalo de tiempo en el que este fenómeno se desarrolla suele ser 3 - 5 segundos después de la perturbación. En este sentido las PMUs no entregan una medición directa del ángulo interno del generador, por lo que se hace uso de algoritmos que involucran a las mediciones sincrofasoriales para la obtención de dicho ángulo [17].

2.2.4.6. Estabilidad oscilatoria

La estabilidad oscilatoria es la capacidad de un sistema eléctrico de potencia para mantener el sincronismo cuando es sometido a pequeñas perturbaciones como variaciones entre la carga y generación, las cuales producen pequeños cambios en el ángulo del generador, velocidad y potencia. Esta estabilidad se manifiesta usualmente debido a la falta del torque de amortiguación en un intervalo de tiempo de 10 - 20 segundos después de la perturbación.

Las oscilaciones que se presentan pueden ser de naturaleza local o global. Los problemas locales (oscilaciones de modo local) están asociados con oscilaciones entre los rotores de unos pocos generadores cercanos entre sí. Estas oscilaciones tienen frecuencias en el rango de 0.7 a 2.0 Hz. Los problemas globales (oscilaciones de modo inter-área) son causados por interacciones entre grandes grupos de generadores. Estas oscilaciones tienen frecuencias en el rango de 0.1 a 0.7 Hz. Existen otros dos tipos de problemas de oscilación provocados por los controladores de diferentes componentes del sistema (modos de control) o por los componentes de rotación del sistema del eje turbina-generador [1], [8], [17].

Las técnicas para analizar las oscilaciones en un sistema eléctrico de potencia son: el análisis modal herramienta más utilizada para determinar los modos oscilatorios y sus correspondientes frecuencias, Transformada de Fourier, Análisis Prony, Transformada de Hilbert-Huan, Filtro de Kalman, Transformada de Wavelet, entre otras. Estos algoritmos matemáticos permiten la estimación de los modos oscilatorios a partir de las mediciones de las PMUs [17].

2.2.4.7. Otras aplicaciones

- Sistemas de protección con entradas fasoriales.
- Estudios post-contingencia.
- Determinación de los parámetros de líneas de transmisión, generadores que conforman la red eléctrica.
- Control centralizado de transformadores desfasadores, cambiadores de tomas.

2.2.5. Periodicidad de adquisición de datos de las PMUs

Para la adquisición de información de las PMUs la velocidad de muestreo está restringida a su velocidad de transmisión, cantidad de PMUs instaladas, ancho de banda, capacidad de procesamiento de las aplicaciones [12].

En la actualidad los sincrofasores capturan datos de fasores alrededor de 20 a 120 muestras por segundo, las PMUs modernas incluso pueden tomar hasta 240 mediciones de muestra por segundo [20]. En el Sistema Nacional Interconectado se instalaron PMUs Arbiter en las cuales la tasa máxima de muestreo es de 60 muestras por segundo, por lo que se obtiene un paquete de información con casi 216.000 muestras por hora, lo que equivale alrededor de 5.184.000 muestras por día. Estos números crecen exponencialmente con la adición de nuevas PMUs y se pueden acumular hasta 1,5 Terabytes (TB) de datos en un lapso de un mes. Para no saturar los canales con información adicional que las PMUs pueden calcular directamente y enviarla desde el punto de medición, se configuró cada una de las PMUs para que únicamente envíen los valores de las mediciones de frecuencia, la tasa de cambio de la frecuencia y los fasores de voltaje y corriente, mientras que la aplicación o software correspondiente al PDC en la cual se puede manejar los datos de todas las PMUs instaladas en el SNI, se configuró para calcular valores de potencia, voltajes y corrientes de línea [12], [21].

En la Figura 2.5. se puede observar un ejemplo de la configuración de la PMU ZHORAY - MILAGRO 2.

Figura 2.5. Configuración de la PMU ZHORAY - MILAGRO 2 [12]

En la Figura 2.6. se puede observar la aplicación del PDC con los datos correspondientes a la PMU MOLINO - PASCUALES 1.

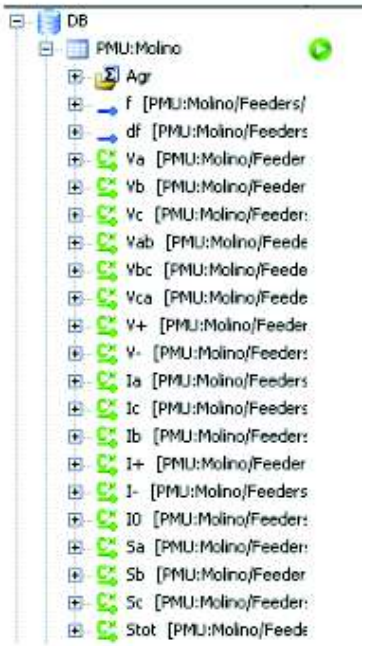


Figura 2.6. Aplicación del PDC con los datos correspondientes a la PMU MOLINO - PASCUALES 1 [12]

2.2.6. Criterios para la Ubicación de las Unidades de Medición Sincrofasorial en el Sistema Nacional Interconectado (S.N.I.)

Para determinar la ubicación de las PMUs en los sistemas eléctricos de potencia, uno de los criterios que se considera es la observabilidad topológica, la cual consiste en emplear la menor cantidad de PMUs, maximizando la redundancia y manteniendo la observabilidad global del sistema. Otro criterio que se considera es el de dar prioridad a las barras más relevantes del sistema eléctrico, y de esta forma enfocar el monitoreo de las variables eléctricas de dichas barras, permitiendo la obtención de mucha información de interés, por ejemplo, los datos asociados a las barras centrales o principales, interconexiones, barras de alta generación o de carga, entre otras [2], [22].

Varios autores han propuesto diferentes metodologías que resuelven la problemática de la ubicación de PMUs en un sistema eléctrico, un ejemplo de estas propuestas se encuentra en [1], [2], [22] todos estos estudios contienen un análisis costo/beneficio, a través de diferentes modelos matemáticos para minimizar la cantidad de PMUs a instalarse. Para la ubicación de las PMUs se identifica los lugares que aseguren observabilidad estática y también los lugares considerados como más relevantes dentro de la operación en un

sistema eléctrico de potencia, de tal forma que se permita monitorear la ocurrencia de fenómenos eléctricos.

En principio, la ubicación de PMUs en el Sistema Nacional Interconectado se realizó de acuerdo con el criterio de barras relevantes, y aprovechando la experiencia operativa [2].

A continuación, se presenta algunos criterios de selección para la ubicación de las PMUs:

- Las barras del anillo troncal del sistema de transmisión del SEP.
- Las barras de los grandes centros de carga del SEP.
- En los nodos asociados a las interconexiones.
- La barra de bornes del generador.
- Las barras de alto y bajo voltaje, para el caso de transformadores de gran capacidad.
- Las subestaciones críticas en la operación del SEP.

2.2.7. Ubicación de las Unidades de Medición Sincrofasorial en el Sistema Nacional Interconectado (S.N.I.)

Las PMUs que se definieron para ser instaladas en el Sistema Nacional Interconectado corresponden a la marca Arbiter modelo 1133 A, para la configuración de este equipo se cuenta con el software Power Sentinel CSV. A su vez se definieron puntos de medición ubicados en las subestaciones del SNI, en base a los criterios mencionados y estudios eléctricos realizados por un grupo de especialistas de CENACE, una vez determinadas las ubicaciones con la coordinación de CELEC EP TRANSELECTRIC se procedió al montaje y puesta en servicio de las PMUs [23].

En la Tabla 2.2. se presenta las PMUs que se encuentran en servicio en el Sistema Nacional Interconectado, y en Figura 2.7. se observa la ubicación geográfica de las PMUs [22].

Tabla 2.2. Ubicación de PMUs en el S.N.I. Ecuatoriano

SUBESTACIÓN	PMU	POSICIÓN	VOLTAJE
AGOYÁN	1	BAÑOS 1	138 kV
C. ESMERALDAS	2	G1	13,8 kV
C. JIVINO III	3	T1	69 kV

C. TRINITARIA	4	TV1	13,8 kV
D. PERIPA	5	PORTOVIEJO 1	138 kV
LOJA	6	VILLONACO	69 kV
MILAGRO	7	SAN IDELFONSO 1	138 kV
MOLINO	8	AT1	138 kV
MOLINO	9	PASCUALES 1	230 kV
MOLINO	10	TOTORAS	230 kV
MONTECRISTI	11	JARAMIJÓ	138 kV
PASCUALES	12	CHONE 1	138 kV
PASCUALES	13	MOLINO 1	230 kV
PASCUALES	14	MOLINO 2	230 kV
POMASQUI	15	JAMONDINO 2	230 kV
POMASQUI	16	JAMONDINO 3	230 kV
C. PUCARÁ	17	T2	138 kV
QUEVEDO	18	ATT	138 kV
QUEVEDO	19	PASCUALES 1	230 kV
SALITRAL	20	ATR	138 kV
STO. DOMINGO	21	BABA	230 kV
STO. DOMINGO	22	ESMERALDAS	138 kV
S. ELENA	23	C. S. ELENA III	69 kV
SANTA ROSA	24	POMASQUI 1	230 kV
SANTA ROSA	25	SANTO DOMINGO 1	230 kV
SANTA ROSA	26	TOTORAS 1	230 kV
SANTA ROSA	27	TOTORAS 2	230 kV
TOTORAS	28	SANTA ROSA 1	230 kV
ZHORAY	29	MILAGRO 2	230 kV

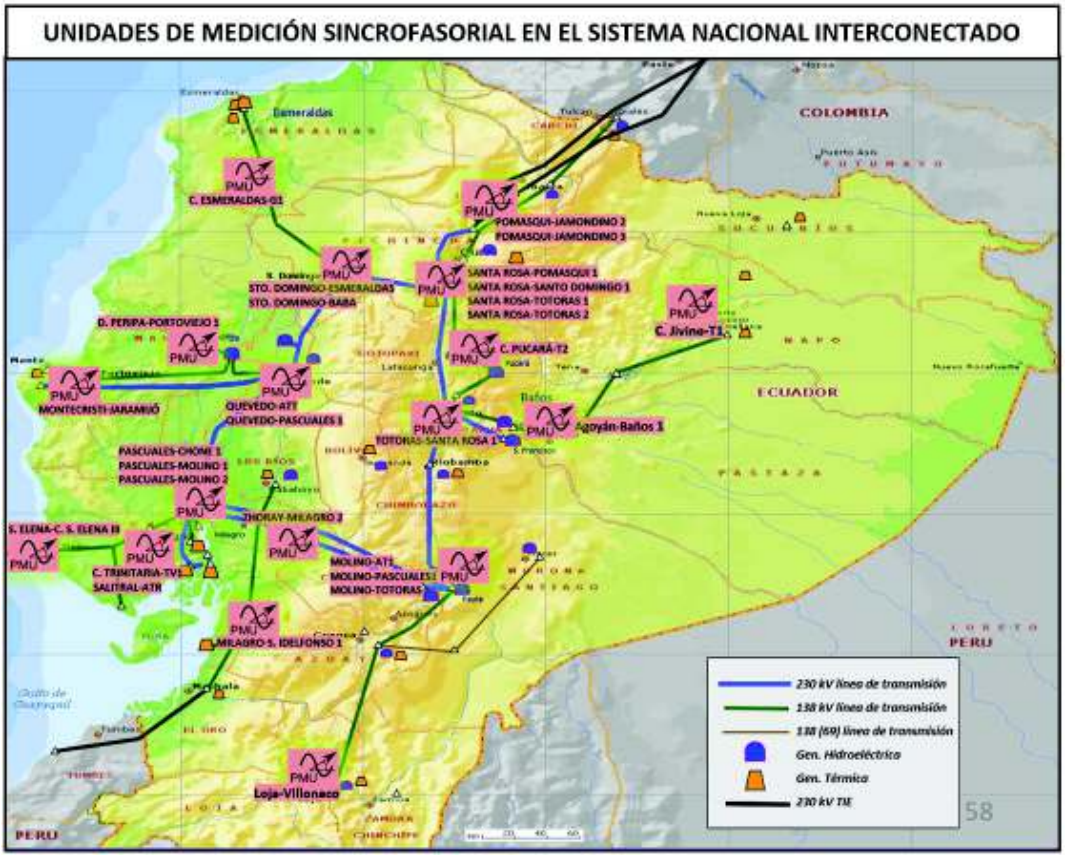


Figura 2.7. Ubicación geográfica de las PMUs

2.3. Preprocesamiento de datos en series temporales

Las bases de datos actuales del mundo real son altamente susceptibles a los datos atípicos, faltantes y al ruido, debido a varios factores, por lo que presenta importantes dificultades al ejecutar cualquier tipo de análisis. Existen varias técnicas de preprocesamiento de datos como: limpieza de datos, integración de datos, reducción de datos. En los últimos años, esta área se ha vuelto de gran importancia debido a que los algoritmos de los diferentes análisis requieren datos significativos y manejables para operar correctamente y proporcionar conocimiento útil, predicciones o descripciones [24].

2.3.1. Limpieza de datos en series temporales

La limpieza de datos es el proceso que consiste en detectar, diagnosticar, resolver inconsistencias y la imputación de datos anómalos. Si a los datos no se aplica la técnica de limpieza, es poco probable que se confíe en los resultados de cualquier análisis que se haya aplicado. Específicamente, esta disertación se centra en la limpieza de datos de series temporales. El entrenamiento de un modelo en una serie de tiempo que contiene datos anómalos generalmente da como resultado parámetros erróneos y sesgados.

A continuación, se explicará los tipos de datos anómalos que serán tratados en la limpieza de datos y se exponen en la Tabla 2.3. [24], [25], [26].

Tabla 2.3. Tipos de datos anómalos

TIPOS DE DATOS ANÓMALOS	
Tipo de datos	**Característica**
Datos faltantes	Los datos faltantes o también llamados datos NaN (Not a Number) se refiere a que no existen valores para una observación específica en un conjunto de datos univariantes o cuando no hay valores de datos para una variable particular de un conjunto de datos multivariantes. La falta de datos se debe principalmente a errores en la recopilación de datos.
Datos atípicos	Los datos atípicos son valores de datos que se desvía significativamente del resto del conjunto de datos, como si hubiera sido generado por un mecanismo diferente.
Datos Ruidosos	El ruido es una señal aleatoria que tiene diferentes frecuencias que van desde frecuencias cero (DC) hasta frecuencias infinitas que se suman a la señal original, lo que le da una densidad espectral de potencia constante, esto significa que la señal contiene todas las frecuencias y todas ellas muestran la misma potencia, por lo que el ruido se considera función de su dinámica de tiempo y frecuencia.

2.3.2. Detección e imputación de datos anómalos

El paso de detección consiste en identificar diferentes tipos de datos anómalos, mientras que el paso de imputación consiste en decidir sobre posibles correcciones para los valores anómalos encontrados [25]. Las técnicas o enfoques de detección e imputación de datos se eligen en función a los tipos de datos anómalos, las cuales puedes ser:

2.3.2.1. Enfoques estadísticos autorregresivos

En los enfoques estadísticos se hacen suposiciones sobre la normalidad de los datos, es decir, los valores de datos normales se generan mediante un modelo estadístico (estocástico) y los datos que no siguen el modelo o se desvían considerablemente de sus valores predichos son valores atípicos. Las anomalías afectan la estructura, los parámetros y la varianza de los modelos. Muchos modelos estadísticos, y en particular los modelos autorregresivos especifican que la variable de salida depende linealmente de sus propios valores anteriores en las series de tiempo [25], [27].

Los modelos autorregresivos proporcionan una descripción parsimoniosa (lenta) de un proceso estocástico (magnitudes aleatorias que varían con el tiempo), donde los parámetros y las constantes del modelo se derivan de los datos [28]. Los modelos tienen dos partes polinomiales, una función autorregresiva que es estacionaria y una función de promedio móvil. Existen varios tipos de modelos estadísticos autorregresivos los más conocidos son AR, ARIMA, SARIMA. La efectividad de los métodos estadísticos depende en gran medida de si las suposiciones hechas para el modelo estadístico son ciertas para los datos dados [29].

2.3.2.2. Enfoques probabilísticos

Los enfoques probabilísticos usan funciones de distribución de probabilidad para ajustar los datos y calcular los parámetros. Estos enfoques identifican valores atípicos como los datos cuya probabilidad es menor que algún umbral elegido, con respecto a la distribución estimada de los datos. Las anomalías en este caso son los datos que se desvían considerablemente de otros miembros de la población. Hay dos tipos de enfoques probabilísticos: paramétrico y no paramétrico. Los métodos paramétricos utilizan funciones de distribución predefinidas que se pueden describir utilizando un número finito de parámetros, mientras que los métodos no paramétricos se estiman por funciones de distancia o densidad a partir de los datos de la serie temporal [25], [29].

2.3.2.3. Enfoques paramétricos

Los enfoques paramétricos son muy susceptibles al ruido y al sobreajuste en los datos, siempre asumen una distribución específica de los datos. Este tipo de modelos suponen que los datos provienen de una familia de distribuciones conocidas [28]. Buzzi-Ferraris y Manenti desarrollaron un enfoque que utiliza las funciones de distribución Gaussianas o la desviación media absoluta (DMA) para la detección de valores atípicos. Todos los puntos de datos con valores superiores al umbral correspondientes a una probabilidad de 0,95 (error del 5%) se consideran valores atípicos. Estas técnicas no tienen en cuenta el número de muestras en el conjunto de datos, lo que arroja muchos falsos positivos para grandes conjuntos de datos. La principal desventaja de los métodos paramétricos es que la mayoría de las distribuciones son univariadas, y la distribución de las observaciones debe conocerse de antemano, además, no existe una regla óptima para elegir o calcular un umbral de rechazo [25], [26].

2.3.2.4. Enfoques no paramétricos

Los métodos basados en la distancia y basados en la densidad son enfoques no paramétricos ampliamente utilizados para la detección de valores atípicos [25].

2.3.2.5. Enfoques basados en distancia

Los enfoques basados en distancia usan la distancia entre un punto y sus vecinos para determinar si el dato es anómalo. Estos enfoques son eficientes en conjuntos de datos multidimensionales, pero son computacionalmente costosos (generalmente el tiempo de ejecución es n^2, donde n es el número de muestras del conjunto de datos) [26].

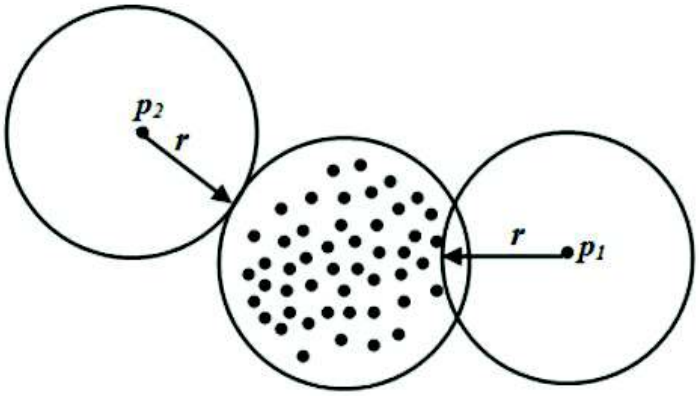


Figura 2.8. Ejemplo de un algoritmo basado en la distancia [25]

En la Figura 2.8. se presenta un ejemplo de un algoritmo basado en la distancia, donde el radio r se calcula a partir de la distribución espacial de los datos. Un dato p se considera como un valor atípico si es mayor al porcentaje $\propto$ que corresponde a todos los otros puntos que tienen una distancia menor que r, el umbral $\propto$ es un parámetro elegido. Los valores atípicos se representan mediante p_1 y p_2 en la Figura 2.8., con un umbral elegido del 1% para 50 puntos de datos. Los enfoques basados en la distancia pueden combinarse con técnicas de agrupamiento, como el vecino más cercano a k para identificar valores atípicos, pero la identificación de una buena medida de distancia es difícil en conjuntos de datos reales [25], [30].

2.3.2.6. Enfoques basados en la densidad

Los enfoques basados en la densidad encuentran anomalías al observar la densidad local del vecindario de un punto. La densidad de un punto de datos se mide por el número de objetos dentro de un área determinada (o volumen). Las técnicas basadas en la densidad detectan los valores atípicos frente a los puntos restantes de la distribución utilizando diferentes medidas, como los factores atípicos locales, la estimación del núcleo y la ventana de Parzen [30].

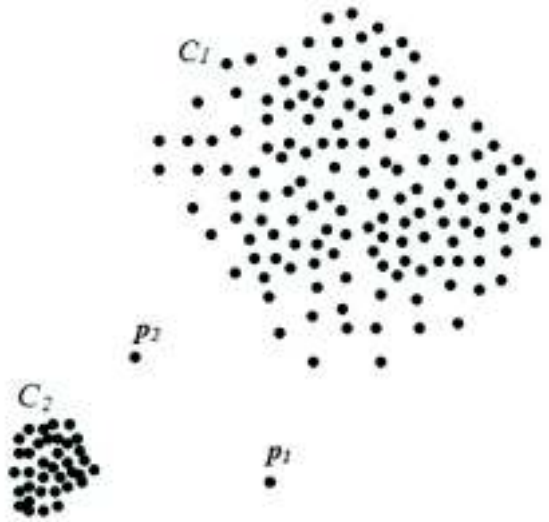


Figura 2.9. Ejemplo de un algoritmo basado en la densidad que utiliza un factor atípico local [25]

En la Figura 2.9. se presenta un ejemplo de un algoritmo basado en la densidad que utiliza un factor atípico local, con dos clústeres C_1 y C_2. De acuerdo con los enfoques de detección de valores atípicos basados en la distancia, todos los puntos en C_2, p_1 y p_2 son atípicos porque el clúster C_1 es predominante. El resultado es erróneo porque hay dos clústeres.

Los datos se califican usando el método basado en la distancia. Los puntos p_1 y p_2 son atípicos porque sus respectivas distancias a C_1 y C_2 son mayores que el radio de los conglomerados. El factor atípico local y las técnicas de estimación del núcleo utilizan un enfoque local basado en la distancia para el cálculo de su distribución. Sin embargo, los métodos de ventana de Parzen interpolan los datos para estimar la distribución de la que se obtuvo la muestra. Los enfoques basados en densidad son computacionalmente costosos para grandes conjuntos de datos y arrojan falsos positivos porque se enfocan en determinar los valores extremos n superiores, donde n es un parámetro elegido [25].

2.3.2.7. Machine Learning

Los enfoques de aprendizaje automático se pueden definir ampliamente como métodos computacionales que utilizan la experiencia para mejorar el rendimiento o para hacer predicciones precisas. La experiencia se refiere a la información pasada disponible para el aprendizaje, que generalmente toma la forma de datos electrónicos recopilados y puestos a disposición para el análisis. Un algoritmo de aprendizaje automático está intrínsecamente relacionado con el análisis de datos y las estadísticas, que tienen por objetivo aprender automáticamente a reconocer patrones complejos y seleccionar decisiones inteligentes basadas en datos.

De manera más general, las técnicas de aprendizaje son métodos basados en datos que combinan conceptos fundamentales en informática con ideas de estadísticas, probabilidad y optimización [26], [31]. Hay dos tipos de tareas de aprendizaje automático: aprendizaje supervisado y aprendizaje no supervisado.

2.3.2.8. Aprendizaje supervisado

Es básicamente un sinónimo de clasificación. La supervisión en el aprendizaje recibe un conjunto de ejemplos etiquetados como datos de entrenamiento y hace predicciones para todos los puntos no vistos. Este es el escenario más común asociado con problemas de clasificación y regresión [25], [31].

2.3.2.9. Aprendizaje no supervisado

Es esencialmente un sinónimo de agrupamiento. El proceso de aprendizaje no supervisado recibe exclusivamente datos de entrenamiento sin etiqueta, y hace predicciones para todos los puntos no vistos. Dado que, en general, no hay ningún ejemplo etiquetado disponible en esa configuración, puede ser difícil evaluar cuantitativamente el rendimiento del aprendizaje. La agrupación y la reducción de dimensionalidad son ejemplos de aprendizaje no supervisados [25], [31].

3. METODOLOGÍA

Este capítulo presenta una descripción detallada de las metodologías implementadas en la aplicación, cada una de estas contienen diferentes enfoques para la evaluación, detección e imputación de datos anómalos. Prácticamente todos los algoritmos o modelos contienen diferentes propuestas, las cuales usan las desviaciones de patrones de los datos para calcular las anomalías de un conjunto de datos establecido.

Finalmente, se describe las diferentes etapas, secciones o fases metodológicas de la aplicación desarrollada para el preprocesamiento de los datos de las mediciones sincrofasoriales a través de la herramienta de software MATLAB. La aplicación consta de las siguientes etapas: selección de la señal, estadística descriptiva, tratamiento de datos NaN, filtrado de la señal, tratamiento de datos atípicos, resultados de la señal preprocesada. Una vez finalizada la ejecución de todo el procedimiento de la técnica de limpieza de datos de una señal, se realiza el cálculo de los errores, este parámetro depende de la combinación de los métodos escogidos en cada una de las etapas de la aplicación, para la obtención de un mejor resultado.

3.1. Estadística descriptiva en series temporales

La estadística descriptiva permite recopilar, analizar, interpretar o explicar y representar los datos. Un modelo estadístico es un conjunto de funciones matemáticas que describen el comportamiento de los datos en términos de variables aleatorias y sus distribuciones de probabilidad asociadas [32].

3.1.1. Descripciones estadísticas básicas de los datos

3.1.1.1. Medidas de Tendencia Central

Media

La medida numérica más común y efectiva del "centro" de un conjunto de datos es la media aritmética, o simplemente llamada media, que corresponde al valor promedio de una distribución [26], [32]. Sea x_1, x_2,..., x_N un conjunto de N valores u observaciones para un atributo numérico X, la Ecuación 3.1. representa la media de este conjunto de valores:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Ecuación 3.1. Ecuación de la media

Mediana

La mediana representa el valor de la variable que se ubica en el centro del conjunto de datos ordenados. Dado un número impar de observaciones, la mediana es el valor más medio de todo el conjunto. En el caso de que el número de observaciones sea par, la mediana se obtiene tomando el promedio del medio [26],[33].

Moda

La moda para un conjunto de datos es el valor que ocurre con mayor frecuencia. Es posible que la mayor frecuencia corresponda a varios valores diferentes, lo que da como resultado más de una moda. Los conjuntos de datos con uno, dos o tres modas se denominan, respectivamente, unimodal, bimodal y trimodal. En general, un conjunto de datos con dos o más modas es multimodal. En el otro extremo, si cada valor de datos ocurre solo una vez, entonces no hay moda [26], [33].

3.1.1.2. Medidas de dispersión de datos

Varianza y desviación estándar

Estas dos medidas están directamente relacionadas entre sí, ya que la desviación estándar es la raíz cuadrada de la varianza. La varianza muestra la distribución o dispersión de datos alrededor de la media y la desviación estándar representa una cantidad razonable para que un punto de datos específico difiera de la media, es decir, expresa cuánto es probable que varíe de su valor medio [34]. Por tanto, una desviación estándar baja significa que las observaciones de datos tienden a ser muy cercanas a la media, mientras que una desviación estándar alta indica que los datos se distribuyen en un amplio rango de valores [26].

La varianza de N observaciones, x_1, x_2,..., x_N , para un atributo numérico X es la que se muestra en la Ecuación 3.2.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \overline{x}^2$$

Ecuación 3.2. Ecuación de la varianza

Donde $\overline{x}$ es el valor de la media de las observaciones definida en la Ecuación 3.1. La desviación estándar, σ, de las observaciones es la raíz cuadrada de la varianza, σ^2.

Rango

El rango es la diferencia entre el máximo valor y el mínimo valor de una distribución de datos [33].

Cuartiles y rango intercuartil

Los cuantiles son medidas que dividen una distribución de datos, en conjuntos consecutivos esencialmente iguales. El 2-cuantile es el punto que divide las mitades inferior y superior de la distribución de datos, es decir, corresponde a la mediana. Los 4-cuantiles son los tres puntos de datos que dividen la distribución de datos en cuatro partes iguales; cada parte representa un cuarto de la distribución de datos, se les conoce más comúnmente como cuartiles. Los 100-cuantiles se conocen como percentiles; dividen la distribución de datos en 100 conjuntos consecutivos de igual tamaño. La mediana, los cuartiles y los percentiles son las formas más utilizadas de cuantiles [26], [32].

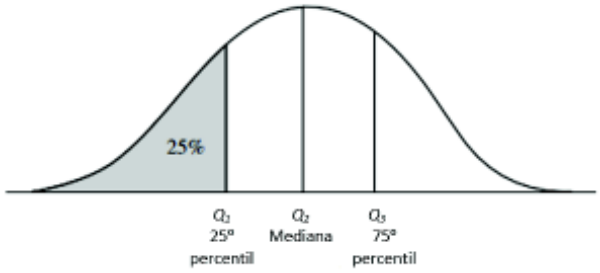


Figura 3.1. Gráfico de la distribución de datos

En la Figura 3.1. los cuartiles dan una indicación del centro, la extensión y la forma de una distribución. El primer cuartil, indicado por Q_1, es el percentil 25º. Corta el 25% más bajo de los datos. El tercer cuartil, indicado por Q_3, es el percentil 75º, corta el 75% más bajo (o el 25% más alto) de los datos. El segundo cuartil es el percentil 50º. Como la mediana, da el centro de la distribución de datos [35].

La distancia entre el primer y el tercer cuartil es una medida simple de dispersión que proporciona el rango cubierto por la distribución central de los datos. Esta distancia se denomina rango intercuartílico (IQR) y se define como se muestra en la Ecuación 3.3.

$$IQR = Q_3 - Q_1$$

Ecuación 3.3. Ecuación del rango intercuartílico

Máximo valor

Muestra el valor más alto en un conjunto de datos [35].

Mínimo valor

Muestra el valor más bajo en un conjunto de datos [35].

3.1.2. Gráficas de las descripciones estadísticas básicas de datos

Las técnicas visuales/gráficas permiten las descripciones estadísticas básicas del conjunto de datos. Estos incluyen histogramas, gráficos de dispersión, BoxPlot y gráfico de cuantil. Dichos gráficos son necesarios para la inspección visual de datos debido a que muestran la distribución de estos, por tanto, es útil para el preprocesamiento de datos [26].

Histograma

Trazar histogramas es un método gráfico para resumir la distribución de los datos dados, el gráfico resultante se conoce más comúnmente como gráfico de barras. El rango de valores se divide en subintervalos consecutivos disjuntos. Los subintervalos, denominados cubos o contenedores, son subconjuntos de la distribución de datos.

El histograma muestra cómo los datos se dispersan en el rango de los datos (los números entre el valor más pequeño y el más grande) y si los datos son simétricos (la distribución es similar en el lado derecho e izquierdo) o sesgados (las categorías parecen disminuir o aumentar en un lado de la distribución, haciendo que el histograma sea asimétrico). La Figura 3.2. muestra un ejemplo simple de cómo los histogramas pueden ilustrar simetría y sesgo [36].

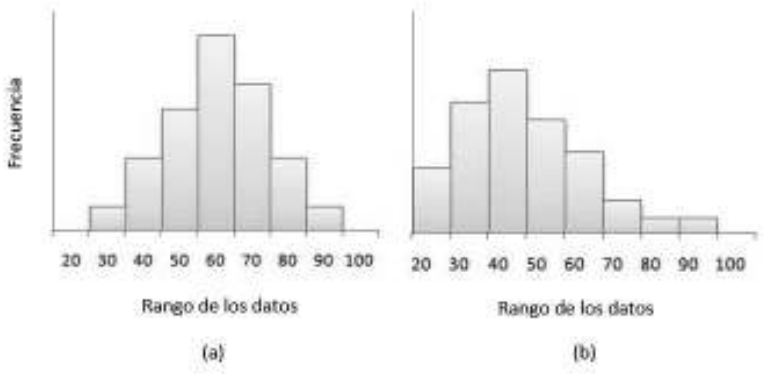


Figura 3.2. Histograma de un conjunto simétrico de datos (a) y un conjunto de datos sesgados a la derecha (b)

Gráfico de Dispersión

Un diagrama de dispersión es un gráfico bidimensional, donde cada observación del conjunto de datos tiene dos componentes, tales datos se llaman bivariados. Por lo general, se designa a cada observación un par de componentes (X, Y) y se trazan como puntos en el plano, un diagrama de dispersión mapea el valor de X unidades en el eje horizontal, y Y unidades en el eje vertical.

Este tipo de diagrama es un método útil para proporcionar una primera observación de los datos bivariados, y de esta forma ver grupos de puntos y valores atípicos, o para explorar la posibilidad de relaciones de correlación [36]. En la Figura 3.3. se puede observar un ejemplo de diagrama de dispersión para un conjunto de datos bivariados correspondientes a las variables entre el precio unitario y casas vendidos.

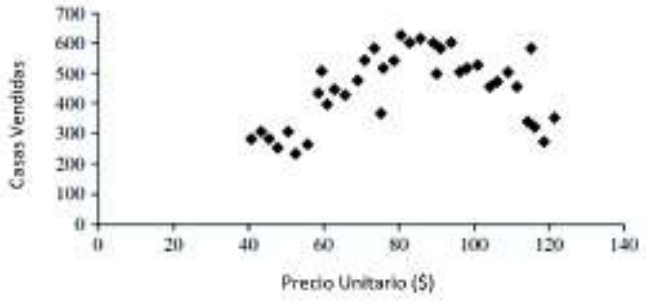


Figura 3.3. Ejemplo de diagrama de dispersión

Diagrama de caja (BoxPlot)

También llamada diagrama de caja y bigotes (BoxPlot) es una pantalla visual simple y conveniente para ilustrar la distribución de los datos a través de percentiles. Los tres cuartiles (percentiles Q_1, Q_2, Q_3) están destinados a dividir los datos en cuatro partes del mismo tamaño, con un cuarto del conjunto de datos en cada uno. El diagrama de caja muestra cinco características importantes de los datos, incluidos tres valores percentiles. La Figura 3.4. muestra la estructura del diagrama de caja, una caja segmentada con líneas que sobresalen de la parte inferior y superior [36]. Dicho diagrama de caja incorpora el resumen de cinco números de la siguiente manera:

- Por lo general, los extremos de la caja están en los cuartiles (Q_1, Q_3), de modo que la longitud de la caja es el rango intercuartílico (IQR) mencionada anteriormente en la Ecuación 3.3.
- La mediana está marcada por una línea dentro del cuadro (Q_2).
- Dos líneas (llamadas bigotes) fuera de la caja se extienden a las observaciones más pequeñas (mínimo) y más grandes (máximo).

Una aplicación común del diagrama de caja es identificar los valores atípicos, mediante la individualización de los valores que caen al menos $1{,}5 * IQR$ por encima del tercer cuartil o por debajo del primer cuartil [26].

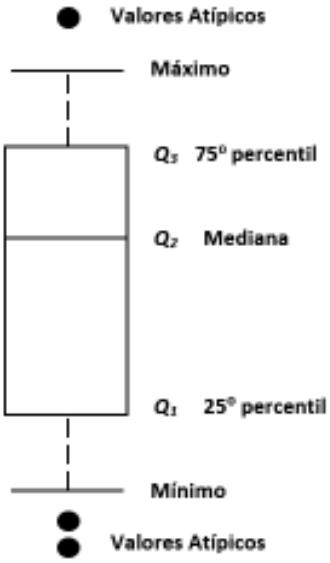


Figura 3.4. Estructura del diagrama de caja

Gráfico Cuantil

Un gráfico de cuantiles es una manera simple y efectiva de observar por primera vez una distribución de datos univariada. En primer lugar, muestra todos los datos para el atributo dado (lo que permite al usuario evaluar tanto el comportamiento general como las ocurrencias inusuales). Traza información de los cuantiles del conjunto de datos. Se tiene en cuenta que el percentil 0,25 corresponde al cuartil Q_1, el percentil 0,50 es la mediana y el percentil 0,75 es Q_3. En una gráfica de cuantiles se permite comparar diferentes distribuciones basadas en sus cuantiles [26]. Por ejemplo, dados los gráficos de cuantiles de los datos de ventas para tiempo diferentes, se compara el Q_1, mediana, Q_3, tal como se muestra en la Figura 3.5.

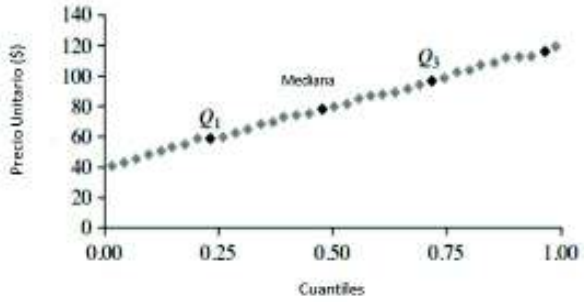


Figura 3.5. Ejemplo del gráfico cuantil

3.2. Metodologías para la detección e imputación de los datos NaN

Este tipo de métodos analizan los datos faltantes de una serie temporal en un caso particular, es muy importante llenar esta variable mediante una estimación razonable de un valor de datos adecuado, es de suma importancia para el posterior análisis de dicha serie temporal. La operación de decidir qué datos usar para llenar estos espacios en blanco se denomina imputación de datos. Este término significa que se asigna datos al espacio en blanco en función de alguna heurística razonable. La selección de la técnica adecuada para manejar los valores perdidos depende de hacer la suposición correcta sobre el patrón de "faltantes" en el conjunto de datos [34].

A continuación, se describe brevemente algunos de los algoritmos implementados.

3.2.1. Interpolación lineal

La interpolación es un proceso de definición de una función que toma valores en puntos específicos y es comúnmente utilizada para los problemas de datos faltantes, ya que crea nuevos valores de los datos dentro del rango de observaciones individuales [37]. La forma más simple de interpolación es la interpolación lineal, donde la aproximación es una variación lineal entre los valores del conjunto de datos. En el caso de una serie temporal para valores perdidos de datos contiguos, la interpolación lineal estima y rellena los valores faltantes dibujando segmentos sólidos de línea recta entre los puntos de datos adyacentes o vecinos, y así calcula el valor perdido a lo largo de dicha línea [38], [39], [40].

Dado n puntos en el plano, (x_i, y_i), $i = 1, \dots, n$, con distintas x_i, existen varias fórmulas diferentes para el polinomio en x de grado menor que n cuyo gráfico pasa por los puntos. Este polinomio se llama polinomio de interpolación debido a que reproduce los datos dados:

$$S(x) = y_i, \quad i = 1, \dots, n.$$

Suponiendo que se tiene los puntos x_i y x_{i+1}, con los valores y_i y y_{i+1} de un conjunto de datos, respectivamente y se considera que $x_i \leq x < x_{i+1}$. Falta el valor del punto x, donde x está en algún lugar entre x_i y x_{i+1}. Para estimar y se emplea la definición de interpolación lineal tal como se muestra en el Ecuación 3.4.

$$S_i(x) = y_i + \frac{(x - x_i)(y_{i+1} - y_i)}{x_{i+1} - x_i} \qquad i = 1, 2, \dots, n-1$$

Ecuación 3.4. Ecuación de la interpolación lineal

3.2.2. Interpolación cúbica “Spline”

El modelo de interpolación cúbica “Spline” se utiliza para estimar y completar los valores faltantes en un conjunto de datos de series temporales, este método ajusta un polinomio cúbico para interpolar entre dos puntos. Los coeficientes del polinomio se determinan usando los valores de puntos que son adyacentes a los dos datos faltantes, para estimar la observación faltante por el valor de la spline. La interpolación cúbica spline tiene alta convergencia, alta estabilidad y crea una función de interpolación suave [41], [42].

Dado un conjunto de puntos (x_i, y_i) donde $i = 0, 1, \dots, n$, esta interpolación consiste en hacer pasar por cada dos puntos un polinomio de tercer grado $S(x)$, precisando condiciones de continuidad de la primera y la segunda derivadas, y a su vez condición de suavidad de la segunda derivada. Sean (x_i, y_i) y (x_{i+1}, y_{i+1}) dos puntos consecutivos; se construye un polinomio de tercer grado en el i-ésimo intervalo como se indica en la Ecuación 3.5.

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \qquad i = 1, 2, \dots, n-1$$

Ecuación 3.5. Ecuación del polinomio de tercer grado

Considerando que la función se encuentra dentro del intervalo:

$$x_i \leq x < x_{i+1}$$

Donde h denota la longitud del subintervalo:

$$h = x_{i+1} - x_i$$

Se hace uso de estas ecuaciones y las propiedades de splines cúbicos indicadas a continuación para expresar los coeficientes desconocidos del polinomio de tercer grado $S(x)$, en términos de valores conocidos. Una vez que se han calculado las pendientes, el interpolador se puede evaluar eficientemente para reproducir dichos datos [43], [44].

- Interpolación en los puntos.

$$S_i(x) = y_i \quad , \quad S_i(x_{i+1}) = y_{i+1} \qquad i = 1, 2, \dots, n-1$$

- Continuidad de la primera derivada [92].

$$S_i'(x_i) = S_{i-1}'(x_i) \qquad i = 1, 2, \dots, n-1$$

- Continuidad y suavidad de la segunda derivada [92].

$$S_i''(x_i) = S_{i-1}''(x_i) \qquad i = 1, 2, \dots, n-1$$

Entonces los coeficientes del polinomio cúbico se expresan como se indica en la Ecuación 3.6.

$$a_i = \frac{(S''_{i+1}(x_{i+1}) - S''_i(x_i))}{6h}$$

$$b_i = \frac{S''_i(x_i)}{2}$$

$$c_i = \frac{y_{i+1} - y_i}{h} - \frac{(S''_{i+1}(x_{i+1}) + 2S''_i(x_i))\, h}{6}$$

$$d_i = y_i$$

Ecuación 3.6. Coeficientes del polinomio de tercer grado

3.2.3. Interpolación cúbica "Akima"

La interpolación Akima es una variante de la interpolación cúbica spline, este método se basa en un ajuste local para definir los coeficientes del polinomio de tercer grado, y se determina mediante los puntos contiguos al intervalo de interpolación. La curva se divide en segmentos y cada segmento se ve influenciado por un conjunto de puntos de datos vecinos más cercanos [45].

Cada segmento del punto x_i al punto x_{i+1} es un polinomio cúbico, y la idea principal es calcular la pendiente del segmento en el punto x_i explícitamente, por medio de una expresión que depende del punto, sus dos predecesores inmediatos y sus dos sucesores inmediatos. El método es por lo tanto local, y mover un punto afecta a los demás segmentos cercanos a la curva (útil cuando algunos puntos de datos son valores atípicos). Los puntos están conectados con segmentos rectos (secantes) cuyas pendientes se denotan con m, donde $m = (y - y_1)/(x - x_1)$. La principal característica del algoritmo Akima es la expresión para calcular la pendiente en el punto x_i, la cual se muestra en la Ecuación 3.7., [46].

$$t_i = \frac{|m_{i+1} - m_i| m_{i-1} + |m_{i-1} - m_{i-2}| m_i}{|m_{i+1} - m_i| + |m_{i-1} - m_{i-2}|}$$

Ecuación 3.7. Ecuación de la pendiente mediante Akima

Por lo que se hace uso de la Ecuación 2.7., para aplicar a los puntos (x_i, y_i) y (x_{i+1}, y_{i+1}) y de esta manera calcular las pendientes t_i y t_{i+1}. Estos seis términos (cuatro coordenadas y dos pendientes) se utilizan para calcular el polinomio cúbico del segmento x_i a x_{i+1}. El polinomio en sí tiene la forma de la Ecuación 2.5., donde $h = x_{i+1} - x_i$ es la longitud del

subintervalo, por lo cual los coeficientes del polinomio de tercer grado se encuentran en función de los 6 términos mencionados como se muestra en la Ecuación 3.8, [46].

$$a_i = \frac{t_i + t_{i+1} - 2(y_{i+1} - y_i)/h}{h^2}$$

$$b_i = \frac{3(y_{i+1} - y_i)/h - 2t_i - t_{i+1}}{h}$$

$$c_i = t_i$$

$$d_i = y_i$$

Ecuación 3.8. Coeficientes del polinomio de tercer grado

3.2.4. Modelo Autorregresivo

En un modelo autorregresivo AR, un valor en el tiempo t se basa en una combinación lineal de valores previos al de la variable, el término autorregresivo indica que es una regresión de la variable contra sí misma. El modelo AR es flexible para manejar una amplia gama de patrones en series de tiempo diferentes [47]. La forma básica del proceso autorregresivo se define en la Ecuación 3.9.

$$x_t = \sum_{k=1}^{p} a_k x_{t-k} + \epsilon_t$$

Ecuación 3.9. Ecuación del modelo autorregresivo AR

Donde $a_1, a_2, \ldots, a_p$ son los coeficientes autorregresivos del modelo y ϵ_t es una secuencia de variables aleatorias independientes con media equivalente a 0 y varianza σ^2 (ruido blanco). Por lo que el modelo AR establece que el valor de x al instante t es una combinación lineal de p valores previos de x más un elemento que es ruido blanco. Los coeficientes AR o coeficientes de autorregresión se pueden calcular de varias formas, uno de los métodos más utilizados es el algoritmo de Yule-Walker, el cual calcula dichos coeficientes a partir de coeficientes de autocorrelación parcial o también llamados "coeficientes de reflexión", este método relaciona los parámetros AR con la función de autocorrelación r_k conocida (o estimada) de x_t, la cual define los coeficientes de autocorrelación en la etapa n, como se muestra en la Ecuación 3.10. [48], .

$$r_k = \frac{1}{N}\sum_{n=1}^{N-k}[x(n+k) - \bar{x}][x(n) - \bar{x}]; \qquad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x(n) \qquad N > k$$

Ecuación 3.10. Función de autocorrelación

Donde k es el orden del modelo, N tamaño del conjunto de datos. Si el orden del modelo es demasiado bajo, solo se captura una parte de la señal. Por otro lado, si el orden de un modelo es demasiado alto, se captura la señal completa, pero también se modela la medida del ruido. Para determinar el orden óptimo del modelo AR se selecciona el orden que minimiza el criterio de Akaike "AIC", el cual se define en la Ecuación 3.11. Donde L es el valor máximo del modelo [49], [50].

$$AIC = 2k - 2\ln(L)$$

Ecuación 3.11. Función de autocorrelación

Por consiguiente, los coeficientes de reflexión son calculados basados en los coeficientes de autocorrelación como se indica en la Ecuación 3.12.

$$\begin{bmatrix} r_0 & r_1 & r_2 & \dots & r_{p-1} \\ r_1 & r_0 & r_1 & \dots & r_{p-2} \\ r_2 & r_1 & r_0 & \dots & r_{p-3} \\ \dots & \dots & \dots & r_0 & \dots \\ r_{p-1} & r_{p-2} & r_{p-3} & \dots & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ \vdots \\ r_p \end{bmatrix}$$

Ecuación 3.12. Método Yule-Walker

Por lo tanto, después de estimar el coeficiente de autocorrelación r_n mediante la Ecuación 3.10. y el coeficiente de autorregresión a_n, mediante la Ecuación 3.12. y aplicarlos a la Ecuación 3.9., se puede estimar x_t.

3.2.5. Media Móvil

La media móvil es un método que se basa en el cálculo de la media de un número n de observaciones, en un intervalo donde se mueve su punto de origen a través de una ventana deslizante de longitud k. La media se calcula y registra para este número de observaciones desde el principio hasta el final de la serie, cada una de las observaciones en el cálculo del promedio móvil recibe un peso igual cuando se calcula el promedio simple, como se indica en la Ecuación 3.13.

$$\mu = \frac{1}{k}\sum_{i=1}^{k} x_i \qquad k = k_b + k_f + 1$$

Ecuación 3.13. Media móvil

Donde k es la longitud de la ventana, especificada como un vector de dos elementos enteros positivos [k_b k_f], la cual contiene el elemento en la posición actual del intervalo, k_b elementos hacia atrás y k_f elementos hacia delante. Los valores de series de tiempo se

promedian sobre cada uno de estos intervalos, cuando el tamaño de la ventana se detiene en los puntos finales debido a que no hay suficientes elementos para llenar la ventana, por lo cual el promedio se toma solo sobre los elementos que llenan la ventana [28], [35].

3.3. Metodologías para filtrar ruido en series temporales

En las señales temporales existe ruido, el cual se constituye por señales indeseables que se introducen a lo largo del trayecto de transmisión, de manera que altera la señal deseada. Este ruido es generado por causas internas y externas al sistema, y constituye uno de los principales factores que limitan un análisis íntegro de la señal. Uno de los puntos importantes para una adecuada limpieza de datos de la señal es eliminar el ruido, el cual es una señal aleatoria generalmente de alta frecuencia, pero con un ancho de banda muy amplio, que se suma a la señal original. Primero se procede a la visualización en el dominio de la frecuencia haciéndose uso de la Transformada de Fourier o la Transformada Wavelet para posteriormente filtrar el ruido de la señal [37], [51].

3.3.1. Transformada de Fourier

La transformada de Fourier (FT) se emplea para convertir señales del dominio del tiempo al dominio de la frecuencia y la inversa de la transformada convierte del dominio de la frecuencia al dominio del tiempo. Esta transformación es reversible y mantiene la misma energía. En el presente estudio se usa la transformada discreta de Fourier (DFT), la cual es un caso particular de la transformada de Fourier para secuencias de longitud finita, que permite estimar el espectro en frecuencias específicas (por consiguiente, se obtiene un espectro discreto de la señal). La DFT se calcula sobre el intervalo temporal $0 < n < N - 1$, donde los datos a transformar consisten en N puntos espaciados uniformemente de duración finita, ya sea con tendencia periódica o no periódica [52].

Dada una secuencia de N muestras de x_n, donde $n = 0,1, \dots, N-1$, entonces la transformada de Fourier discreta (DFT) se define como se indica en la Ecuación 3.14.

$$x_k = \int_{n=0}^{N-1} x_n \, e^{-j2\pi kn/N}, \qquad k = 0,1, \dots, N-1$$

Ecuación 3.14. Transformada Discreta de Fourier

Donde x_k por lo general se denominan como "coeficientes de Fourier" o "armónicos". La secuencia x_n puede ser calculada utilizando la inversa de la transformada discreta de Fourier (IDFT), como se observa en la Ecuación 3.15.

$$x_n = \frac{1}{N}\int_{n=0}^{N-1} x_k \, e^{j2\pi kn/N}, \qquad n = 0,1,\dots,N-1$$

Ecuación 3.15. Inversa de la Transformada Discreta de Fourier

El análisis de Fourier se emplea para visualizar, descomponer la señal en el dominio de la frecuencia mediante la obtención de los coeficientes de Fourier x_k, con los cuales se puede identificar el ruido en la señal y posteriormente aplicar los filtros que se exponen a continuación para proceder a eliminarlo.

3.3.1.1. Filtros

Un filtro es una operación matemática que toma una señal de entrada y la modifica produciendo otra señal con el objetivo de atenuar ciertas características de la señal. A continuación, se detallan los filtros Yulewalk, Butterworth y FIR.

Yulewalk

El filtro digital Yulewalk se basa en un ajuste de mínimos cuadrados en el dominio del tiempo. Para filtrar la señal hace uso de la función de transferencia que se expone en la Ecuación 3.16. y para calcular los respectivos coeficientes se utiliza las ecuaciones de Yule-Walker modificadas que se presentan en [53], junto con los coeficientes de correlación calculados por la transformada discreta de Fourier de la señal.

$$H(z) = \frac{B(z)}{A(z)} = \frac{b(1) + b(2)z^{-1} + \cdots + b(n+1)z^{-n}}{a(1) + a(2)z^{-1} + \cdots + a(n+1)z^{-n}}$$

Ecuación 3.16. Función de transferencia del filtro Yulewalk

Butterworth

El filtro analógico Butterworth presenta una respuesta de frecuencia lo más plana posible (con mínimas ondulaciones) en la banda de paso hasta la frecuencia de corte, un módulo monótono decreciente hasta cero en la banda de rechazo con una fase muy suave, lo cual es importante cuando se considera la distorsión. La función de transferencia del filtro Butterworth se define como se indica en la Ecuación 3.17. Donde N es el orden del filtro, ω_c es la frecuencia de corte [54].

$$|H(j\omega)|^2 = \frac{1}{1 + \left(\frac{\omega}{\omega_c}\right)^{2N}}$$

Ecuación 3.17. Función de transferencia del filtro Butterworth

FIR

El filtro digital de Respuesta Finita al Impulso (FIR) se caracteriza por tener una respuesta de impulso de duración finita. Los coeficientes del filtro FIR ya sean simétricos o asimétricos proporcionan una fase lineal; estos coeficientes son las respuestas al impulso. Una de las ventajas de los filtros FIR es que se puede diseñar con una fase exactamente lineal. Una respuesta con una fase lineal es muy importante, debido a que se reduce la distorsión en la fase [55]. Un filtro FIR se caracteriza por la función de transferencia que se presenta en la Ecuación 3.18. donde la secuencia $h(k)$ corresponde a los coeficientes del filtro y N es el orden del filtro.

$$H(z) = \sum_{k=0}^{N-1} h(k)\, z^{-k}$$

Ecuación 3.18. Función de transferencia del filtro FIR

3.3.2. Transformada de Wavelet

La transformada de Wavelet es una nueva técnica que permite obtener la descomposición local tiempo-frecuencia de la señal. Esta transformada compara la señal con versiones desplazadas y escaladas de una función llamada “wavelet madre”, la correlación entre estas es medida a través de coeficientes, los cuales proporcionan una descripción local de las características de la señal. A diferencia de la transformada de Fourier, esta realiza un análisis simultáneo en el dominio del tiempo y en el dominio de la frecuencia en señales que pueden incluir transitorios o señales variables en el tiempo [56]. La transformada wavelet para una función $f(t)$ está dada por la Ecuación 3.19.

$$f(t) = \sum_{k} \sum_{j} a_{j,k}\, \psi_{j,k}(t)$$

Ecuación 3.19. Transformada Wavelet

Donde j y k son índices de integración y el $\psi_{j,k}(t)$ es la función Wavelet madre, existiendo diferentes tipos de wavelets como, por ejemplo: Haar, Daubechies, Coiflets, Symlet y Biorthogonal, la elección de esta función depende de su aplicación. El conjunto de coeficientes $a_{j,k}$ son llamados como la transformada discreta de Wavelet de $f(t)$. La transformada discreta de Wavelet (DWT) descompone la señal en varias bandas de frecuencia en cada punto en el tiempo a lo largo de la señal. DWT no produce ninguna información redundante debido a que los coeficientes de la transformada no se duplican.

Una de las principales aplicaciones de la DWT es eliminar el ruido de una señal, es muy efectivo en comparación a otros análisis [57].

3.3.2.1. Filtrado de ruido mediante la DWT

Este método recurre a un algoritmo piramidal el cual hace uso de la transformada discreta de Wavelet. En la DWT, una señal (S) para analizarla pasa a través de un banco de filtros en n niveles de descomposición. En cada nivel, el banco de filtros consiste en un filtro pasa alto y pasa bajo, su estructura se indica en la Figura 3.6.

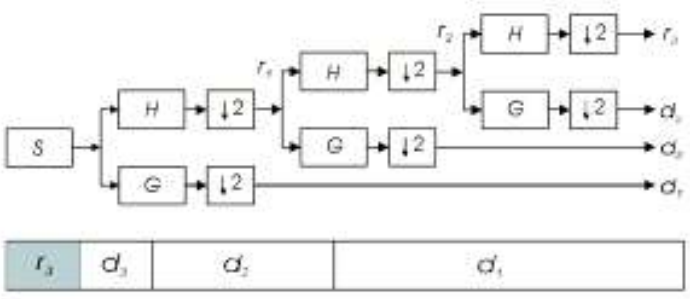


Figura 3.6. Esquema piramidal del banco de filtros

Primero se descompone la señal de entrada (S) a través de: un filtro pasa bajos (H) el cual determina los coeficientes de aproximación que contienen las componentes de baja frecuencia de la señal, un filtro pasa alto (G) el cual determina los coeficientes que contienen las componentes de alta frecuencia de la señal, y a su vez se modifica la razón de muestreo de ambas salidas por dos. En el siguiente nivel el proceso es iterado en la rama pasa bajos (r). Teóricamente, el nivel de descomposición puede ser infinito, pero se ha determinado que el cuarto nivel de descomposición es el más indicado para eliminar el ruido de una señal. Una vez que se obtiene el conjunto de coeficientes wavelet los cuales se correlacionan con las diferentes bandas de frecuencia, se los compara con un pequeño valor de umbral de wavelet, aquellos que tienen un valor por debajo del umbral se establece como cero, caso contrario se mantiene, al establecer estos coeficientes en cero, esencialmente eliminan el ruido de una señal [58].

3.4. Metodologías para la detección e imputación de los datos atípicos

Las series temporales puede contener datos que no cumplan con el comportamiento general o modelo del conjunto de datos, estos valores son llamados atípicos. Los valores atípicos podrían causar problemas críticos debido a que una pequeña cantidad de este tipo

de datos amenaza con perjudicar gravemente el contenido intrínseco de los datos al cambiar dramáticamente la tendencia de las estimaciones.

Se han implementado algunos métodos de suavizado que permiten detectar e imputar la influencia de los valores atípicos y aproximar los datos de la señal a una función estimada por patrones importantes de la serie temporal, modificando los valores atípicos y ajustando a una curva suave y fina en relación con la señal original. Otro de los métodos de detección son los basados en la agrupación, los cuales establecen que los objetos de datos normales pertenecen a agrupaciones grandes y densas, mientras que los valores atípicos pertenecen a agrupaciones pequeñas o dispersas, o no pertenecen a ninguna agrupación [26], [59].

A continuación, se describe brevemente algunas de las técnicas de suavizado y de agrupamiento que detectan e imputan los valores atípicos en una serie temporal.

3.4.1. Regresión lineal local "LOWESS"

Es un método de regresión no paramétrico, el cual se basa en ajustar polinomios lineales locales en cada punto, para así obtener los valores de ajuste de la serie temporal. La regresión lineal local LOWESS asume que la variable explicativa x y la variable de respuesta y están relacionadas mediante la Ecuación 3.20.

$$y_i = g(x_i) + \varepsilon_i$$

Ecuación 3.20. Ecuación general de la Regresión Local

donde g es un polinomio lineal en función de la variable x_i, y ε_i una variable aleatoria con media de cero, que indica la variación de y alrededor de g. Los valores ε_i se utilizan para estimar el ajuste y_i en cada x_i, y se calculan ajustando los polinomios lineales, mediante el uso del algoritmo de mínimos cuadrados ponderados. Este algoritmo determina residuos y asigna pesos $W(x_i)$ a cada punto de datos del vecindario, los puntos de datos cercanos a x_i tienen más peso que los que están más lejos. Los vecindarios locales están además determinados por la longitud de la ventana. Luego se realiza nuevamente un ajuste polinomial local, pero ahora se asigna a cada observación un nuevo peso que es el producto del peso en el ajuste inicial y el peso asignado a su residual a partir de ese ajuste inicial. Por lo tanto, las observaciones que muestran grandes residuos en el ajuste inicial se reducen en el segundo ajuste. El proceso anterior se repite varias veces dando como resultado el estimador LOWESS, el cual define de manera iterativa las ponderaciones de robustez y los suaviza varias veces [60], [61], [62].

3.4.2. Regresión cuadrática local “LOESS”

Otro método de suavizado para imputar los valores atípicos de una serie temporal es la regresión cuadrática local LOESS. El marco de este método es similar a LOWESS, y se basa en ajustar polinomios cuadráticos locales en cada punto, para así obtener los valores de ajuste y suavizar la variable dependiente en función de la variable independiente, las cuales se relacionan mediante la Ecuación 3.20. En el método de LOESS de la misma forma que el método LOWESS, utiliza el algoritmo de los mínimos cuadrados ponderados para ajustar funciones cuadráticas de la variable x_i en los vecindarios locales, se obtienen los pesos de cada punto de datos del vecindario en una función suave. La curva de respuesta final es una combinación de estas curvas individuales. El número de vecindarios, especificado por las observaciones que se encuentran dentro de la longitud de la ventana, determina la suavidad de la curva [62], [63].

3.4.3. Savitzky-Golay “SGOLAY”

El algoritmo de Savitzky-Golay es un ajuste polinomial de mínimos cuadrados móviles. Este método realiza una regresión polinomial de orden superior en los datos especificados por la ventana móvil, los polinomios que se usan generalmente son los de segundo y cuarto grado, debido a que el proceso de aproximación se vuelve inestable con polinomios de mayor grado. El suavizado Savitzky-Golay permite conservar la altura y el ancho del pico de la señal, y lograr un alto nivel de suavizado sin atenuar las características de los datos. El punto de partida de este método es utilizar polinomios locales para describir el conjunto seleccionado de datos. Se escoge un punto central como el punto que es aproximado por el polinomio. Una serie de puntos contiguos a la izquierda y a la derecha se utilizan para ajustar el polinomio [62], [63]. La ecuación polinomial debe resolverse repetidamente para cada punto de datos en cuestión. La función Savitzky-Golay se observa en la Ecuación 3.21.

$$g_i = \sum_{n=-n_L}^{n_R} C_n f_i + n$$

Ecuación 3.21. Función Savitzky-Golay

Donde $f_i = f(t_i)$ cada valor de f_i es reemplazado por g_i, la cual es una combinación de sí misma y sus puntos vecinos, n_L es el número de puntos a la izquierda del punto x_i, mientras que n_R son los puntos a la derecha de x_i. La idea del método Savitzky-Golay es encontrar los coeficientes C_n, mediante el ajuste de mínimos cuadrados dentro de una ventana móvil,

lo que permite aproximar un polinomio de orden superior para cada punto f_i, y finalmente establecer la función g_i. Donde g_i es el valor del polinomio en la posición x_i.

3.4.4. K-Means

K-means es una herramienta muy popular del aprendizaje no supervisado. Este algoritmo de agrupación se basa en la distancia y proporciona información sobre los datos mediante el fraccionamiento del conjunto de datos en clústeres (agrupaciones), de modo que los datos de un clúster son más similares entre sí que los datos de otro clúster, por consiguiente, este análisis se torna sensible a valores atípicos [64].

El procedimiento k-means segmenta un conjunto de datos en k número de agrupaciones no superpuestas. Inicialmente, se definen los k centroides (un centroide es la media de los puntos en ese clúster), uno para cada agrupación. El siguiente paso es analizar cada punto dentro del conjunto de datos y agruparlo con el centroide más cercano, hasta que la distancia sea mínima entre estos dos puntos. El centroide se actualiza según los puntos asignados en cada agrupación, este proceso se repite hasta que la posición de los centroides se mantenga constante [65]. El algoritmo k-means intenta minimizar la función objetivo que se observa en la Ecuación 3.22.

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Ecuación 3.22. Función objetivo del algoritmo k-means

Donde E es la suma del error cuadrado para todos los puntos x_i y los centroides c_j. En otras palabras, para cada objeto en cada agrupación, la distancia desde el punto de datos hasta el centro de la agrupación es cuadrada. Esta función objetivo trata de hacer que las k agrupaciones resultantes sean tan compactas y separadas como sea posible. En este estudio para obtener el número óptimo de k agrupaciones del conjunto de datos se utilizó el criterio Calinski-Harabasz este método se basa en definir las agrupaciones mediante la maximización del índice VRC_k, el cual mide la varianza entre agrupaciones y la varianza dentro del agrupación, a través de la Ecuación 3.23.

$$VRC_k = \frac{B_k(n-k)}{W_k(k-1)}$$

Ecuación 3.23. Criterio Calinski-Harabasz

Donde VRC_k es el criterio de relación de varianza, B_k es la varianza entre agrupaciones, W_k es la varianza dentro de las agrupaciones. Los clústeres bien definidos tienen una gran varianza entre agrupaciones y una pequeña varianza dentro de la agrupación [66].

3.5. Criterios de evaluación de metodologías

Para comparar y medir el desempeño de los algoritmos presentados para la imputación de datos de las señales temporales se emplea dos indicadores: la raíz cuadrada del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2), los cuales miden la desviación entre la señal original y su versión reconstruida, de tal manera que esto aporte en la elección del algoritmo indicado para su aplicación.

3.5.1. Raíz cuadrada del error medio cuadrático "RMSE"

La raíz del error cuadrático medio (RMSE) es una medida de precisión que mide el grado de desviación entre los valores predichos por un modelo y los valores realmente observados de la serie temporal. Si el RMSE es equivalente a 0 significa que el modelo establecido predice las observaciones con una precisión exacta [32]. El valor RMSE se calcula mediante la Ecuación 3.24.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

Ecuación 3.24. Ecuación RMSE

Donde y_i son los valores de las observaciones del conjunto de datos, $\hat{y}_i$ son los valores que estima el modelo definido y n es el número total de observaciones.

3.5.2. Coeficiente de determinación "R^2"

El coeficiente de determinación (R^2) mide la capacidad predictiva del modelo ajustado, es decir, determina la bondad de ajuste. El valor que proporciona este coeficiente permite interpretar que tan bien se ajusta el modelo a los datos, de tal forma que si este tiene un valor de 1 el modelo estimado proporciona un ajuste perfecto de los datos, caso contrario se reduce la capacidad predictiva del modelo [67]. El coeficiente de determinación es adimensional y se calcula mediante la Ecuación 3.25.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Ecuación 3.25. Ecuación de R^2

Donde y_i son los valores del conjunto de datos, $\hat{y}_i$ son los valores de respuesta del modelo ajustado, $\bar{y}$ es la media del conjunto de datos y n es el número total de observaciones.

3.6. Diagrama de bloques de la aplicación desarrollada para la limpieza de datos de las mediciones sincrofasoriales

La aplicación ha sido desarrollada con la finalidad de ofrecer al usuario facilidades al momento de interactuar con la ejecución del programa, de tal forma que mediante la interfaz gráfica creada le sea posible manejar cada una de las etapas en las que se efectuará la limpieza de datos. En la primera etapa “Selección de la señal”, brinda la opción de escoger de una manera práctica una de las diferentes señales temporales registradas por las PMUs instaladas en el SNI. La etapa “Estadística Descriptiva” permite observar el comportamiento de los datos en términos de variables aleatorias y sus respectivas gráficas de descripciones estadísticas básicas del conjunto de datos. “Tratamiento de datos NaN” consiste en llenar los datos faltantes en función de algunos de los métodos específicos propuestos para esta etapa. “Filtrado de la señal” elimina el ruido. “Tratamiento de datos atípicos” esta etapa consiste en detectar e imputar la influencia de los valores atípicos. “Resultados de la señal preprocesada” permite obtener la señal temporal preprocesada y el cálculo de errores de la combinación de los métodos escogidos. En la Figura 3.7. se puede observar la secuencia de las etapas a seguir para obtener la señal temporal preprocesada mediante la aplicación desarrollada.

Figura 3.7. Bosquejo del diagrama de bloques del funcionamiento general de la aplicación

3.7. Descripción de App Designer de MATLAB

App Designer integra dos tareas principales: diseñar los componentes visuales de la interfaz gráfica de usuario (GUI - Graphical User Interface) y programar el comportamiento de la aplicación, las cuales están estrechamente vinculadas por lo que si se realiza cambios en el diseño afecta inmediatamente a la programación de la aplicación. Incluye una versión totalmente integrada del editor MATLAB, fue introducido en la versión R2016a [68]. En la Figura 3.8. y la Figura 3.9. se puede observar el entorno y los componentes de la App Designer, tanto en la vista del diseño como en la vista del código.

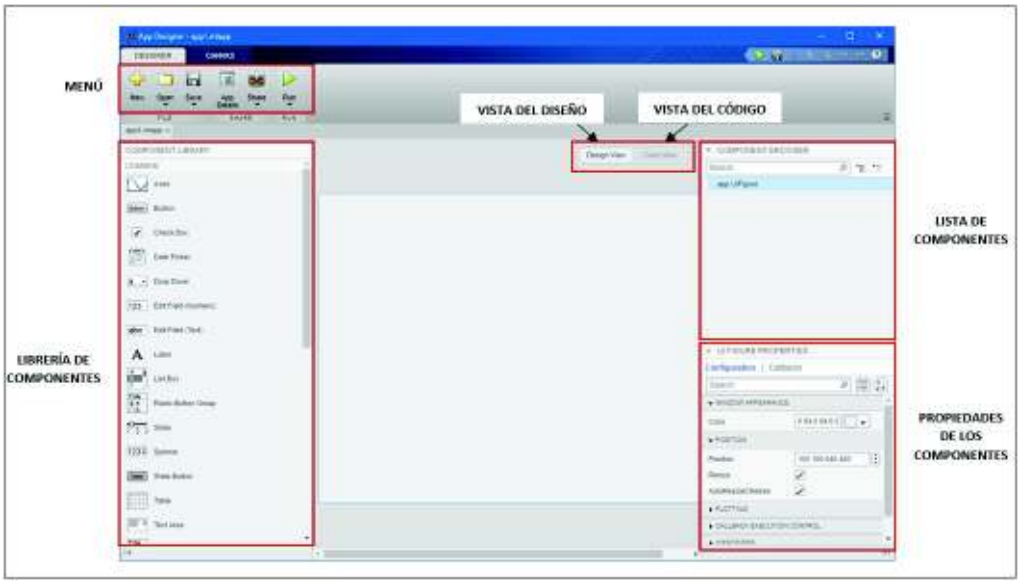


Figura 3.8. Vista del diseño de la App Designer

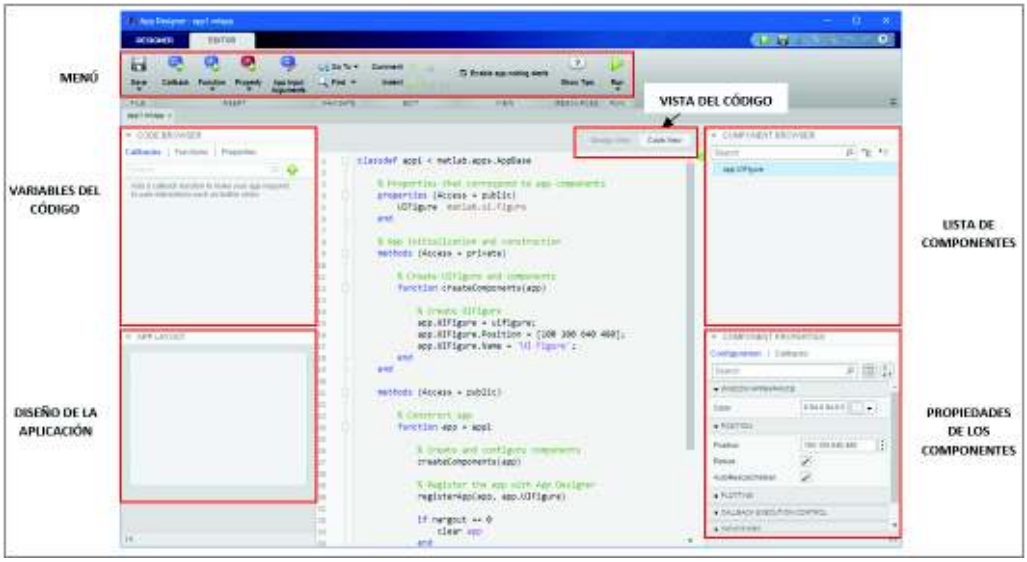


Figura 3.9. Vista del código de la App Designer

A continuación, en la Tabla 3.1. se puede observar la librería de componentes de la App Designer, la cual incluye un gran conjunto de componentes interactivos para diseñar aplicaciones modernas. Esta simplifica el proceso de diseño de las componentes visuales de una interfaz gráfica.

Tabla 3.1. Librería de componentes de App Designer

COMPONENTES	DESCRIPCIÓN
Componentes comunes	Incluye axes para crear gráficos y varios componentes que responden a las interacciones, como botones, controles deslizantes, listas desplegables.
Contenedores y herramientas de figura	Incluye paneles y pestañas para agrupar componentes, así como barras de menú.
Instrumentación	Incluye indicadores para visualizar el estado, así como mandos e interruptores para seleccionar los parámetros de entrada.

3.8. Algoritmo de cálculo

El programa desarrollado para el preprocesamiento de datos de las mediciones sincrofasoriales implementa un algoritmo de cálculo que propone la técnica de limpieza de datos mediante el uso del lenguaje de programación M.

El algoritmo general de este programa consiste en la extracción y la lectura de la base de datos desde un formato ".xlsx" de Excel a un archivo ".mat" de MATLAB, las diferentes señales temporales de frecuencia, derivada de frecuencia, voltaje, corriente y ángulos de fase de cada una de las PMUs se seleccionan e inicializan en variables y_i, en seguida se procede a analizar la variable mediante el cálculo de sus medidas de tendencia central y de dispersión, se reconoce la existencia de los datos NaN del conjunto de datos y_i para ser eliminados y reemplazados, una vez que la señal es continua se calcula el espectro de frecuencia y los coeficientes para ser comparados con el umbral y de esta manera ser filtrada, se detecta los datos atípicos y se aproxima los datos de la señal a una función estimada, finalmente se calcula los errores de todo el proceso entre la señal preprocesada y los valores realmente observados de la serie temporal, obteniendo finalmente los resultados.

A continuación, en la Figura 3.10. se presenta el diagrama de flujo correspondiente al programa implementado para el presente proyecto.

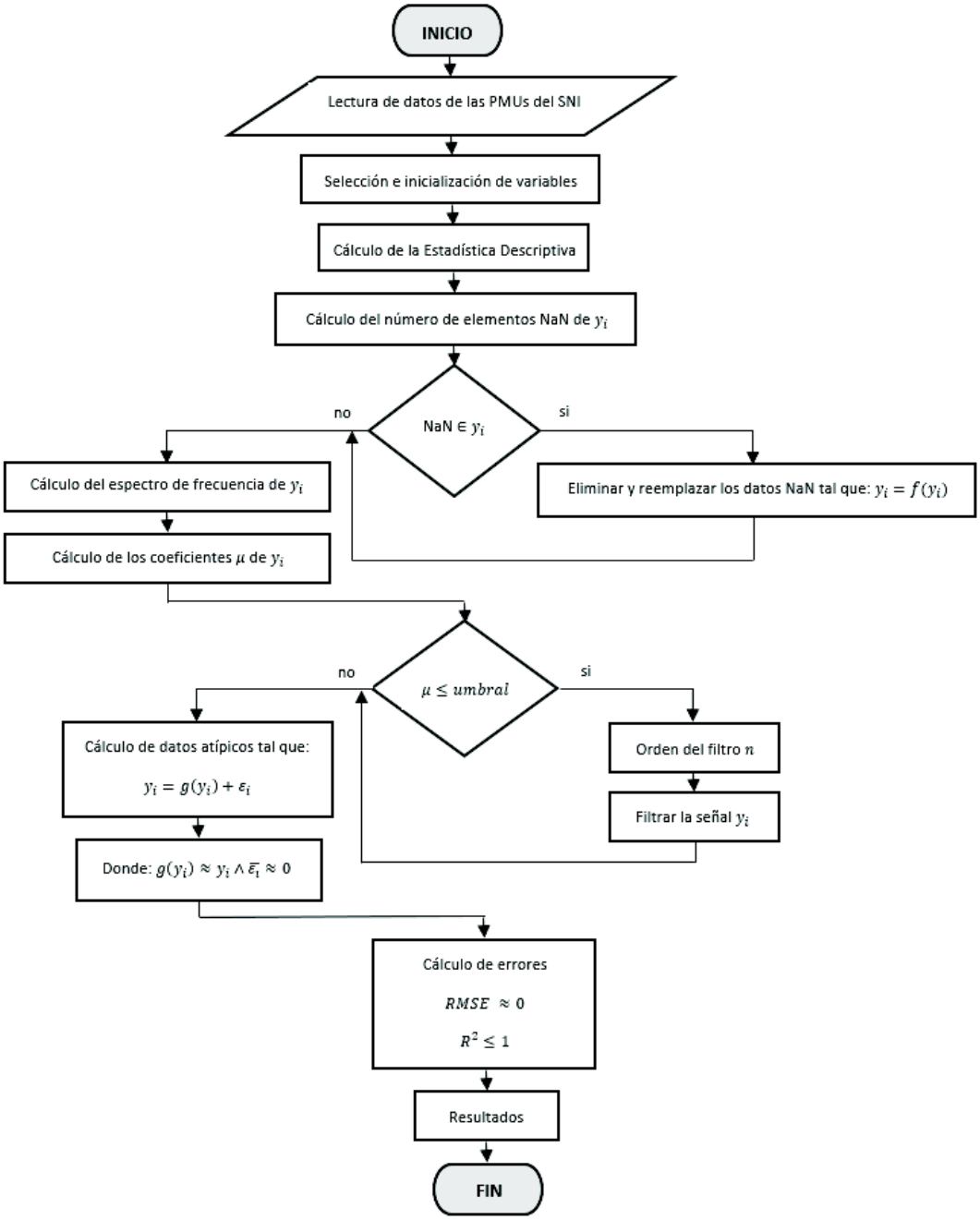


Figura 3.10. Diagrama de flujo del algoritmo de cálculo

3.9. Base de Datos

Una base de datos consiste en una colección de datos interrelacionados. La base de datos empleada en el presente proyecto constituye una parte fundamental para el desarrollo de la aplicación. El esquema físico de la base de datos se basa en un modelo relacional, es decir, consiste en una colección de tablas, a cada una de las cuales se le asigna un nombre único.

Cada tabla consta de un conjunto de atributos (columnas o campos) y generalmente almacena un gran conjunto de tuplas (registros o filas). Este esquema de distribución de los datos facilita el manejo y la depuración de la base de datos.

La base de datos proporcionada contiene información de 29 PMUs instaladas en el SNI según se indicó en la Tabla 2.2., cada una de las PMUs obtienen las diferentes mediciones sincrofasoriales de frecuencia, tasa de frecuencia, fasores de secuencia positiva del voltaje y corriente de su respectiva ubicación en la red eléctrica. Por lo que se obtiene una tabla característica correspondiente a cada PMU, la cual consta de 6 columnas y 47942 filas, y contienen los datos de las diferentes mediciones sincrofasoriales a una tasa de muestreo de 60 muestras por segundo en un intervalo de tiempo de 800 [s]. A continuación, en la Figura 3.11. se presenta un ejemplo de la PMU AGOYÁN-BAÑOS 1 con su correspondiente tabla.

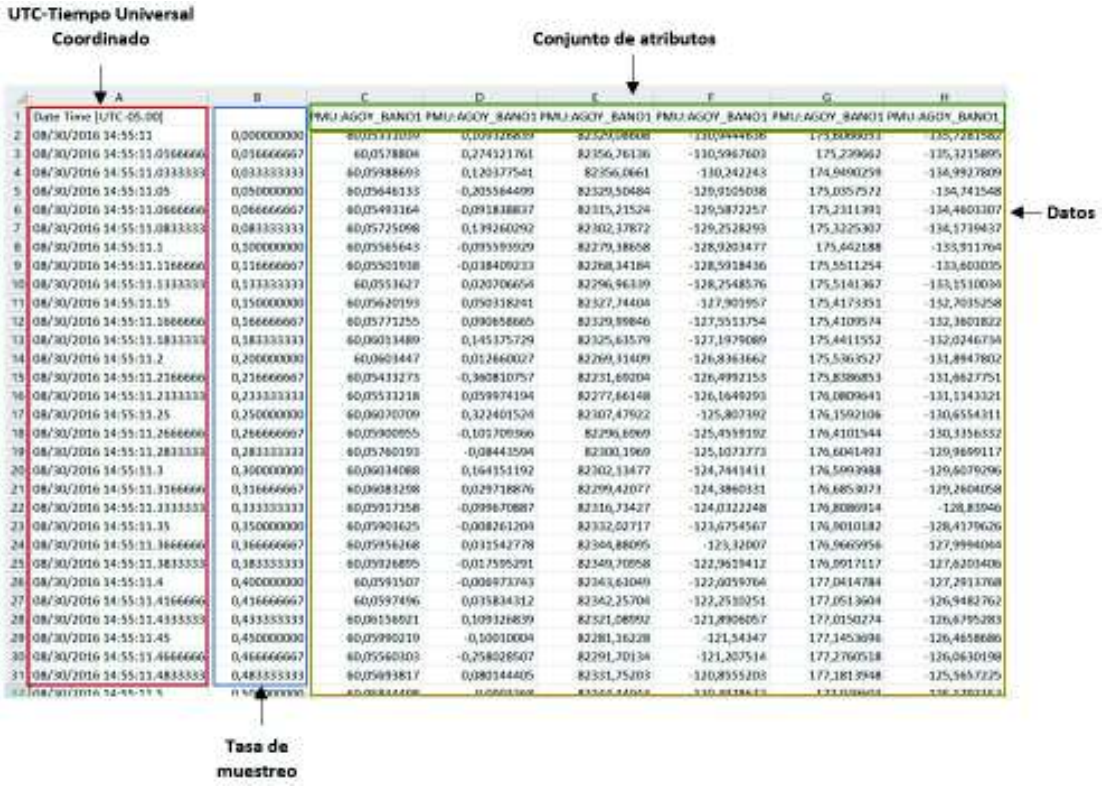


Figura 3.11. Tabla de datos de la PMU AGOYÁN-BAÑOS 1

La base de datos junto con las subrutinas desarrolladas en MATLAB, recolectan, depuran y procesan los datos de entrada de acuerdo con los parámetros establecidos en la aplicación.

3.10. Interfaz gráfica para la limpieza de datos de las mediciones sincrofasoriales

La estructura de la interfaz gráfica se expone en la Figura 3.12., como se puede observar está compuesta por seis secciones, las cuales permiten obtener como resultado final la limpieza de las diferentes mediciones sincrofasoriales por cada una de las PMUs instaladas en el SNI.

Figura 3.12. Interfaz gráfica para la limpieza de mediciones sincrofasoriales

A continuación, se describe el desarrollo general del funcionamiento de cada una de las etapas: selección de señal, estadística descriptiva, tratamiento de datos NaN, filtrado de la señal, tratamiento de datos atípicos, resultado de la señal preprocesada. Con la finalidad de comprender la lógica del proceso establecido en cada sección incluyendo sus respectivas restricciones.

3.10.1. Sección 1: SELECCIÓN DE SEÑAL

Esta sección brinda al usuario la opción de seleccionar la señal a tratar, es la primera pestaña de la aplicación, la cual consiste en un esquema compuesto por un desplegable que indica cada una de las diferentes PMUs del SNI mencionadas anteriormente y a su vez muestra una tabla deslizante que contiene la opción de seleccionar una de las mediciones sincrofasoriales ya sea la frecuencia, derivada de frecuencia, voltaje, corriente y ángulos de fase correspondiente a la PMU seleccionada, como se puede observar en la Figura 3.13. y la Figura 3.14.

Figura 3.13. Sección 1: Selección de Señal

Figura 3.14. Sección 1: Selección de Señal

Al accionar el botón “CARGAR”, se procede con la rutina de limpiar todas las figuras que se hayan encontrado en los ejes correspondientemente, también se lee el desplegable y la caja deslizante para determinar la PMU y la señal temporal que fueron seleccionadas, posteriormente llama a la función “seleccionar” creada para guardar los datos en las respectivas variables globales, finalizando con la gráfica de la señal en el espacio

correspondiente. En la Figura 3.15. se puede observar la gráfica de la señal de frecuencia de la PMU POMASQUI-JAMONDINO 2

Figura 3.15. Sección 1: Selección de Señal

Previamente a la ejecución de esta sección, se procede a almacenar, analizar y depurar la base de datos mediante un Script .m llamado “GuardarDatos”, el cual archiva la base de datos del formato original “.xlsx” de Excel al formato binario “.mat” de MATLAB, en la Figura 3.16. se observa la base de datos en el nuevo formato binario. Este proceso se realiza con la finalidad de agilizar y optimizar la lectura de la base de datos. A su vez los datos de las diferentes mediciones sincrofasoriales de las PMUs se almacenan en variables globales y_i, y los datos correspondientes al tiempo se almacenan en la variable global x.

Si la señal seleccionada es el ángulo del voltaje o el ángulo de la corriente, se ejecuta un proceso adicional, el cual permite corregir el ángulo, este método consiste en reconstruir los valores de fase originales añadiendo múltiplos apropiados de 2π, se lo realiza debido a que el algoritmo que calcula los valores de fase originales de la señal varía entre $-\pi$ y π.

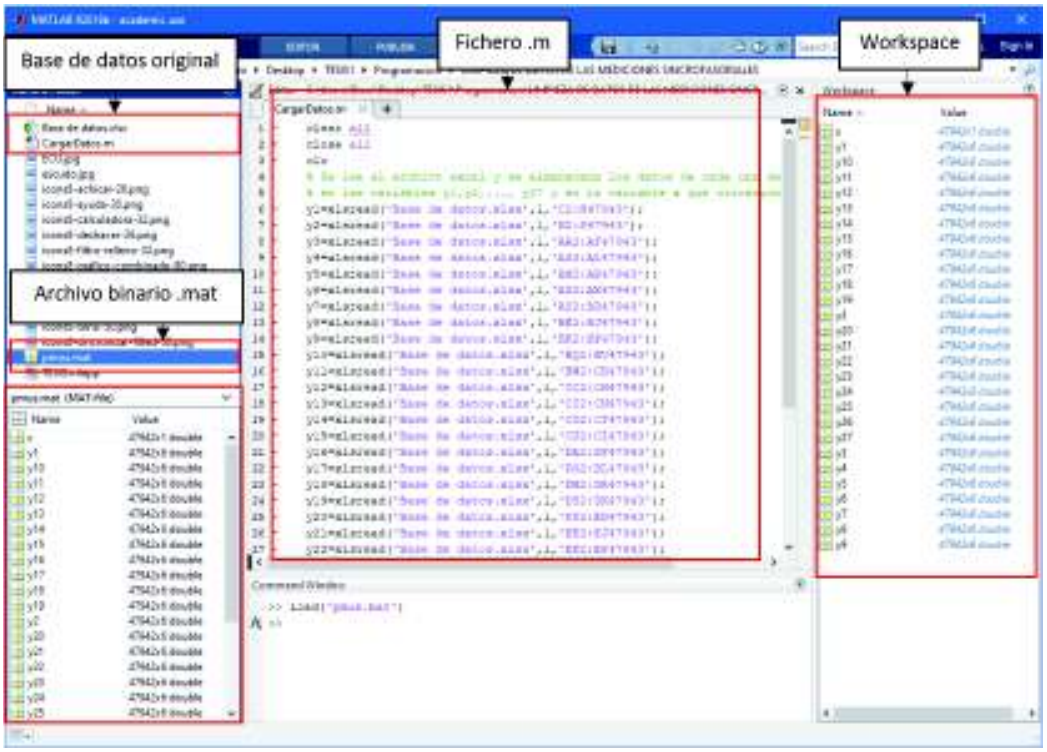


Figura 3.16. Ventana de MATLAB

3.10.2. Sección 2: ESTADÍSTICA DESCRIPTIVA

Una vez que la primera sección ha sido efectuada, se habilita la segunda sección llamada "Estadística Descriptiva", la misma que se encuentra conformada por los botones "Calcular" y "Retroceder". En la figura 3.17. se presenta la estructura de la presente sección.

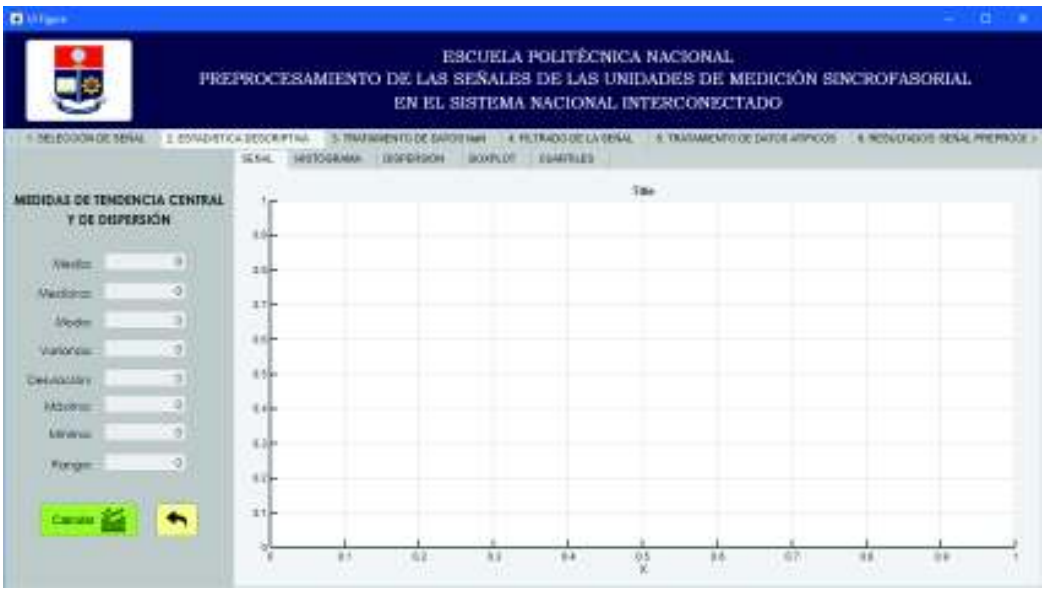


Figura 3.17. Sección 2: Estadística Descriptiva

En esta etapa al ejecutar el botón “Calcular”, la rutina inicia llamando a la función “estadística”, la cual elimina los datos NaN del conjunto de datos para calcular las medidas de tendencia central y de dispersión como son la media, mediana, moda, varianza, desviación, valor máximo, valor mínimo y rango de la señal escogida en la etapa previa, todos estos valores obtenidos son presentados en su respectiva caja. Además, esta sección también contiene cinco pestañas adicionales y en cada una de ellas se encuentra una gráfica que permite visualizar las descripciones de la estadística básica del conjunto de datos, estos gráficos son: histograma, gráfico de la señal con sus respectivas medidas de tendencia central, diagrama de dispersión, BoxPlots y gráfico cuantil. Finalmente, el botón “Retroceder” deshabilita la presente sección y permite regresar a la anterior sección. En la Figura 3.18. se presenta un ejemplo de los resultados que se obtienen en esta etapa.

Figura 3.18. Sección 2: Estadística Descriptiva

3.10.3. Sección 3: TRATAMIENTO DE DATOS NaN

Después del análisis, interpretación y representación del conjunto de datos en la etapa anterior, ahora se habilita la tercera sección “Tratamiento de datos NaN”. Esta sección consiste en la detección, eliminación e imputación de datos faltantes NaN, y para ello se implementaron las siguientes metodologías: interpolación lineal, interpolación cúbica - spline, interpolación cúbica - akima, modelo autorregresivo y media móvil. Las cuales fueron expuestas en el capítulo anterior. Cada uno de estos métodos utilizan los datos disponibles de la señal temporal para estimar los parámetros del modelo con el fin de

reemplazar los datos faltantes con estimaciones adecuadas o sustitutivas. En la figura 3.19. se presenta la estructura de la presente sección.

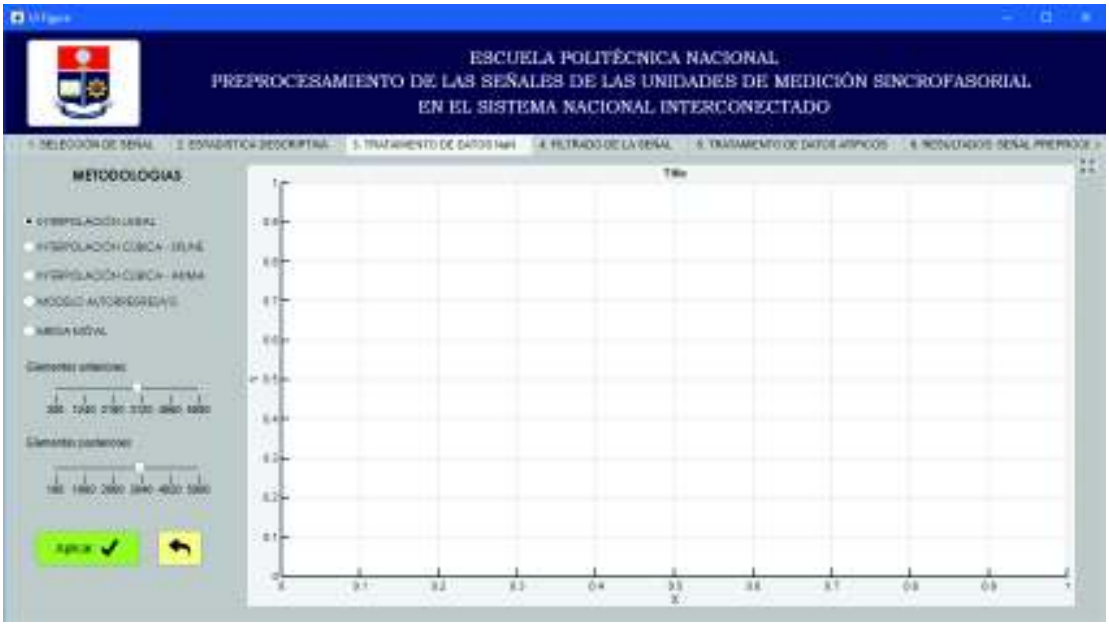


Figura 3.19. Sección 3: Tratamiento de datos NaN

Esta etapa contiene un grupo de radio botones con las metodologías mencionadas, por defecto la interpolación lineal es seleccionada de entre todas estas, salvo que el usuario opte por otra metodología. Las barras deslizantes son parámetros del algoritmo media móvil, donde la barra deslizante para determinar el número de elementos anteriores con respecto a la posición actual tiene un rango entre 300 a 5000 puntos, y para el número de elementos posteriores con respecto a la posición actual tiene un rango de 100 a 5000 puntos, estos rangos fueron determinados heurísticamente.

Al ejecutar el botón “Aplicar” la rutina inicia con la detección de los datos NaN en todo el conjunto de datos, una vez identificados se procede a reemplazarlos por datos estimados por la metodología seleccionada por el usuario. Estos algoritmos emplean los elementos vecinos para completar los datos faltantes, dando como resultado una señal temporal continua, principal requisito para la ejecución de la siguiente sección. El botón “Retroceder” deshabilita la presente sección y permite regresar a la anterior sección. En la Figura 3.20. se observa un ejemplo del tratamiento de datos NaN en la señal de frecuencia a través de la metodología media móvil, donde se halla un conjunto de datos vacíos en el intervalo de tiempo de 575 [s] a 580 [s].

Figura 3.20. Sección 3: Tratamiento de datos NaN

3.10.4. Sección 4: FILTRADO DE LA SEÑAL

Una vez que la señal es continua, es decir, sin datos NaN se procede a la siguiente sección "Filtrado de la señal". Esta etapa tiene como objetivo eliminar el ruido de la señal. Por lo que se presenta dos diferentes tipos de análisis, uno de ellos es Fourier y otro de los análisis es Wavelet. La estructura de esta sección se expone en la Figura 3.21.

Figura 3.21. Sección 4: Filtrado de la señal

Al ejecutar el primer botón de “Filtrar”, inicia la rutina con el cálculo de la transformada discreta de Fourier, con la finalidad de obtener la respuesta en frecuencia de la señal seleccionada. Este análisis permite configurar y definir cualquiera de los tres diferentes tipos de filtros pasa bajos Yulewalk, Butterworth y FIR, los cuales permiten filtrar adecuadamente la señal. En cada uno de los filtros existe la opción de elegir el orden de este, mediante diversas pruebas empíricas se definió un rango de valores adecuados para el orden del filtro, por tanto, el rango del filtro Yulewalk varía entre 15 a 30, el rango del filtro Butterworth varía entre 4 a 20 y el rango del filtro FIR varía entre 5 a 20, y se restringió los valores que se encuentran fuera de los rangos definidos. Por defecto se encuentran definidos los valores adecuados del orden de cada filtro. Otros parámetros definidos para los filtros fueron la frecuencia de corte y la frecuencia de sampling.

En la Figura 3.22. se puede observar los resultados de la señal de frecuencia filtrada mediante el análisis de Fourier junto con el filtro Yulewalk de orden 15, de esta manera se exponen las gráficas de la respuesta en frecuencia de señal y el filtro Yulewalk, y a su vez la gráfica de la señal original y la señal filtrada.

Figura 3.22. Sección 4: Filtrado de la señal

El algoritmo de wavelet fue descrito en el capítulo anterior, y básicamente descompone la señal en un número determinado de niveles, en cada nivel se separa la alta frecuencia de la baja frecuencia y finalmente se procede a filtrar la señal.

Al ejecutar el segundo botón de “Filtrar”, inicia la rutina con el cálculo de la transformada discreta de Wavelet, la cual descompone la señal a través de la familia Daubechies wavelet

(utilizadas normalmente para eliminar el ruido, debido a la propiedad de ortogonalidad que posee) en el número de niveles de descomposición seleccionado por el usuario. El rango de este parámetro varía de 2 a 5, se restringieron los valores que no pertenecen a este rango, pero el cuarto nivel de descomposición es normalmente utilizado para la aplicación de eliminar ruido. Finalmente se procede a comparar los valores entre el umbral y los coeficientes Wavelet calculados en la descomposición, si estos coeficientes se encuentran por debajo del umbral son reemplazados por cero, obteniendo, así como resultado la señal filtrada.

En la Figura 3.23 se observa los resultados de la señal de frecuencia filtrada mediante el análisis de wavelet, de tal manera se exponen las gráficas de la señal original, el ruido de la señal, la señal filtrada.

Figura 3.23. Sección 4: Filtrado de la señal

3.10.5. Sección 5: TRATAMIENTO DE DATOS ATÍPICOS

Está sección es de suma importancia, debido a que este tipo de datos sesga considerablemente los resultados en cualquier análisis. Los valores atípicos son datos diferentes a los valores restantes del conjunto de datos. Después de la ejecución de las etapas previas, la presente etapa consiste en la detección e imputación de los datos atípicos del conjunto de datos, la cual finaliza con el desarrollo de la técnica de limpieza de datos.

Para tratar esta problemática se implementaron las siguientes metodologías: Regresión lineal local “LOWESS”, Regresión cuadrática local “LOESS”, Savitzky-Golay “SGOLAY” y k-means, las cuales fueron expuestas anteriormente. Todos estos métodos se basan en ajustar una función suave y fina a la señal original, con excepción del método k-means, el cual se basa en un algoritmo de agrupación. En la Figura 3.24. se presenta la estructura correspondiente a esta sección. Esta pestaña se divide en dos partes, la primera contiene las tres primeras metodologías mencionadas y la segunda parte contiene la metodología K-means, a su vez incluye dos pestañas adicionales que presentan las respectivas gráficas.

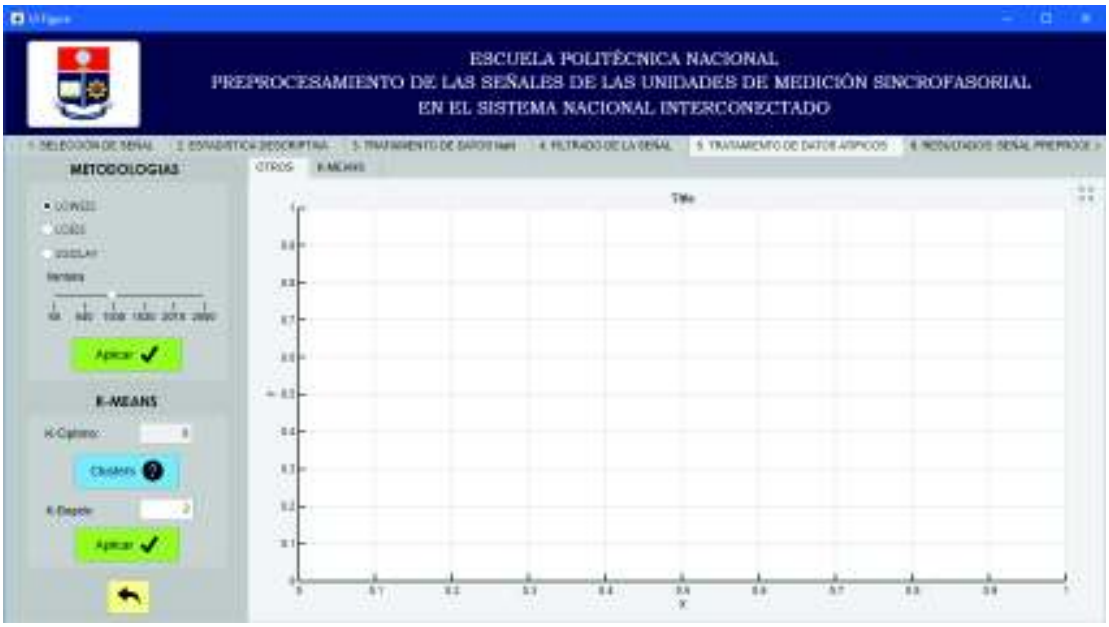


Figura 3.24. Sección 5: Tratamiento de datos atípicos

Al ejecutar el primer botón de “Aplicar”, inicia la rutina con la lectura de los botones para efectuar el método escogido por el usuario, cada uno de estos métodos se calculan sobre una ventana deslizante que tiene por longitud el número de puntos definido por el usuario, a través de la barra deslizante. El rango de la ventana deslizante varía entre 50 y 2500 puntos, para valores menores a 50 el algoritmo es más resistente a datos atípicos y para valores mayores a 2500 se distorsiona la forma de la señal original, estos valores se determinaron heurísticamente.

En la Figura 3.25. se expone la gráfica con la señal de frecuencia original junto con la señal de frecuencia suavizada y sin datos atípicos, sobre una ventana deslizante de aproximadamente 1400 puntos.

Figura 3.25. Sección 5: Tratamiento de datos atípicos

A continuación, se presenta el desarrollo de la metodología K-means, la cual consiste en un algoritmo de agrupación que se basa en la distancia, fracciona el conjunto de datos en k clústeres para la obtención de información sobre estos. Al ejecutar el botón de "Clusters", mediante el criterio de Calinski-Harabasz, se calcula el número k óptimo de clústeres en los que el conjunto de datos podría ser dividido para un mejor resultado. El usuario tiene la opción de elegir el número de k clústeres en un rango de 2 a 12, este parámetro fue determinado de forma empírica, y por defecto k tiene un valor de 2 clústeres para cualquier análisis.

Al ejecutar el botón "Aplicar", la rutina inicia con la lectura de la caja que contiene el número de k clústeres seleccionado por el usuario y se procede al cálculo del algoritmo K-means, una vez separado en k clústeres el conjunto de datos se procede a evaluar los datos atípicos de cada uno de los clústeres usando el criterio de percentiles, el cual considera por datos atípicos a los valores inferiores a $Q_1 - 1.5 * IQR$ o superiores a $Q_3 + 1.5 * IQR$, al ser identificados los valores atípicos se eliminan y reemplazan por un método de imputación.

En la Figura 3.26. se observa cómo se efectúa el cálculo del K-Óptimo, debido a que este proceso tarda se presenta una ventana adicional, la cual indica el porcentaje de carga del proceso hasta finalmente obtener el resultado.

Figura 3.26. Sección 5: Tratamiento de datos atípicos

En la Figura 3.27. se observa el resultado del k-Óptimo con un valor de 6 clústeres para el conjunto de datos de la señal de frecuencia, y se ingresa este valor en el campo del k-Elegido para proceder a ejecutar el algoritmo K-means, debido a que este proceso tarda se presenta una ventana adicional, la cual indica el porcentaje de carga del proceso hasta finalmente obtener el resultado.

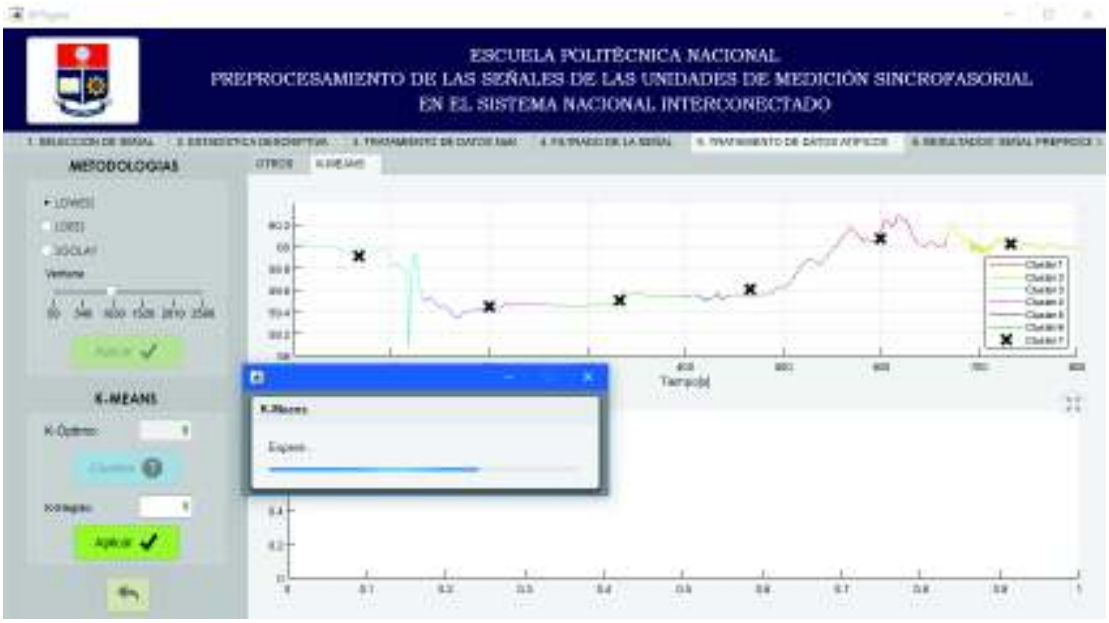


Figura 3.27. Sección 5: Tratamiento de datos atípicos

En la Figura 3.28. se expone dos gráficas de resultados, una de ellas contiene la señal de la frecuencia dividida en k clústeres con sus respectivos centroides, y la otra gráfica contiene la señal de frecuencia con la exclusión de sus datos atípicos.

Figura 3.28. Sección 5: Tratamiento de datos atípicos

3.10.6. Sección 6: SEÑAL PREPROCESADA

Esta sección es la última de todo el proceso, expone las gráficas correspondientes a la señal original seleccionada en la primera etapa y la señal preprocesada mediante las técnicas de limpieza de datos presentada en la quinta etapa. A su vez muestra el cálculo de la raíz cuadrada del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2), estos errores miden la desviación entre la señal original y su versión preprocesada, de tal manera que esto permite determinar y comparar el aporte en la elección de los diferentes algoritmos utilizados en cada etapa.

En la Figura 3.29. se puede observar la estructura de esta sección, la cual está conformada por las dos gráficas mencionadas, la tabla del cálculo de errores y los botones llamados “Resultados” y ”Guardar”.

Figura 3.29. Sección 6: Señal Preprocesada

Al ejecutar el botón “Resultados”, la rutina inicia con la gráfica tanto de la señal original como la señal preprocesada mediante las técnicas de limpieza de datos y calcula los errores que existen entre ambas señales, en la Figura 3.30. se expone un ejemplo de la presente sección con referencia a la señal de frecuencia.

Figura 3.30. Sección 6: Señal Preprocesada

Al ejecutar el “Guardar”, la rutina exporta y guarda el conjunto de valores de los datos de la señal preprocesada mediante las técnicas de limpieza de datos, en un archivo con extensión “.dat” en el workspace. En la Figura 3.31. se observa el archivo PMU-AGOYÁN-BAÑOS1-SEÑAL-FRECUENCIA.dat, el cual contiene el vector de valores de la señal preprocesada.

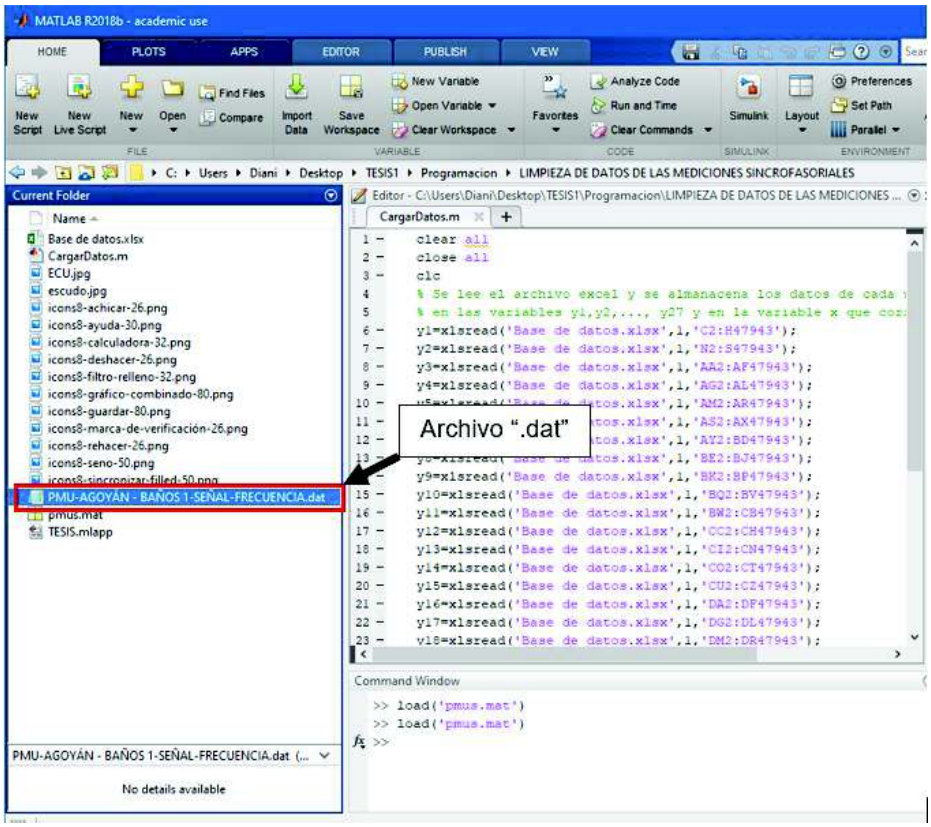


Figura 3.31. Ventana de MATLAB

Además, existen dos botones adicionales: el botón “Retroceder” deshabilita la presente sección y permite regresar a la anterior sección y el botón “Seguir” reinicia todo el proceso para un nuevo cálculo del preprocesamiento de datos de otra señal.

4. RESULTADOS Y DISCUSIÓN

En este capítulo se expone los resultados obtenidos de las simulaciones, mediante aplicación desarrollada para la limpieza de datos de las diferentes mediciones sincrofasoriales de las PMUs instaladas en el Sistema Nacional Interconectado. Para ello, se presenta diferentes casos de estudio que permitirán determinar y verificar cuales, de las diferentes metodologías empleadas en cada una de las secciones de la aplicación, obtuvieron una mejor calidad de los resultados, para tratar los datos ruidosos, faltantes, atípicos e inconsistentes de las diferentes series temporales, y de esta manera obtener una base de datos fidedigna para su posterior uso.

4.1. Depuración de la base de datos

La base de datos que fue empleada en el presente estudio técnico contiene información de 29 PMUs instaladas en el Sistema Nacional Interconectado. Cada Unidad de Medición Sincrofasorial registra una tabla que contiene 6 columnas, las cuales representan un conjunto de atributos correspondientes a las diferentes mediciones sincrofasoriales de frecuencia, tasa de frecuencia, fasores de secuencia positiva de las ondas sinusoidales de voltaje y corriente, y también contiene 47942 filas, que corresponden a los valores estimados del atributo en el tiempo. Cada medición se ha obtenido a una tasa de muestreo de 60 muestras por segundo en un intervalo de tiempo de 800 [s].

Este esquema ha facilitado la depuración de la base de datos, por lo que primero se realizó una observación general de las tablas correspondientes a cada PMU, cada una de estas se encuentran organizadas de acuerdo con la Tabla 2.2. El proceso de depuración se lo realizó mediante un Script .m llamado “GuardarDatos”, el cual se encarga de extraer e importar la base de datos desde formato original “.xlsx” de Excel al formato binario “.mat” de MATLAB, este Script no modifica la base de datos original del formato “.xlsx” de Excel, inicia con la rutina de la lectura y extracción de la base de datos, al momento de importar las tablas al formato binario “.mat” de MATLAB se obtiene un total de 30 tablas debido a que la tabla de la PMU de SANTA ROSA - TOTORAS 2 se encuentra duplicada, por lo cual, se procede a eliminar la copia de esta tabla. A su vez, existen dos tablas que se encuentran completamente vacías, la cuales corresponden a las tablas de la PMU de la PMU de C. ESMERALDAS y de C. TRINITARIA - TV1, adicionalmente la primera tabla vacía a eliminar presenta sus columnas dispersas en las demás tablas correspondientes a otras PMUs.

Finalmente, se obtiene como resultado un archivo “.mat”, el cual contiene 27 tablas correspondientes al resto de PMUs de acuerdo con la Tabla 2.2., cada una de estas tablas

contiene las diferentes señales temporales en el siguiente orden: frecuencia [Hz], derivada de la frecuencia [Hz/s], voltaje [V], ángulo del voltaje [rad], corriente [A], ángulo de la corriente [rad].

4.2. Casos de estudio

La aplicación para el preprocesamiento de señales mediante las técnicas de limpieza de datos presenta una gran cantidad de opciones en el uso y combinación de las diferentes metodologías expuestas en cada sección, por lo que se definió un conjunto reducido de pruebas o casos de estudio que permiten exponer los resultados de forma general, y se presentan a continuación:

4.2.1. Caso 1: SEÑAL DE FRECUENCIA

4.2.1.1. Resultado de la sección 1: SELECCIONAR SEÑAL

Para el presente caso de estudio se escogió por señal a la frecuencia de la PMU AGOYÁN - BAÑOS 1, como se puede observar en la Figura 4.1.

Figura 4.1. Sección 1: Selección de Señal

En la Figura 4.2., se obtiene una mejor apreciación de la señal temporal de frecuencia, la cual a simple vista presenta discontinuidades, y un pico pronunciado al tiempo de 117 [s], estas observaciones serán tratadas en las secciones posteriores.

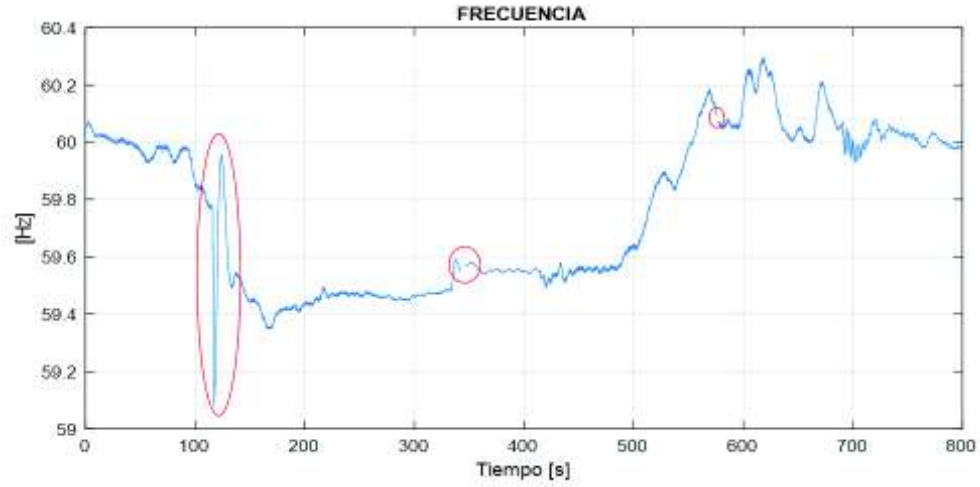


Figura 4.2. Señal de la frecuencia de la PMU AGOYÁN - BAÑOS 1

4.2.1.2. Resultado de la sección 2: ESTADÍSTICA DESCRIPTIVA

Previo al uso de cualquier tipo de metodología sobre la señal de frecuencia, se procede a tener una imagen general de los datos. En la Tabla 4.1. se observa los resultados de las medidas de tendencia central y de dispersión, las cuales resumen y describen al conjunto de datos que representan la señal de frecuencia.

Tabla 4.1. Medidas de tendencia central y de dispersión

Medidas de tendencia central y de dispersión	Valor [Hz]
Media	59,77
Mediana	59,8
Moda	59,47
Varianza	0,07258
Desviación	0,2694
Máximo	60,3
Mínimo	59,08
Rango	1,214

Como se puede observar en la Figura 4.3. se expone la gráfica de la frecuencia con sus respectivas medidas de tendencia central, con un valor promedio de 59,77 [Hz], el valor que se ubica en el centro es de 59,8 [Hz], y el valor que ocurre con mayor frecuencia es de 59,47 [Hz] en el conjunto de datos. En este caso el valor de la mediana es mayor que el valor de la media, por lo que la distribución de los datos se encuentra sesgada hacia abajo.

Figura 4.3. Sección 2: Estadística Descriptiva

En la Figura 4.4. se expone el resultado del histograma de la señal de frecuencia, en el cual se observa dos picos pronunciados por lo que presenta la característica bimodal, a su vez se aprecia que la distribución no es simétrica y los valores más frecuentes de la señal comprenden los rangos de 59,4 a 59,6 [Hz] y 59,95 a 60,1 [Hz], mientras que los valores atípicos según la gráfica se consideran en un rango de 59,05 a 59,35 [Hz].

Figura 4.4. Sección 2: Estadística Descriptiva

En la Figura 4.5. se presenta el resultado del gráfico de dispersión de la señal de frecuencia, y se observa grupos de puntos en el tiempo, de los cuales sobresale un grupo específico de puntos en el tiempo de 114 a 119 [s], los cuales son considerados como valores atípicos.

Figura 4.5. Sección 2: Estadística Descriptiva

El resultado de la gráfica BoxPlot de la señal de la frecuencia se expone en la Figura 4.6., en la cual el conjunto de datos de la señal temporal se divide en cuatro partes, los extremos de la caja se encuentran en el cuartil Q_1= 59,5 [Hz] y el cuartil Q_3 =60,01 [Hz], mientras que el cuartil Q_2 = 59,8 [Hz] indica la mediana, las dos líneas (llamadas bigotes) fuera de la caja se extienden a la observación mínima equivalente a 59,08 [Hz] y a la observación máxima equivalente a 60,3 [Hz]. En este gráfico no se identificó valores atípicos, debido a que ningún valor cae el 1,5 por debajo o por encima del rango intercuartílico, y el conjunto de datos presenta una desviación estándar muy baja equivalente a 0,2694 [Hz], lo cual significa que las observaciones de datos tienden a ser muy cercanas a la media.

Mientras que en la Figura 4.7. se observa el gráfico cuantil de la señal de frecuencia, el cual se encuentra relacionada con la gráfica BoxPlot, y presenta como resultado la señal con el percentil 0,25 corresponde al cuartil Q_1, el percentil 0,50 es la mediana y el percentil 0,75 es Q_3.

Figura 4.6. Sección 2: Estadística Descriptiva

Figura 4.7. Sección 2: Estadística Descriptiva

4.2.1.3. Resultado de la sección 3: TRATAMIENTO DE DATOS NaN

Para el tratamiento de datos NaN de la señal de frecuencia seleccionada, primero se examinó la existencia de datos NaN en el conjunto de valores de los datos, por lo que se obtuvo como resultado la presencia de estos, tal como se indica en la Tabla 4.2.

Tabla 4.2. Contenido de datos NaN en la señal

Cantidad de datos NaN	Intervalo de tiempo [s]
312	342 a 347,1833
107	575,7667 a 577,5333
Total: 419 datos NaN	

A continuación, se procedió a la detección e imputación de datos NaN de la señal de frecuencia, por lo cual se empleó cada uno de los métodos expuestos: el método de interpolación lineal obtiene como resultado la señal expuesta en la Figura 4.8., el método de interpolación cúbica “Spline” obtiene como resultado la señal expuesta en la Figura 4.9., el método de interpolación cúbica “Akima” obtiene como resultado la señal expuesta en la Figura 4.10., el modelo autorregresivo obtiene como resultado la señal expuesta en la Figura 4.11., y finalmente el método de media móvil, el cual utilizó una ventana de 500 puntos anteriores y 783 puntos posteriores para el cálculo del algoritmo, obtiene como resultado la señal expuesta en la Figura 4.12., en este último método si se escogen valores inadecuados puede dar resultados no deseados como aproximaciones incompletas o con picos pronunciados.

Figura 4.8. Sección 3: Tratamiento de datos NaN

Figura 4.9. Sección 3: Tratamiento de datos NaN

Figura 4.10. Sección 3: Tratamiento de datos NaN

Figura 4.11. Sección 3: Tratamiento de datos NaN

Figura 4.12. Sección 3: Tratamiento de datos NaN

La metodología seleccionada para la señal de la frecuencia se determinó de forma empírica, este proceso consistió en el reemplazo de una porción del conjunto de datos de la señal por datos NaN. Con la finalidad de aplicar cada una de las metodologías para el tratamiento de datos NaN, y así comparar los datos originales de la señal con los datos estimados por cada una de las metodologías. En la Figura 4.13. se puede observar a la

señal original de frecuencia con el contenido de un nuevo conjunto de datos NaN que comprenden el intervalo de 275 a 279,2 [s].

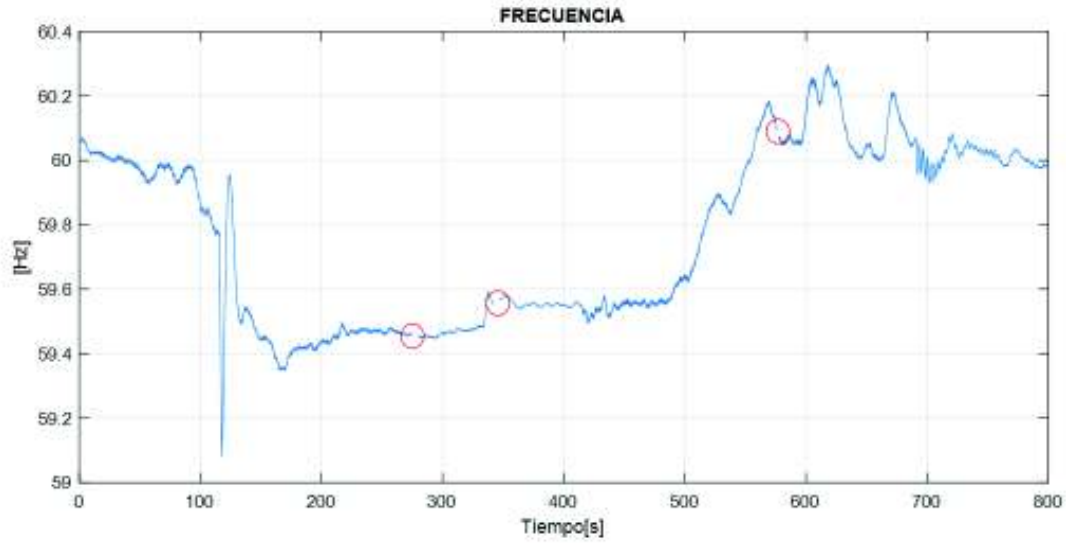


Figura 4.13. Señal de la frecuencia agregada un intervalo de datos NaN

El resultado de aplicar cada una de las diferentes metodologías para el tratamiento de datos NaN de la nueva señal de frecuencia se expone en la Figura 4.14.

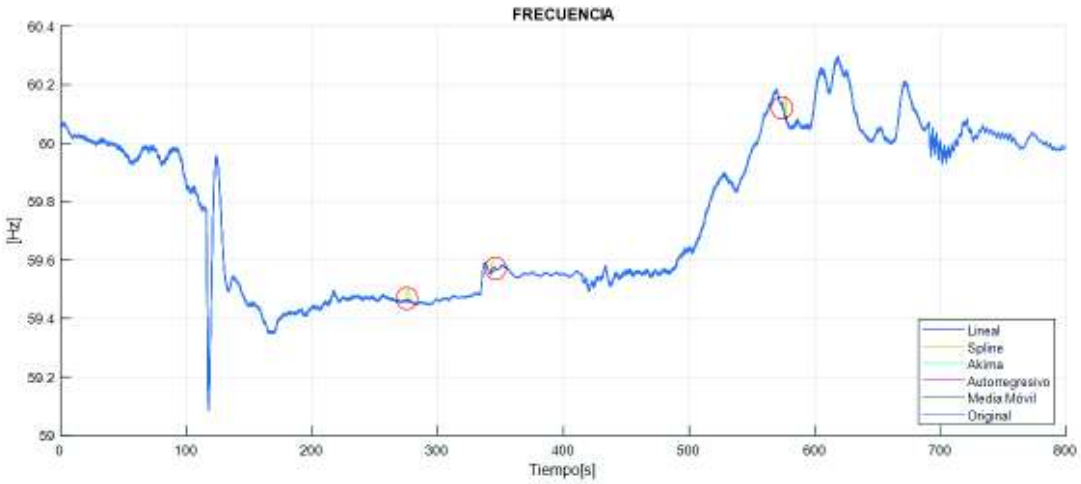


Figura 4.14. Señal de la frecuencia agregada un intervalo de datos NaN

En la Figura 4.15 se observa un enfoque del resultado de la estimación de cada uno de los modelos para la imputación de datos NaN del nuevo conjunto de datos NaN de la señal de frecuencia. A simple inspección se determina que el método que se aproximó más a los datos originales de señal de frecuencia fue el modelo autorregresivo.

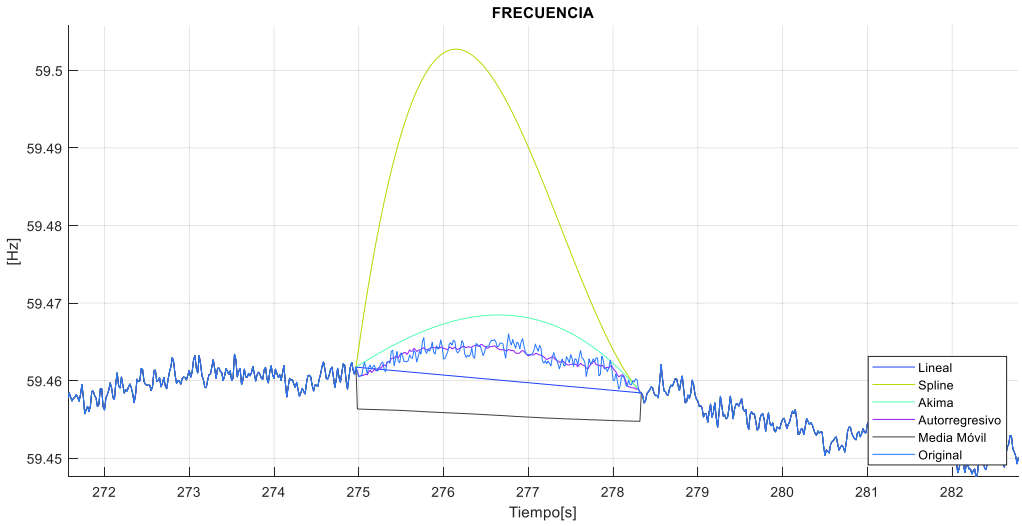


Figura 4.15. Señal de la frecuencia agregada un intervalo de datos NaN

En la Tabla 4.3. se presenta el cálculo de la raíz cuadrada del error medio cuadrático de los diferentes métodos empleados para el tratamiento de datos NaN de la señal de frecuencia, como se puede apreciar el modelo autorregresivo presenta como resultado un valor RMSE muy bajo, por lo que en comparación a los demás modelos este resulta ser el óptimo para su presente aplicación.

Tabla 4.3. Cálculo de los errores de los métodos para el tratamiento de datos NaN en la señal de la frecuencia

Metodología	RMSE
Lineal	0,0032
Spline	0,0259
Akima	0,0032
Autorregresivo	0,0008
Media Móvil	0,0075

4.2.1.4. Resultado de la sección 4: FILTRADO DE LA SEÑAL

Una vez que la señal es continua, se procede a filtrar esta. En esta sección, previo al uso del análisis de Fourier se definió internamente los parámetros de los filtros como son la frecuencia de corte con un valor de 5 [Hz] para las diferentes señales con la excepción de la señal de la derivada de frecuencia, ya que esta tiene un valor de 15 [Hz], y otro de los parámetros que se definió fue la frecuencia de sampling, la cual tiene un valor de 60 [Hz] para todas las señales.

A continuación, en la Figura 4.16. se presenta los resultados obtenidos del análisis de Fourier junto con el filtro Yulewalk de orden 15. En la primera gráfica de este análisis se obtiene la respuesta de la señal en el dominio de la frecuencia, en la segunda gráfica se observa la respuesta del filtro pasa bajo en el dominio de la frecuencia, y como se puede apreciar presenta un pequeño pico en su forma. Sí se selecciona un orden menor al del rango definido no filtra la señal y en cambio un orden mayor al del rango definido tiende a ser inestable, en la tercera gráfica se presenta la señal original y en la cuarta gráfica la señal filtrada.

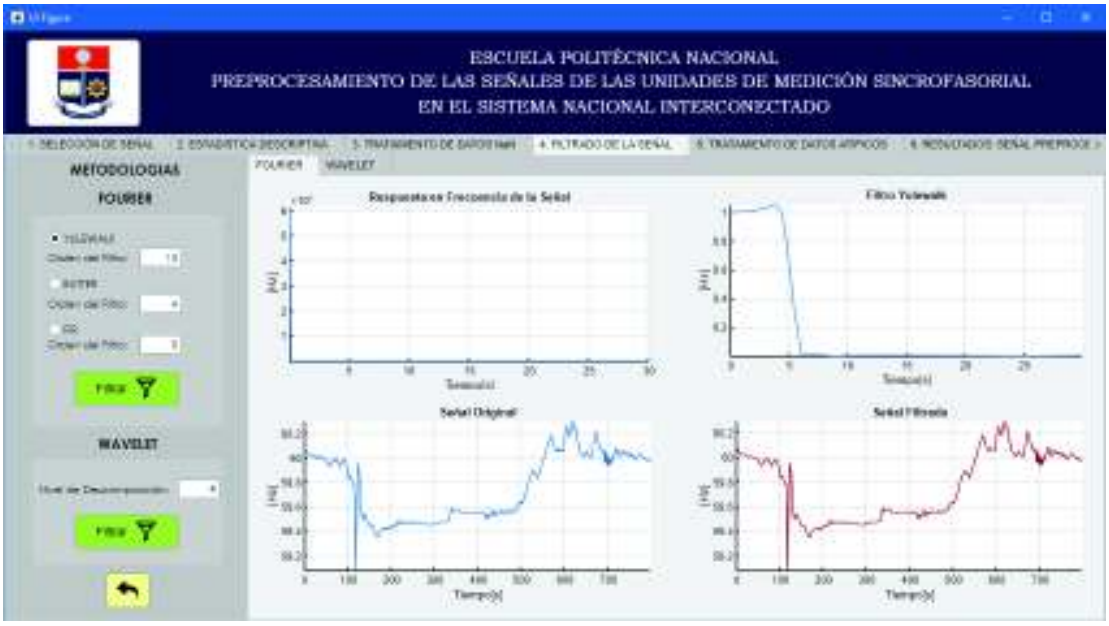


Figura 4.16. Sección 4: Filtrado de la señal

En la Figura 4.17. se presenta los resultados obtenidos del análisis de Fourier junto con el filtro Butterworth de orden 4. En la primera gráfica de este análisis se obtiene la respuesta de la señal en el dominio de la frecuencia, en la segunda gráfica se observa la respuesta del filtro pasa bajo en el dominio de la frecuencia, y como se puede apreciar a diferencia del anterior filtro este es suave no presenta ningún pico en su forma y decae rápidamente después de haber llegado a la frecuencia de corte, en la tercera gráfica se presenta la señal original y en la cuarta gráfica la señal filtrada.

Figura 4.17. Sección 4: Filtrado de la señal

En la Figura 4.18. se presenta los resultados del análisis de Fourier junto con el filtro FIR de orden 5. En la primera gráfica se obtiene la respuesta de la señal en el dominio de la frecuencia, en la segunda gráfica se observa la respuesta del filtro pasa bajo en el dominio de la frecuencia, y como se puede apreciar la diferencia de este filtro es que presenta una forma plana y lisa, y demora en decaer después de haber llegado a la frecuencia de corte, en la tercera gráfica se presenta la señal original y en la cuarta gráfica la señal filtrada.

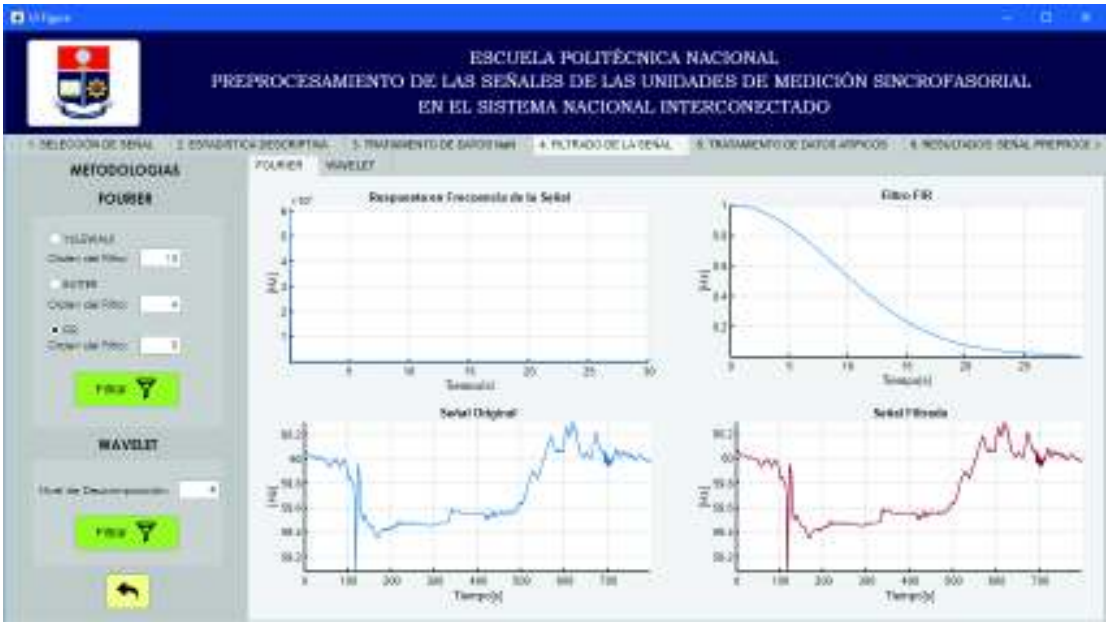


Figura 4.18. Sección 4: Filtrado de la señal

Otro de los métodos para eliminar el ruido es el análisis de wavelet, por lo cual se efectuó la descomposición wavelet en cuatro niveles para la señal de la frecuencia, y de esta manera se adquirió la señal filtrada. En la Figura 4.19. se expone los resultados que se obtuvieron de este análisis, la primera gráfica contiene la señal original, la segunda gráfica el residuo o ruido eliminado de la señal y finalmente la tercera gráfica la señal filtrada.

Figura 4.19. Sección 4: Filtrado de la señal

En todos los resultados que se obtuvieron de los métodos usados para filtrar el ruido de la señal de frecuencia, se observó que efectivamente las altas frecuencias disminuyen en la señal. Para proceder a comparar estos métodos y así escoger el más adecuado, se tomó en cuenta a la raíz cuadrada del error medio cuadrático. En la Tabla 4.4. se presenta el cálculo RMSE de los diferentes métodos empleados. Como se puede observar el análisis de Fourier y el filtro FIR presentan el valor más bajo de RMSE en comparación a los demás modelos, por esto resulta ser el modelo óptimo para su presente aplicación.

Tabla 4.4. Cálculo de los errores de los métodos para el filtrado de la señal de la frecuencia

Metodología	RMSE
Fourier y filtro Yulewalk	0,0018
Fourier y filtro Butterworth	0,0017
Fourier y filtro FIR	0,0014
Wavelet	0,0019

4.2.1.5. Resultado de la sección 5: TRATAMIENTO DE DATOS ATÍPICOS

A continuación, se presenta el resultado de los métodos de suavizado y un método de agrupación basado en la distancia, los cuales permiten detectar e imputar la influencia de los valores atípicos y aproximar los datos de la señal a una función estimada.

La metodología Regresión lineal local “LOWESS”, calcula y ajusta un polinomio lineal sobre la señal de frecuencia, mediante una ventana deslizante de 758 puntos de datos, tras realizar varias pruebas empíricas se logró determinar este valor de ventana como el óptimo para el cálculo de este algoritmo. Se expone el resultado en la Figura 4.20., la señal que se presenta en color rojo es la nueva función estimada. Para valores menores del rango definido para la ventana deslizante, el algoritmo es más resistente a datos atípicos y para valores mayores del rango definido se pierde la forma de la señal original.

Figura 4.20. Sección 5: Tratamiento de datos atípicos

La Regresión cuadrática local “LOESS”, calcula y ajusta un polinomio cuadrático sobre la señal de frecuencia, mediante una ventana deslizante de 1030 puntos de datos, tras realizar varias pruebas empíricas se logró determinar este valor de ventana como el óptimo para el cálculo de este algoritmo. Se expone el resultado en la Figura 4.21., la señal que se presenta en color rojo es la nueva función estimada.

Figura 4.21. Sección 5: Tratamiento de datos atípicos

El método Savitzky-Golay “SGOLAY”, calcula y ajusta un polinomio mediante mínimos cuadrados móviles sobre la señal de frecuencia, en una ventana deslizante de 750 puntos de datos. Se expone el resultado en la Figura 4.22., la señal que se presenta en color rojo es la nueva función estimada.

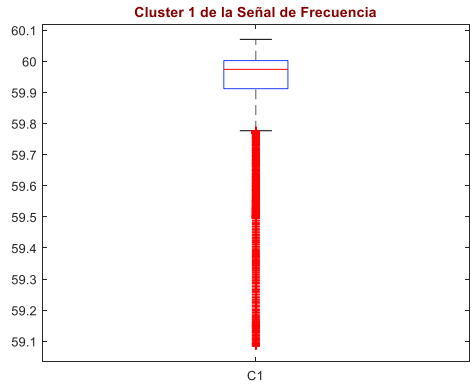
Figura 4.22. Sección 5: Tratamiento de datos atípicos

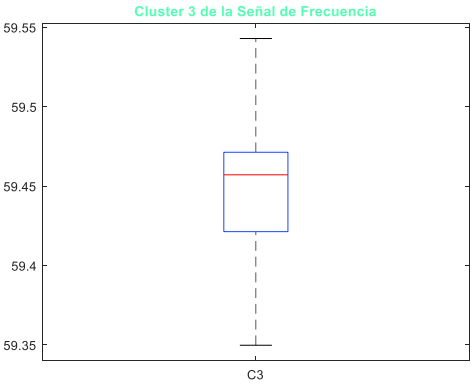
La última metodología k-means es una técnica de agrupamiento que se basa en la distancia, en la cual primero se determinó el k-óptimo y se obtuvo como resultado un valor de 6 clústeres para la señal de la frecuencia. Posteriormente se procedió a evaluar los datos atípicos mediante el algoritmo K-means junto con el criterio BoxPlot, el cual consideró como valores atípicos a los puntos que caen al menos $1{,}5 * IQR$ por encima del tercer cuartil o por debajo del primer cuartil en cada uno de los 6 grupos de la señal de frecuencia. Los resultados se exponen en la Figura 4.23., en la primera gráfica se presenta la señal de frecuencia divida en 6 grupos y en la segunda gráfica el resultado de la señal sin datos atípicos.

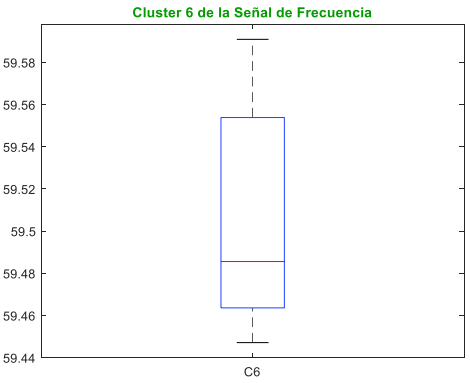
Figura 4.23. Sección 5: Tratamiento de datos atípicos

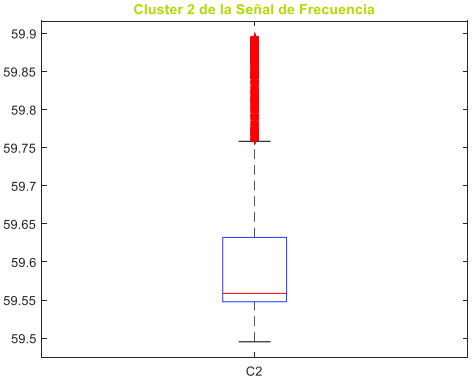
En la Figura 4.24. se presenta los resultados BoxPlot de cada uno de los 6 grupos en el que fue divida la señal de frecuencia. Este análisis muestra los valores atípicos (cruces rojas) previos y posteriores a la imputación de datos mediante el algoritmo K-means. Como se puede observar hay grupos que no presentan valores atípicos y otros en los cuales presentan una gran cantidad de datos atípicos, pero posterior al empleo del análisis existe una disminución de estos, pero no en su totalidad.

Análisis previo a K-means

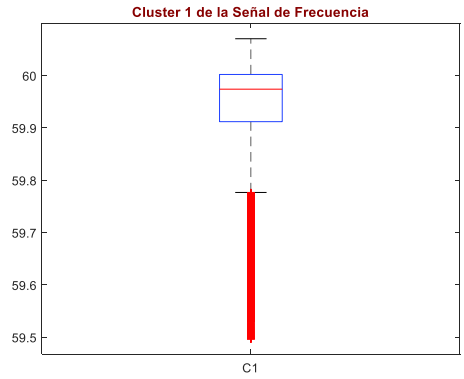


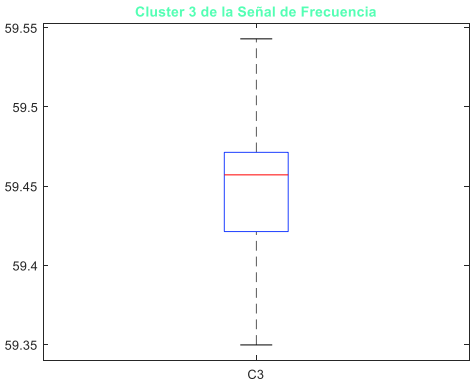


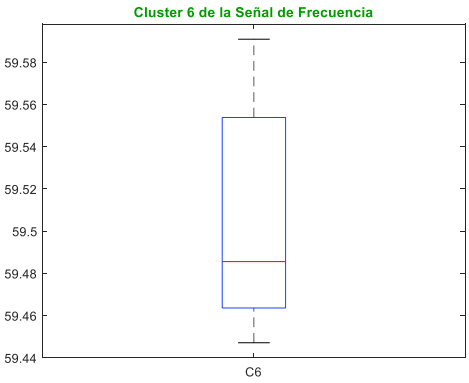


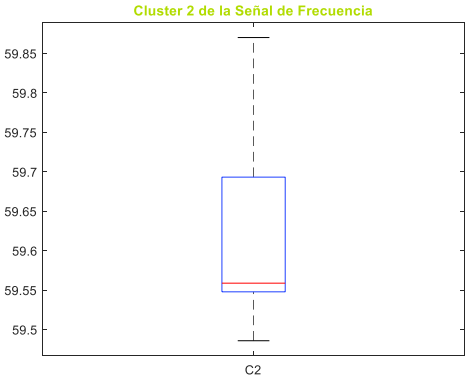


Análisis posterior a K-means









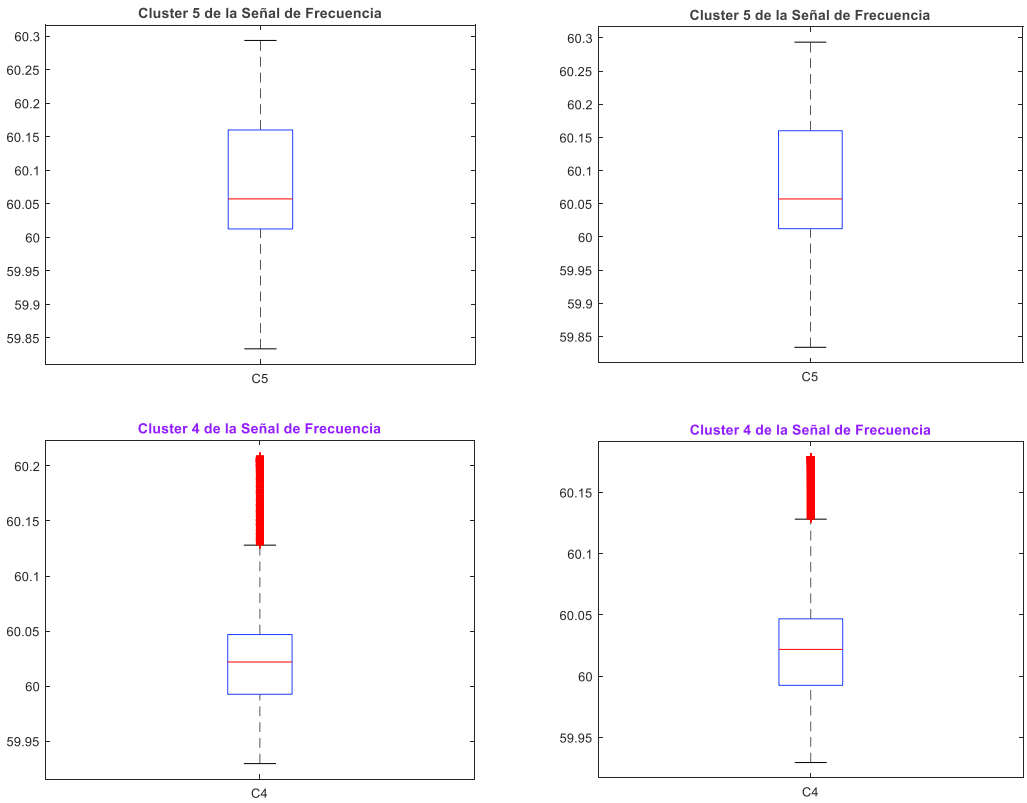


Figura 4.24. Diagrama de caja de cada clúster previo y posterior al análisis K-means

En todos los resultados que se obtuvieron de los métodos usados para el tratamiento de los datos atípicos de la señal de frecuencia, se observó que efectivamente existe una reducción de los datos atípicos. Para proceder a comparar estos métodos y así escoger el más adecuado, se tomó en cuenta a la raíz cuadrada del error medio cuadrático y el coeficiente de determinación. En la Tabla 4.4. se presenta el cálculo RMSE y R^2 de los diferentes métodos empleados. Como se puede observar el método LOESS presentan el valor más bajo de RMSE y casi tiende a ser 1 su coeficiente de determinación, por lo que, en comparación a los demás modelos, este resulta ser el modelo óptimo para su presente aplicación.

Tabla 4.4. Cálculo de los errores de los métodos para el tratamiento de datos atípicos en la señal de frecuencia

Metodología	RMSE	R^2
LOWESS	0,02573	0,9909
LOESS	0,0189	0,9951
SGOLAY	0,0201	0,9944
K-MEANS	0,0397	0,9782

4.2.1.6. Resultado de la sección 6: SEÑAL PREPROCESADA

Después de todo el proceso ejecutado por las técnicas de limpieza de datos para tratar la señal de frecuencia de la PMU AGOYÁN - BAÑOS 1, en esta sección se presenta finalmente el resultado obtenido de la señal junto con la señal original escogida en la sección 1, como se puede observar en la Figura 4.25. Adicionalmente, se incluye en esta sección los valores de la raíz cuadrada del error medio cuadrático y el coeficiente de determinación, los cuales fueron calculados entre la señal final y la original, para de esta manera poder comparar el rendimiento de cada método, y así seleccionar los más adecuados con la finalidad de no perder la forma de la señal original.

En la Figura 4.26. se presenta como se almacenó el nuevo conjunto de valores de los datos de la señal en un archivo PMU-AGOYÁN-BAÑOS1-SEÑAL-FRECUENCIA.dat, después de la finalización de las técnicas de limpieza de datos, con el objetivo de un uso posterior de este archivo para cualquier análisis o estudio.

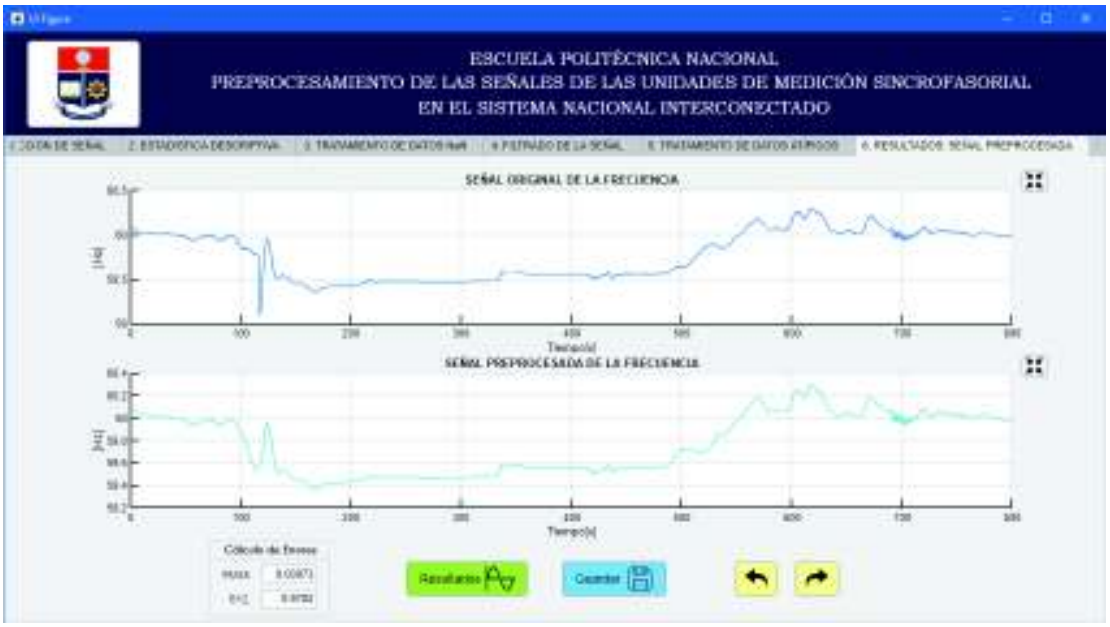


Figura 4.25. Sección 6: Señal Preprocesada

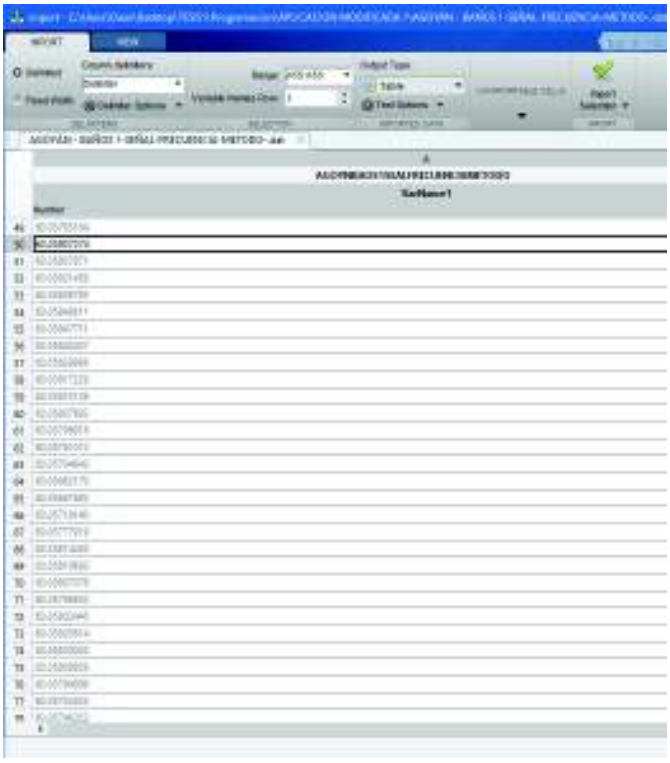


Figura 4.26. Archivo PMU-AGOYÁN-BAÑOS1-SEÑAL-FRECUENCIA.dat

Este proceso permite analizar los resultados obtenidos con respecto al fenómeno físico ocurrido. Desde el punto de vista de la estadística descriptiva mediante el diagrama de caja expuesto en la Figura 4.6., no se logró observar el contenido de datos atípicos en el conjunto de datos de la señal de frecuencia, debido a que de forma general los datos presentan una desviación estándar muy baja equivalente a 0,2694 [Hz], lo cual significa que las observaciones de datos tienden a ser muy cercanas a la media.

Pero mediante el estudio del esquema de alivio de carga se observa que los valores de frecuencia se encuentran en rangos críticos, para una operación normal del sistema eléctrico de potencia, debido a que en un instante el valor mínimo que presenta la señal de frecuencia es de 59,08 [Hz], encontrándose en el tercer paso del esquema de alivio de carga. Otro de los factores que influyen en la obtención de estos valores es el protocolo C37.118-2005 que utiliza la PMU, el cual no considera la dinámica del transitorio durante la ocurrencia de la falla, por lo que procede a estimar datos erróneos hasta que el sistema se estabilice. Entonces mediante el uso de las técnicas de limpieza de datos se puede tratar estos datos como atípicos y así proceder al tratamiento adecuado de la señal de frecuencia.

4.2.2. Caso 1: SEÑAL DE VOLTAJE

4.2.2.1. Resultado de la sección 1: SELECCIONAR SEÑAL

Para el presente caso de estudio se escogió por señal al voltaje de la PMU AGOYÁN - BAÑOS 1, como se puede observar en la Figura 4.27.

Figura 4.27. Sección 1: Selección de Señal

En la Figura 4.28., se obtiene una mejor apreciación de la señal temporal de voltaje, la cual a simple vista presenta discontinuidades, y un pico pronunciado al tiempo de 117 [s], estas observaciones serán tratadas en las secciones posteriores.

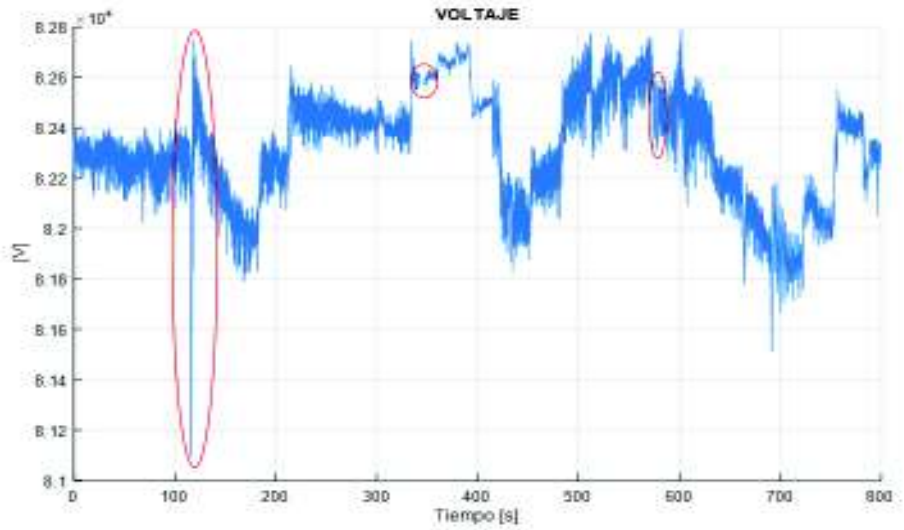


Figura 4.28. Señal de voltaje de la PMU AGOYÁN - BAÑOS 1

4.2.2.2. Resultado de la sección 2: ESTADÍSTICA DESCRIPTIVA

En la Tabla 4.5. se observa los resultados de las medidas de tendencia central y de dispersión, las cuales describen al conjunto de datos que representan la señal de voltaje.

Tabla 4.5. Medidas de tendencia central y de dispersión

Medidas de tendencia central y de dispersión	Valor [V]
Media	8,2321 e+04
Mediana	8,2336 e+04
Moda	8,1093 e+04
Varianza	4,6424 e+04
Desviación	215.464
Máximo	8,2789 e+04
Mínimo	8,1093 e+04
Rango	1695,48

Como se puede observar en la Figura 4.29. se expone la gráfica del voltaje con sus respectivas medidas de tendencia central, el valor promedio y el valor de la mediana se encuentran muy cercanos, pero el valor que ocurre con mayor frecuencia esta considerablemente lejano en este rango de valores.

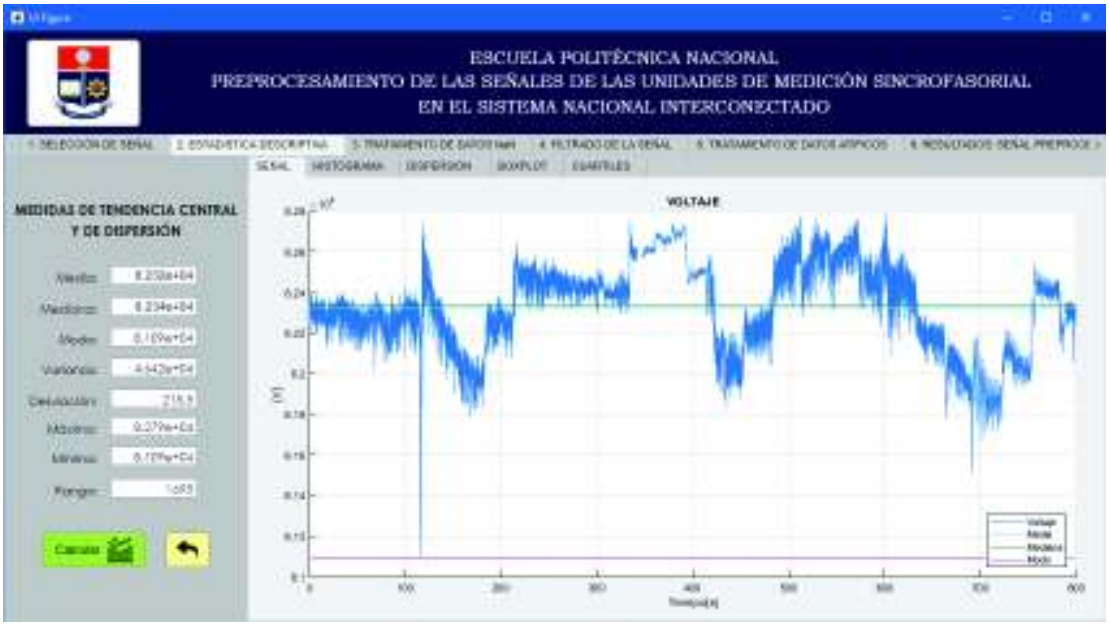


Figura 4.29. Sección 2: Estadística Descriptiva

En la Figura 4.30. se expone el resultado del histograma de la señal de voltaje, en el cual se presenta la característica unimodal, a su vez se aprecia que la distribución es simétrica y los valores más frecuentes comprenden el rango de 8,24 a 8,25 e+04 [V], mientras que los valores atípicos según la gráfica se consideran en un rango de 8,1 a 8,16 e+04 [V].

Figura 4.30. Sección 2: Estadística Descriptiva

En la Figura 4.31. se presenta el gráfico de dispersión de la señal de voltaje, y como se observa que sobresalen dos grupos de puntos en el tiempo, uno de 114 a 119 [s] y otro de 690 a 694 [s], que son considerados como valores atípicos.

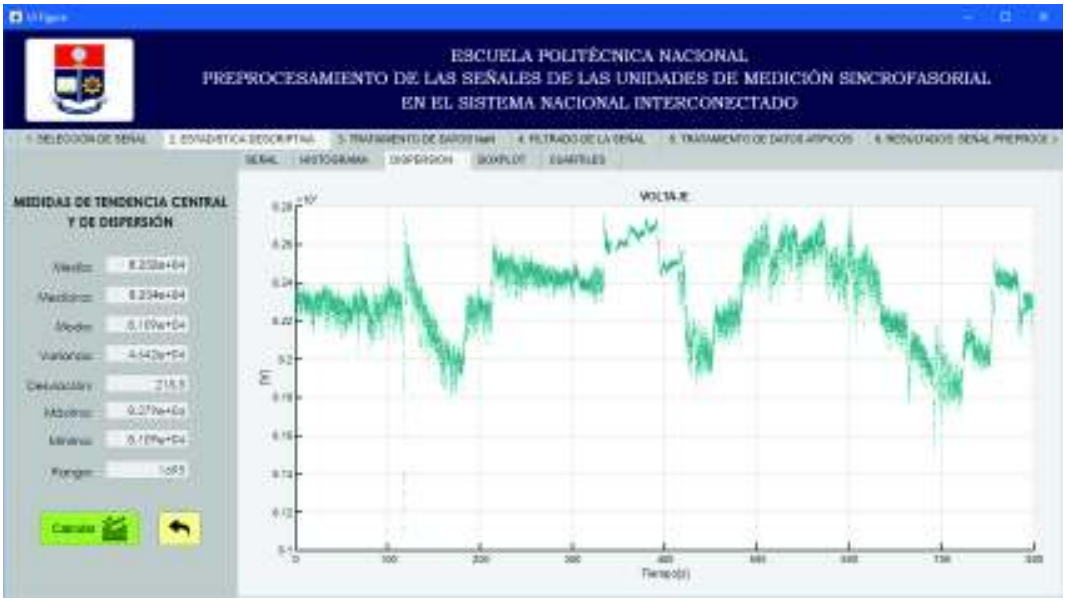


Figura 4.31. Sección 2: Estadística Descriptiva

La Figura 4.32. y 4.33. contienen el resultado de la gráfica BoxPlot y el gráfico cuantil de la señal de voltaje correspondientemente, por lo cual en estas se presenta los percentiles equivalentes a Q_1 = 8,2182 e+04 [V], Q_2 = 8,2336 e+04 [V] y Q_3 = 8,2477 e+04 [V]. El diagrama de caja se extiende entre la observación equivalente a 8,1742 e+04 [V] y 8,2789 e+04 [V], en esta gráfica también se identificaron valores atípicos que se encuentran 1,5 por debajo del rango intercuartílico.

Figura 4.32. Sección 2: Estadística Descriptiva

Figura 4.33. Sección 2: Estadística Descriptiva

4.2.2.3. Resultado de la sección 3: TRATAMIENTO DE DATOS NaN

Para el tratamiento de datos NaN de la señal de voltaje seleccionada, primero se examinó la existencia de datos NaN en el conjunto de valores de los datos, por lo que se obtuvo como resultado la presencia de estos, tal como se indica en la Tabla 4.6.

Tabla 4.6. Contenido de datos NaN en la señal de voltaje

Cantidad de datos NaN	Intervalo de tiempo [s]
312	342 a 347,1833
107	575,7667 a 577,5333
Total: 419 datos NaN	

A continuación, se empleó cada uno de los métodos expuestos en esta sección, de la misma forma que se procedió para escoger el método adecuado para la señal de frecuencia, ahora se realizó para la señal de voltaje obteniendo como resultado la Tabla 4.7., la cual presenta el cálculo de la raíz cuadrada del error medio cuadrático de los diferentes métodos empleados para el tratamiento de datos NaN de la señal de voltaje. Como se puede apreciar el método de media móvil presenta como resultado un valor RMSE más bajo en comparación a los demás modelos y este resulta ser el óptimo para su presente aplicación.

Tabla 4.7. Cálculo de los errores de los métodos para el tratamiento de datos NaN en la señal de voltaje

Metodología	RMSE
Lineal	22,137
Spline	37,507
Akima	22,086
Autorregresivo	24,492
Media Móvil	21,099

El método de media móvil utilizó una ventana de 300 puntos anteriores y 100 puntos posteriores para el tratamiento de los datos NaN de la señal de voltaje, obteniendo como resultado la señal expuesta en la Figura 4.34.

Figura 4.34. Sección 3: Tratamiento de datos NaN

4.2.2.4. Resultado de la sección 4: FILTRADO DE LA SEÑAL

Se utilizó cada uno de los métodos expuestos en esta sección obteniendo como resultado la Tabla 4.8., la cual presenta el cálculo de la raíz cuadrada del error medio cuadrático de los diferentes métodos empleados para el filtrado de la señal de voltaje. Como se puede apreciar el método de Wavelet presenta como resultado un valor RMSE más bajo en comparación a los demás modelos y este resulta ser el óptimo para su presente aplicación.

Tabla 4.8. Cálculo de los errores de los métodos para el filtrado de la señal de voltaje

Metodología	RMSE
Fourier y filtro Yulewalk	19,398
Fourier y filtro Butterworth	18,396
Fourier y filtro FIR	12,195
Wavelet	3,764

Se efectuó la descomposición wavelet en cuatro niveles para la señal de voltaje, y de esta manera se adquirió la señal filtrada. En la Figura 4.35. se expone los resultados que se obtuvieron de este análisis, la primera gráfica contiene la señal original, la segunda gráfica el residuo o ruido eliminado de la señal y finalmente la tercera gráfica la señal filtrada.

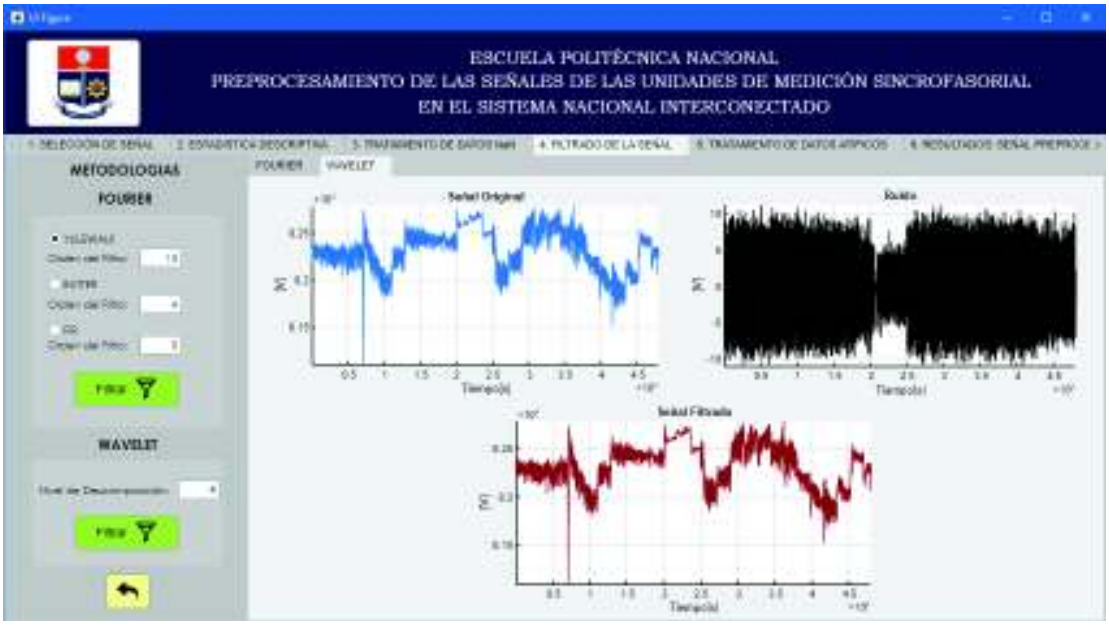


Figura 4.35. Sección 4: Filtrado de la señal

4.2.2.5. Resultado de la sección 5: TRATAMIENTO DE DATOS ATÍPICOS

En todos los resultados que se obtuvieron de los métodos usados para el tratamiento de los datos atípicos de la señal de voltaje, se observó que efectivamente existe una reducción de los datos atípicos. Para proceder a comparar estos métodos y así escoger el más adecuado, se tomó en cuenta a la raíz cuadrada del error medio cuadrático y el coeficiente de determinación. En la Tabla 4.9. se presenta el cálculo RMSE y R^2 de los diferentes métodos empleados. Como se puede observar el método K-means presentan el valor más bajo de RMSE y casi tiende a ser 1 su coeficiente de determinación, por lo que, en comparación a los demás modelos, este resulta ser el modelo óptimo para su presente aplicación.

Tabla 4.9. Cálculo de los errores de los métodos para el tratamiento de datos atípicos en la señal de voltaje

Metodología	RMSE	R^2
LOWESS	51,02	0,944
LOESS	51,19	0,9438
SGOLAY	53,53	0,9385
K-MEANS	20,39	0,9911

En la Figura 4.36. se observa los resultados del método K-means, en el cual primero se determinó el k-óptimo y se obtuvo como resultado un valor de 10 clústeres para la señal de voltaje. En la primera gráfica se presenta la señal de frecuencia divida en 6 grupos y en la segunda gráfica el resultado de la señal sin datos atípicos

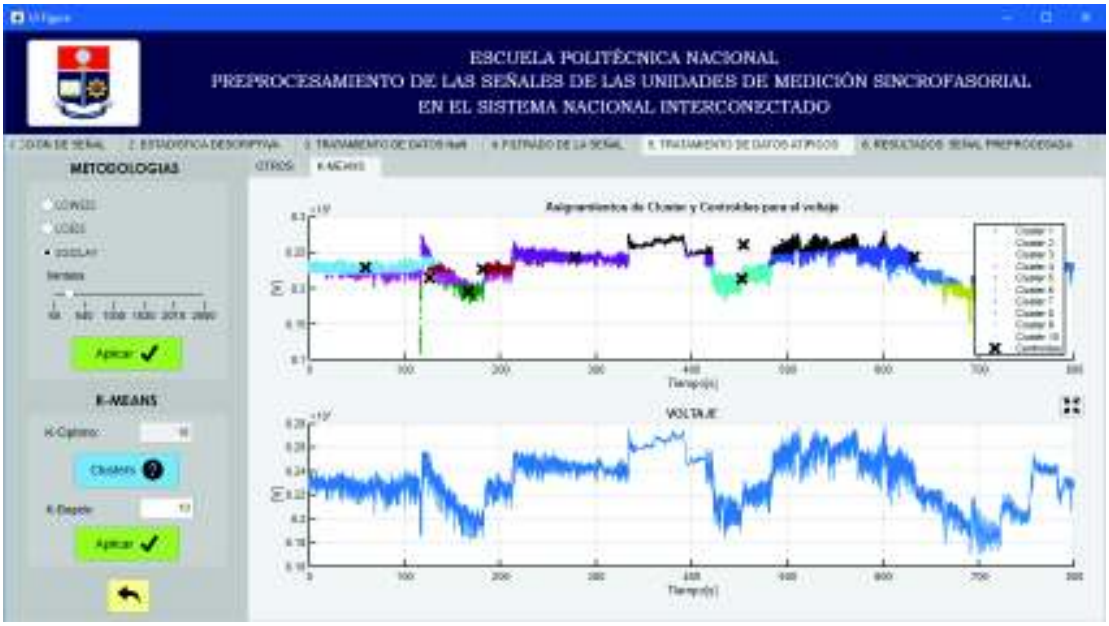


Figura 4.36. Sección 5: Tratamiento de datos atípicos

4.2.2.6. Resultado de la sección 6: SEÑAL PREPROCESADA

Después de todo el proceso ejecutado por las técnicas de limpieza de datos para tratar la señal de voltaje de la PMU AGOYÁN - BAÑOS 1, en esta sección se presenta finalmente el resultado obtenido de la señal junto con la señal original escogida en la sección 1, como se puede observar en la Figura 4.37.

Adicionalmente, se puede observar los valores de la raíz cuadrada del error medio cuadrático y el coeficiente de determinación, los cuales fueron calculados entre la señal final y la original, es decir, el error que se obtuvo durante todo el proceso de las técnicas de limpieza de datos.

Figura 4.37. Sección 6: Señal Preprocesada

En este caso al realizar la limpieza de datos de la señal del voltaje se observó que tanto desde el punto de vista estadístico y con respecto al fenómeno físico se presenta datos erróneos. Es decir, mediante el análisis de la estadística descriptiva se determina la existencia de datos atípicos en la señal de voltaje que se expone en la Figura 4.32., mientras que su interpretación física es la disminución y aumento de los valores de los datos de la señal de voltaje, el rango de los límites del voltaje es de ±5% de su valor nominal, por lo que el valor mínimo y el valor máximo de la sección de voltaje se encuentra cerca de los límites como se presenta en la Ecuación 4.1.

$$V_{minimo\ p.u.=}\frac{\sqrt{3}*81{,}093\ kV}{138\ kV}=1{,}02\ p.u.$$

$$V_{maximo\ p.u.=}\frac{\sqrt{3}*82{,}789\ kV}{138\ kV}=1{,}04\ p.u.$$

Ecuación 4.1. Ecuación de Voltaje en p.u.

Entonces mediante el uso de las técnicas de limpieza de datos se puede tratar estos los diferentes tipos de datos anómalos y así proceder al tratamiento adecuado de la señal de voltaje.

4.2.3. Caso 1: SEÑAL DEL ÁNGULO DEL VOLTAJE

4.2.3.1. Resultado de la sección 1: SELECCIONAR SEÑAL

Para el presente caso de estudio se escogió por señal al ángulo del voltaje de la PMU AGOYÁN - BAÑOS 1, como se puede observar en la Figura 4.38.

Figura 4.38. Sección 1: Selección de Señal

En la Figura 4.39., se obtiene una mejor apreciación de la señal temporal del ángulo de voltaje, la cual a simple vista presenta discontinuidades, estas observaciones serán tratadas en las secciones posteriores.

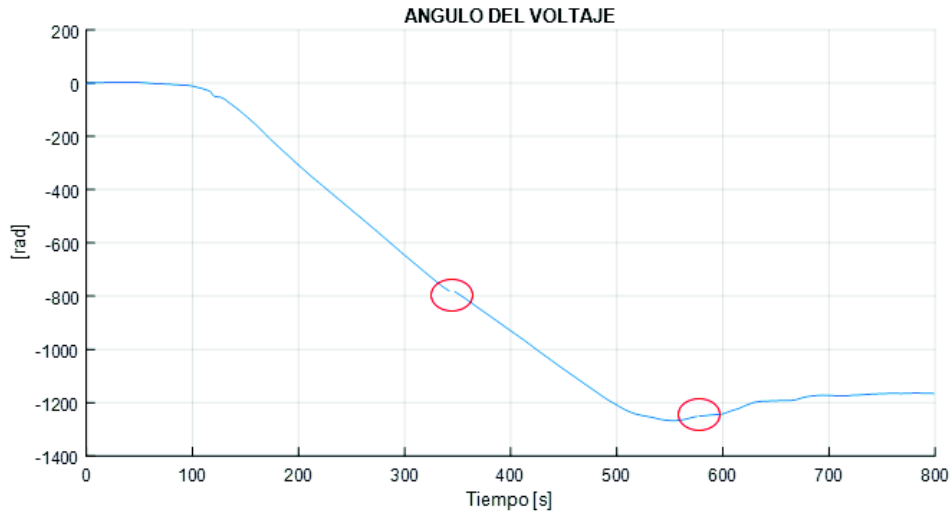


Figura 4.39. Señal del ángulo de voltaje de la PMU AGOYÁN - BAÑOS 1

4.2.3.2. Resultado de la sección 2: ESTADÍSTICA DESCRIPTIVA

En la Tabla 4.10. se observa los resultados de las medidas de tendencia central y de dispersión, las cuales describen y representan la señal del ángulo de voltaje.

Tabla 4.10. Medidas de tendencia central y de dispersión

Medidas de tendencia central y de dispersión	Valor [rad]
Media	-760,286
Mediana	-933,608
Moda	-1267,422
Varianza	223140,75
Desviación	472,377
Máximo	2,785
Mínimo	-1267,422
Rango	1270,207

Como se puede observar en la Figura 4.40. se expone la gráfica del ángulo de voltaje con sus respectivas medidas de tendencia central, el valor promedio y el valor de la mediana y valor de la moda, como se observa los valores se encuentran muy dispersos.

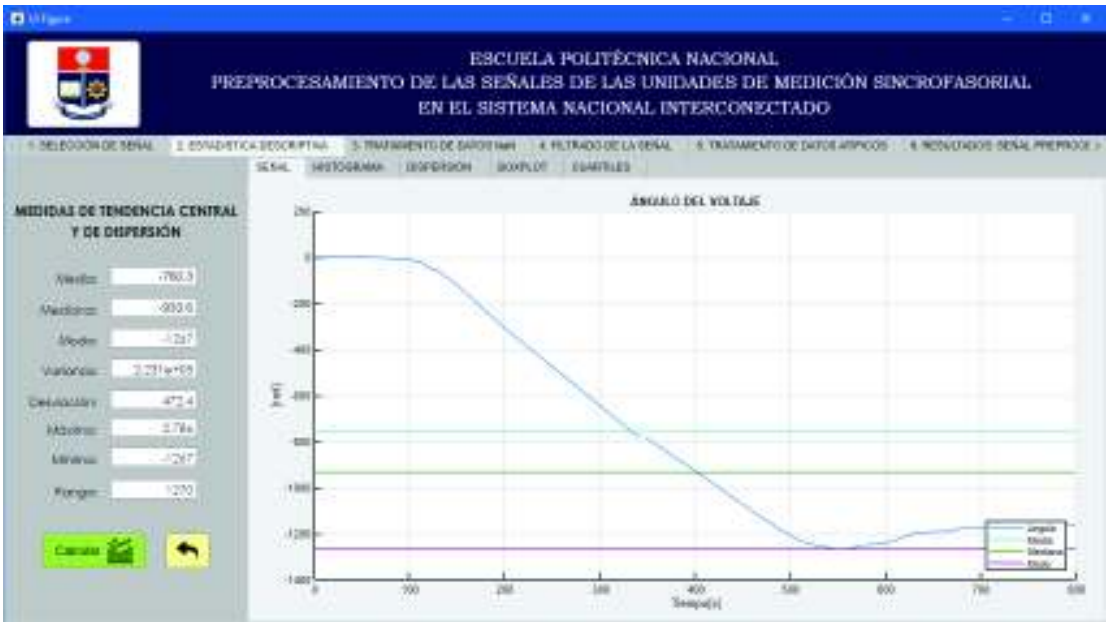


Figura 4.40. Sección 2: Estadística Descriptiva

En la Figura 4.41. se expone el resultado del histograma de la señal del ángulo del voltaje, en el cual se presenta la característica unimodal, a su vez se aprecia que la distribución es no simétrica y los valores más frecuentes comprenden el rango de -1200 a -1150 [rad].

Figura 4.41. Sección 2: Estadística Descriptiva

En la Figura 4.42. se presenta el gráfico de dispersión de la señal del ángulo del voltaje.

Figura 4.42. Sección 2: Estadística Descriptiva

La Figura 4.43. y 4.44. contienen el resultado de la gráfica BoxPlot y el gráfico cuantil de la señal del ángulo de voltaje correspondientemente, por lo cual en estas se presenta los percentiles equivalentes a Q_1 = -1175,115 [rad], Q_2 = -933,608 [rad] y Q_3 = -300,923 [rad]. El diagrama de caja se extiende entre la observación equivalente a -1267,42 [rad] y 2,785 [rad], en esta gráfica no se identificaron valores atípicos.

Figura 4.43. Sección 2: Estadística Descriptiva

Figura 4.44. Sección 2: Estadística Descriptiva

4.2.3.3. Resultado de la sección 3: TRATAMIENTO DE DATOS NaN

Para el tratamiento de datos NaN de la señal del ángulo de voltaje seleccionada, primero se examinó la existencia de datos NaN en el conjunto de valores de los datos, por lo que se obtuvo como resultado la presencia de estos, tal como se indica en la Tabla 4.11.

Tabla 4.11. Contenido de datos NaN en la señal del ángulo de voltaje

Cantidad de datos NaN	Intervalo de tiempo [s]
312	342 a 347,1833
107	575,7667 a 577,5333
Total: 419 datos NaN	

A continuación, se empleó cada uno de los métodos expuestos en esta sección, y de esta forma escoger el método adecuado para la señal del ángulo de voltaje obteniendo como resultado la Tabla 4.12., la cual presenta el cálculo de la raíz cuadrada del error medio cuadrático de los diferentes métodos empleados para el tratamiento de datos NaN de la señal. Como se puede apreciar el método Spline presenta como resultado un valor RMSE más bajo que los demás modelos y este resulta ser el óptimo para su presente aplicación.

Tabla 4.12. Cálculo de los errores de los métodos para el tratamiento de datos NaN en la señal del ángulo de voltaje

Metodología	RMSE
Lineal	0,0057
Spline	0,0018
Akima	0,002
Autorregresivo	0,1283
Media Móvil	9,29

El método de interpolación cúbica “Spline”, fue seleccionado para el tratamiento de los datos NaN de la señal de ángulo del voltaje, obteniendo como resultado la señal expuesta en la Figura 4.45.

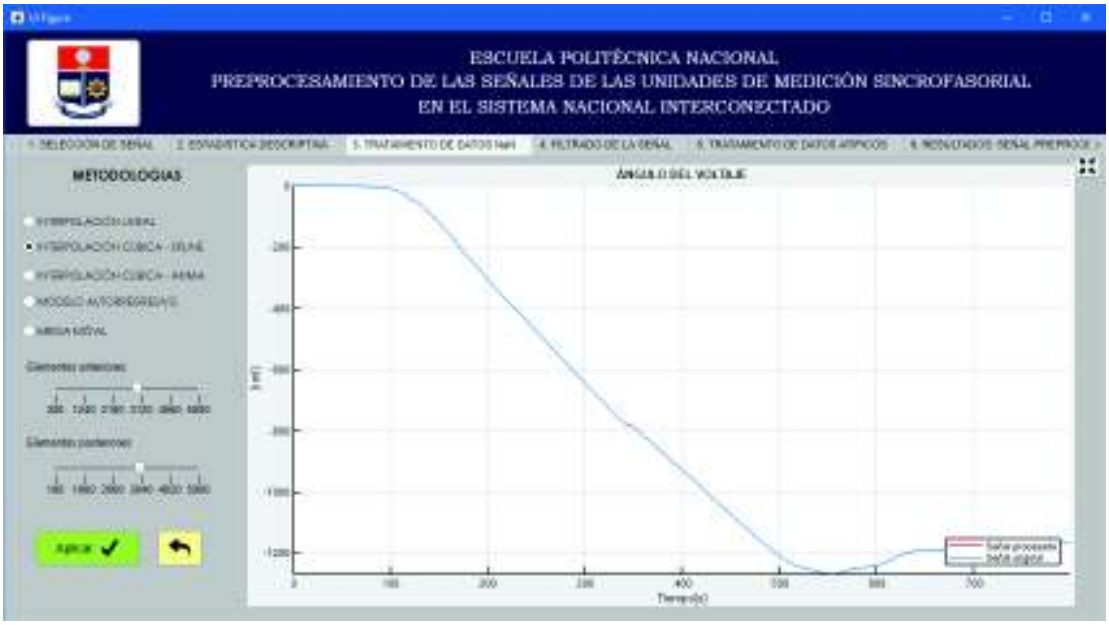


Figura 4.45. Sección 3: Tratamiento de datos NaN

4.2.3.4. Resultado de la sección 4: FILTRADO DE LA SEÑAL

Se utilizó cada uno de los métodos expuestos en esta sección obteniendo como resultado la Tabla 4.13., la cual presenta el cálculo de la raíz cuadrada del error medio cuadrático de los diferentes métodos empleados para el filtrado de la señal del ángulo de voltaje. Como se puede apreciar el método de Fourier junto con el filtro FIR presenta como resultado un valor RMSE más bajo en comparación a los demás modelos y este resulta ser el óptimo para su presente aplicación.

Tabla 4.13. Cálculo de los errores de los métodos para el filtrado de la señal del ángulo de voltaje

Metodología	RMSE
Fourier y filtro Yulewalk	0,0045
Fourier y filtro Butterworth	0,0002
Fourier y filtro FIR	0,0001
Wavelet	0,0004

Se efectuó la descomposición de Fourier junto con el filtro FIR en la señal del ángulo de voltaje, y de esta manera se adquirió la señal filtrada. En la Figura 4.46. se expone los resultados que se obtuvieron de este análisis.

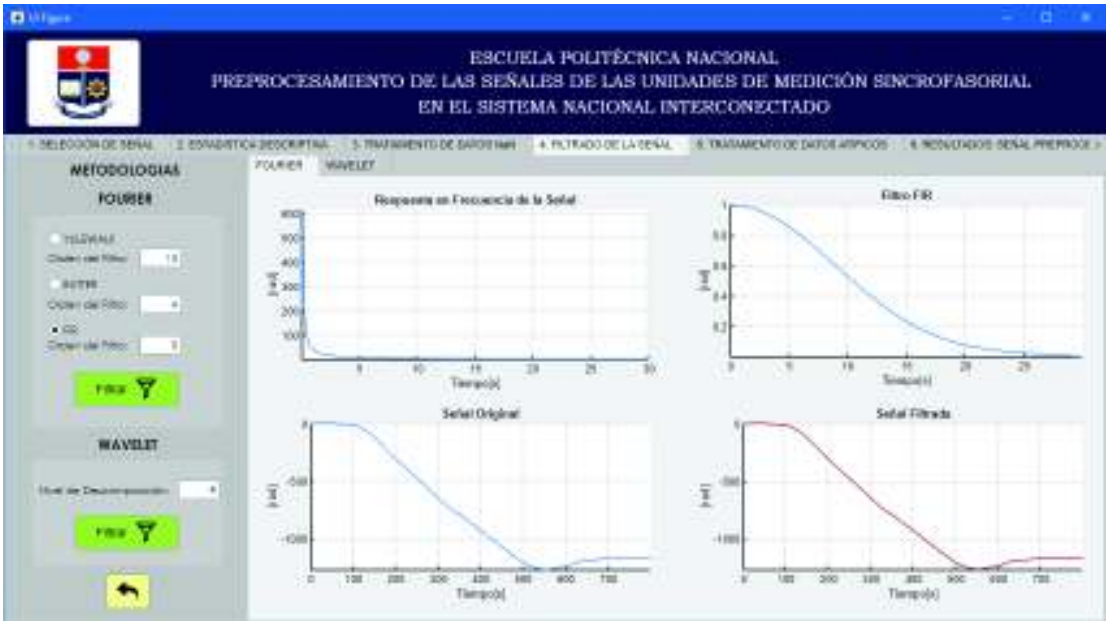


Figura 4.46. Sección 4: Filtrado de la señal

4.2.3.5. Resultado de la sección 5: TRATAMIENTO DE DATOS ATÍPICOS

En todos los resultados que se obtuvieron de los métodos usados para el tratamiento de los datos atípicos de la señal del ángulo de voltaje, se observó que en este caso la señal no presenta datos atípicos y lo que se procede a realizar es a suavizar. Para comparar estos métodos y así escoger el más adecuado, se tomó en cuenta a la raíz cuadrada del error medio cuadrático y el coeficiente de determinación. En la Tabla 4.14. se presenta el cálculo RMSE y R^2 de los diferentes métodos empleados. Como se puede observar el método LOESS presentan el valor más bajo de RMSE y su coeficiente de determinación es igual a 1, por lo que, en comparación a los demás modelos, este resulta ser el modelo óptimo para su presente aplicación.

Tabla 4.14. Cálculo de los errores de los métodos para el tratamiento de datos atípicos en la señal del ángulo de voltaje

Metodología	RMSE	R^2
LOWESS	0,039	1
LOESS	0,0044	1
SGOLAY	0,0095	1
K-MEANS	2,995	1

La Regresión cuadrática local “LOESS”, calcula y ajusta un polinomio cuadrático sobre la señal del ángulo de voltaje, mediante una ventana deslizante de 100 puntos de datos, Se expone el resultado en la Figura 4.47., la señal que se presenta en color rojo es la nueva función estimada.

Figura 4.47. Sección 5: Tratamiento de datos atípicos

4.2.3.6. Resultado de la sección 6: SEÑAL PREPROCESADA

Después de todo el proceso ejecutado por las técnicas de limpieza de datos para tratar la señal del ángulo de voltaje de la PMU AGOYÁN - BAÑOS 1, en esta sección se presenta finalmente el resultado obtenido de la señal junto con la señal original escogida en la sección 1, como se puede observar en la Figura 4.48.

Adicionalmente, se puede observar los valores de la raíz cuadrada del error medio cuadrático y el coeficiente de determinación, los cuales fueron calculados entre la señal final y la original.

Figura 4.48. Sección 6: Señal Preprocesada

Al efectuar la limpieza de datos de la señal del ángulo del voltaje, esta fue tratada desde un punto de vista como una señal temporal sin analizar el sentido físico que en realidad representa. Se efectuó todo el proceso para la obtención de una señal preprocesada lista para el análisis y mediante el estudio de la estabilidad de ángulo se procede a su debida interpretación, para entender esta se necesita saber la interpretación de la diferencia angular, para esto se debe analizarla con respecto a una señal de referencia, entonces se considera como referencia la señal del ángulo de voltaje de la PMU MOLINO - PASCUALES 1, de la cual también se procede a realizar la respectiva limpieza de datos, para poder obtener la diferencia angular con la señal del ángulo de voltaje de la PMU AGOYÁN - BAÑOS 1. A continuación, se observa en la Figura 4.49. la señal de referencia, es decir, la señal del ángulo de voltaje de la PMU MOLINO - PASCUALES 1, aplicada la limpieza de datos.

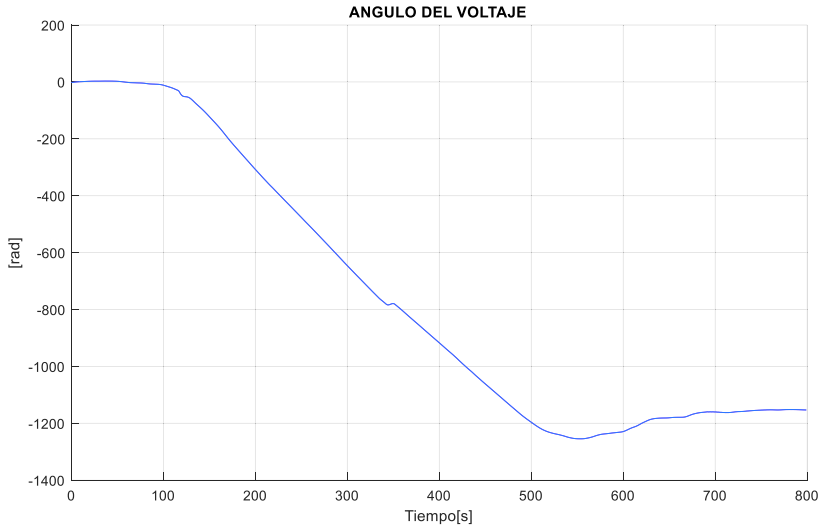


Figura 4.49. Señal preprocesada del ángulo de voltaje de la PMU MOLINO - PASCUALES 1

Ahora se procede a calcular la diferencia angular entre la referencia y la señal del ángulo de voltaje, el resultado de esta se puede observar en la Figura 4.50. La diferencia angular aumenta y disminuye en el tiempo, esto ocurre debido a que la demanda de igual forma sube o baja en el transcurso del tiempo. De tal modo que resulta importante realizar una limpieza adecuada de datos de las señales para el posterior estudio o análisis de las diferentes señales.

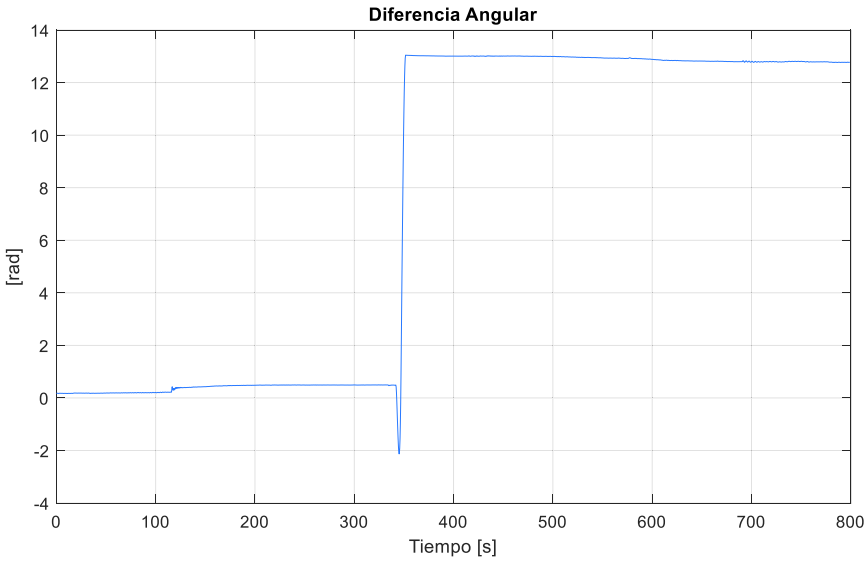


Figura 4.50. Diferencia angular entre la señal del ángulo de voltaje de la PMU MOLINO - PASCUALES 1 y la PMU AGOYÁN - BAÑOS 1

5. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

- Se realizó el preprocesamiento de los datos de las mediciones sincrofasoriales de las PMUs del Sistema Nacional Interconectado, mediante la aplicación de los enfoques que comprenden el proceso de las técnicas de limpieza de datos, tal que, se resolvió las inconsistencias que presentan cada una de las diferentes señales temporales.
- Los enfoques que comprenden las técnicas de limpieza de datos cumplieron una función específica a través de las diferentes metodologías propuestas, lo cual permitió el diagnóstico, detección e imputación de los datos anómalos de las diferentes mediciones sincrofasoriales. Por consiguiente, se influyó en la calidad de los datos, y así se obtuvo una base de datos fidedigna.
- Se desarrolló una aplicación que permite realizar la limpieza de datos de las mediciones sincrofasoriales de las PMUs del SNI. Esta aplicación se divide en etapas, las cuales tratan datos ruidosos, faltantes, atípicos e inconsistentes, de una forma interactiva para el usuario.
- Los datos faltantes NaN implican una reducción del tamaño del conjunto de datos de la señal y, en consecuencia, un aumento en el error de las estimaciones de los parámetros. Por lo que, provocan una falta de respuesta y dan lugar a estimaciones sesgadas, siendo necesario la imputación de estos datos mediante mecanismos que permiten estimar valores de reemplazo, los cuales retienen las características generales del conjunto de datos.
- Los métodos de imputación de los datos atípicos pueden ser altamente influenciados por observaciones extremas, y estos dependen específicamente de la longitud de la ventana. Al seleccionar una menor longitud de la ventana puede evitar la convergencia del algoritmo y ser más resistente a la presencia de valores atípicos, mientras que a mayor longitud de la ventana la curva ajustada puede ser muy suavizada y perder las características esenciales de la forma original de la señal.

- Una tendencia general que presentaron las señales al realizar la limpieza de datos de las mediciones sincrofasoriales fue la obtención de datos erróneos debido a la ocurrencia de una falla en el sistema eléctrico. Esto se debe al protocolo C37.118-2005 que utiliza la PMU, el cual no considera la dinámica del transitorio durante la ocurrencia de una contingencia, por tanto, se estima datos erróneos hasta que el sistema se estabilice. Entonces mediante el uso de las técnicas de limpieza de datos se puede tratar estos datos como atípicos y así proceder al tratamiento adecuado de las diferentes señales.
- La limpieza de datos es un problema que se encuentra en otros campos. Las técnicas desarrolladas en esta disertación pueden ampliarse a otros campos de los datos de series de tiempo y los modelos de imputación para el contexto particular están bien definidos, por los resultados obtenidos.
- Los resultados obtenidos demostraron que existe la necesidad de métodos de agrupamiento como el algoritmo K-means, para rastrear con precisión los cambios en los valores de las diferentes mediciones sincrofasoriales de la red eléctrica, y podrían detectar cualquier escenario de contingencia.
- La red eléctrica convencional ha experimentado una serie de desarrollos, por lo que se ha convertido en una red inteligente moderna. La visión para este desarrollo ha sido integrar una capa de procesamiento de datos, lo que permitirá que la red inteligente tenga una monitorización de la red durante todo el día.

5.2. Recomendaciones

- Previo al uso de la aplicación se recomienda ejecutar una sola vez el Script.m llamado “GuardarDatos”, el cual genera un archivo “pmu.mat” que contiene la base de datos de las mediciones sincrofasoriales de las PMUs del SNI, con esto el proceso al cargar los datos de las señales temporales en la aplicación se agiliza y optimiza.
- Es necesario continuar el orden indicado de las etapas expuestas en la aplicación, debido a que cada etapa se habilita sí su etapa previa fue ejecutada, por tanto, se recomienda seguir este orden del proceso de las técnicas de limpieza de datos para la obtención de los resultados finales.

- Para las diferentes metodologías que emplean el parámetro de la longitud de la ventana se recomienda seleccionar un valor adecuado, ya que de este depende la correcta convergencia del algoritmo para la obtención de óptimos resultados.
- En el método de agrupación basado en la distancia K-means, se recomienda una correcta elección del número de clústeres, ya que dependiendo de esto el algoritmo se vuelve más resistente o sensible a datos atípicos de la señal temporal.

6. REFERENCIAS BIBLIOGRÁFICAS

[1] L. Armas, “Estudio Técnico para la Determinación de la Ubicación Óptima de Unidades de Medición Fasorial (PMU) en el Sistema Nacional Interconectado S.N.I. Basado en Criterios de Observabilidad ante Contingencias”, Escuela Politécnica Nacional, Quito, 2016.

[2] D. Jiménez, “Ubicación Óptima de Unidades de Medición Sincrofasoriales PMU’s para Mejorar la Observabilidad de un Sistema Eléctrico de Potencia”, Escuela Politécnica Nacional, Quito, 2015.

[3] M. Campos, M. Arias, “Ubicación Óptima de Unidades de Medición Fasorial aplicando Swarm Intelligence”, IEEE, 2010.

[4] V. Raju, “An Optimal PMU Placement method for Power System Observability”, Gokaraju Rangaraju Institute of Engineering &Technology, 2016.

[5] J. Ekanayake, K. Liyanage, J. Wu, A. Yokoyama, N. Jenkins, “Smart Grid : Technology and Applications.”, Editorial Wiley, pp. 173–186, 2012.

[6] A. Monti, C. Muscas, F. Ponci, “Phasor Measurement Units and Wide Area Monitoring Systems: From the Sensors to the System”, Editorial ELSEVIER, pp. 2–100, 2016.

[7] J. Momoh, James Momoh, “Smart Grid Fundamentals of Design and Analysis”, John Wiley & Sons, pp. 20–215, 2012.

[8] S. R. Bhide, “Digital power system protection”, PHI, pp. 216–236, 2014.

[9] A. G. Phadke, J. S. Thorp, “Power Electronics and Power Systems Synchronized Phasor Measurements and Their Applications Second Edition”, vol. 2, Springer, pp. 84–220, 2017.

[10] N. Granda, “Esquema Adaptable de Separación Controlada en Islas para Sistemas Eléctricos”, Universidad Nacional de San Juan, Argentina.

[11] P. Mallet, J. A. Frazao, “PMU based situation awareness for smart distribution grids”, Octubre, Grenoble Alpes University, 2015.

[12] A. de la Torre, “Análisis Técnico para la Implementación de un Sistema de Monitoreo de Área Extendida (WAMS) en el Sistema Nacional Interconectado del Ecuador”,

Septiembre, Universidad Politécnica Salesiana, 2013.

[13] P. M. Anderson, “Power system protection”, McGraw-Hill, pp. 548–551, 2007.

[14] S. Ramírez, “Protección de Sistemas Eléctricos”, Universidad Nacional de Colombia, 2012.

[15] K. Chandragupta, V. Ramkumar, “Phasor Measurement Units in Power System Networks”, India, College of Technology, pp. 120–125, 2014.

[16] IEEE, “Std C37.244-2013 - IEEE Guide for Phasor Data Concentrator Requirements for Power System Protection , Control , and Monitoring”, Mayo, 2013.

[17] IEEE, “Dynamic Vulnerability Assessment and Intelligent Control for Sustainable Power Systems”, Wiley & Sons, pp. 9–147, 2002.

[18] J. Cepeda, P. Verdugo, G. Argüello, “Monitoreo de la Estabilidad de Voltaje de Corredores de Transmisión en Tiempo Real a partir de Mediciones Sincrofasoriales”, Revista EPN, 2014.

[19] J. Báez, S. Ordóñez, J. Cepeda, P. Verdugo, F. Quilumba, “Definición de Estudios Eléctricos para Determinar la Ubicación de Unidades de Medición Sincrofasorial en Sistemas Eléctricos de Potencia”, Revista Energía, 2017.

[20] N. Zhou, D. Meng, Z. Huang, G. Welch, “Dynamic state estimation of a synchronous machine using PMU data: A comparative study”, 2015.

[21] A. Mukherjee, “Clustering and Control of Streaming Synchrophasor Datasets”, December, University of North Dakota, 2015.

[22] A. De La Torre, J. Cepeda, “Implementación de un sistema de monitoreo de área extendida WAMS en el Sistema Nacional Interconectado del Ecuador SNI”, Revista INGENIUS, 2013.

[23] J. Cepeda, D. Echeverría, G. Argüello, “Cenace’s experiences on implementing a wide area monitoring system (WAMS) in the Ecuadorian power system”, Revista Energía, 2014.

[24] R. P. Cody, “Cody’s Data Cleaning Techniques”, vol. 2, Editorial SAS, p. 272, 2008.

[25] H. N. Akouemo, “Data Cleaning in the Energy Domain”, Mayo, Marquette University, 2015.

[26] J. Han, M. Kamber, J. Pei, “Data Mining Concepts and Techniques”, vol. 3, Editorial

ELSEVIER, p. 673, 2011.

[27] T. D. Johnson Theodore, “Exploratory Data Mining and Data Cleaning”, vol. 39, Wiley & Sons, pp. 17–200, 2008.

[28] C. C. Aggarwal, “Outlier analysis”, Springer, pp. 1–446, 2013.

[29] A. Charalambides, M. Koutras, “Probability and Statistics Models with Application”, Lóndres, p. 609, 2000.

[30] D. Swagatam, A. Ajith, K. Amit, “Metaheuristic Clustering”, vol. 178, Springer, p. 249.

[31] M. Mohri, “Fundations of Machine Learning”, Cambridge Massachusetts, pp. 147–350, 2012.

[32] J. Freund, J. Wilson, “Statistical Methods”, vol. 2, Academic Press, pp. 1–75, 1966.

[33] T. De Waal, J. Pannekoek, S. Scholtus, S. Netherlands, “Survey Methodology Statistical Data Editing and Imputation”, Wiley & Sons, pp. 1–57, 2011.

[34] B. R. J., “Statistics, A Guide to the Use of Statistical Methods in the Physical Sciences”, Manchester University, pp. 1–20, 1989.

[35] R. Nisbet, J. Elder, G. Miner, “Handbook of Statistical Analysis and Data Mining Applications”, Editorial ELSEVIER, p. 823, 2007.

[36] C. W. Kang, P. H. Kvam, “Basic Statistical Tools for Improving Quality”, Wiley & Sons, pp. 29–49, 2012.

[37] D. Montgomery, C. Jennings, M. Kulahci, “Introduction to Time Series Analysis and Forecasting”, Wiley &. Sons, pp. 18–139, 2008.

[38] M. H. Trauth, “MATLAB® recipes for earth sciences”, vol. 4, Springer, p. 427, 2015.

[39] B. Moler, “Numerical Computing with MATLAB”, SIAM, pp. 93–110, 2007.

[40] T. Mitsa, “Temporal Data Mining”, Taylor & Francis, p. 398, 2010.

[41] D. S. Chi Fung, “Methods for the Estimation of Missing Values in Time Series”, Cowan University, p. 332, 2006.

[42] B. R. Forest, B. Menke, “Environmental Data Analysis with MATLAB”, vol. 2, Editorial ELSEVIER, pp. 1–20, 2012.

[43] J. R. Hauser, “Numerical methods for nonlinear engineering models”, Springer, pp. 1–986, 2009.

[44] H. Anton, C. Rorres, “Elementary Linear Algebra”, vol. 10, Wiley & Sons, pp. 956–995.

[45] U. Kumar, A. Ahmadi, A. K. Verma, P. Varde, “Current Trends in Reliability, Availability, Maintainability and Safety”, Springer, pp. 335–350, 1969.

[46] D. Salomon, “The Computer Graphics Manual”, vol. 1, Springer, pp. 578–600.

[47] A. M. O. Mohamed, “Arid Land Hydrogeology: In Search of a Solution to a Threatened Resource”, Taylor & Francis, pp. 97–100, 2006.

[48] R. Hyndman, A. George, “Forecasting : principles and practice”, Abril, Monash University, 2014.

[49] J. L. Gutierrez, A. 1. Schlögl, “Comparación de métodos autorregresivos para la detección de artefactos en señales de EEG”, University of Technology Graz, Austria, 2010.

[50] C. T. Leondes, “Computer Techniques and Algorithms in Digital Signal Processing”, vol. 75, Academic Press, pp. 1–80, 1987.

[51] B. Badiru, L. Racz, “Handbook of Measurements”, CRC Press, pp. 50–55, 2016.

[52] S. Sriram, S. Nitin, K. Prabhu, M. Bastiaans, “Signal denoising techniques for partial discharge measurements”, IEEE, 2005.

[53] B. Porat, B. Friedlander, “The Modified Yule-Walker Method of ARMA Spectral Estimation”, IEEE, 1983.

[54] H. G. Dimopoulos, “Analog Electronic Filters”, vol. 1, Springer, pp. 1–30, 2012.

[55] D. Schlichthärle, “Digital Filters: Basics and Design”, vol. 2, Springer, pp. 127–140, 2001.

[56] S. Burrus, A. Ramesh Gopinath, H. Guo, “Introduction to Wavelets and Wavelet Transforms”, Prentice-Hall, p. 266, 1998.

[57] C. Drive, “A Wavelet Tour of Signal Processing”, vol. 3, Editorial ELSEVIER, pp. 102–1, 2009.

[58] E. R. Cerda, J. C. J. Correa, O. G. Brambila, “Aplicación de la Transformada Wavelet en la Detección de Defectos Causados por Vibrado en Piezas Cilíndricas Rectificadas”, Sociedad Mexicana de Ingeniería Mecánica, 2006.

[59] M. Gupta, J. Gao, C. Aggarwal, J. Han, “Outlier detection for temporal data”, Morgan

& Claypooll, p. 109, 2014.

[60] R. Baxter, N. Hastings, A. Law, E. J. . Glass, “Linear Regression Analysis”, Wiley & Sons, pp. 162–170, 2008.

[61] B. S. Everitt, “Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences”, Taylor & Francis, pp. 73–80, 2009.

[62] A. W. Bowman, A. Azzalini, “Applied Smoothing Techniques for Data Analysis”, Clarendon Press, pp. 75–91, 1997.

[63] K. Takezawa, “Introduction to Nonparametric Regression”, Wiley & Sons, pp. 231–319, 2006.

[64] J. Wu, “Advances in K-means Clustering: a data mining thinking”, Springer, p. 177, 2012.

[65] M. E. Celebi, “Partitional Clustering Algorithms”, Springer, pp. 42–100, 2015.

[66] S. Lukasik, P. A. Kowalski, M. Charytanowicz, P. Kulczycki, “Clustering using flower pollination algorithm and Calinski-Harabasz”, IEEE, 2016.

[67] J. M. Wooldridge, “Introducción a la econometría”, vol. 4, CENGAGE, pp. 80–83, 2009.

[68] The MathWorks Inc., “Creating Graphical User Interfaces”, pp. 16–20, 2014.

7. ANEXOS

ANEXO I. Manual de usuario de la aplicación implementada.

ANEXO II. Código final implementado en Matlab se anexa en CD.

ANEXO I

Manual de usuario de la aplicación “ Preprocesamiento de las señales de las unidades de medición sincrofasorial”.

Paso previo a la ejecución de la aplicación:

1. Ejecutar el Script.m llamado “GuardarDatos”

Pasos que seguir para simular la señal preprocesada:

1. Ingresar en la pestaña “SELECCIÓN DE SEÑAL”, y seleccionar la señal para realizar limpieza de datos.
2. Ejecutar la pestaña “ESTADÍSTICA DESCRIPTIVA”, la cual proporciona información descriptiva del conjunto de datos de la señal.
3. Continuar con la pestaña “TRATAMIENTO DE DATOS NaN”, y seleccionar cualquiera de las metodologías expuestas para la obtención de resultados.
4. Luego en la pestaña “TRATAMIENTO DE DATOS ATÍPICOS”, seleccionar la respectiva metodología para tratar los valores atípicos de la señal.
5. Finalmente, en la pestaña “RESULTADOS: SEÑAL PREPROCESADA ”, se obtiene el resultado final de la señal preprocesada mediante la técnica de limpieza de datos.

ORDEN DE EMPASTADO