UNIVERSIDAD MICHOACANA DE SAN NICOLAS DE HIDALGO

FACULTAD DE INGENIERÍA ELÉCTRICA

“EFECTO DE LA PANDEMIA SOBRE LA EFICIENCIA DE LOS PREDICTORES”

TESIS

QUE PARA OBTENER EL TITULO DE:

Ingeniero en Computación

PRESENTA:

Jesús Salvador Macedonio Carbajal

ASESOR:

Dr. Jaime Cerda Jacobo

Morelia Michoacán noviembre de 2022

Agradecimientos

Gracias a la Facultad de Ingeniería eléctrica por el apoyo brindado a lo largo de la carrera y por haberme preparado para el camino laboral.

Al Dr. Jaime Cerda Jacobo, por su asesoramiento y guía durante el proceso de realización de este proyecto.

A mis amigos con los que he compartido buenos momentos y que me han ayudado a olvidar los momentos de estrés.

A mis padres y familiares que siempre me han apoyado y alentado a salir adelante.

Dedicatorias

A mis padres, por su apoyo incondicional, por darme la oportunidad que estudiar y llegar a este punto en mi vida, por sus consejos y sus lecciones, las cuales han sido de gran importancia para mi formación como la persona que soy.

Al resto de mi familia, que siempre han estuvieron al pendiente y brindándome ánimo para poder concluir mis estudios, han sido un factor de suma importancia para motivarme y sentirme respaldado a lo largo de toda mi vida.

Índice general

Resumen

Como se sabe, el estallido de la pandemia trajo consigo una serie de regulaciones y restricciones que conllevaron una modificación en el estilo de vida cotidiano que se tenía con anterioridad, teniendo como consecuencia un cambio en los patrones y niveles de consumo de electricidad. De esta forma, los sistemas de predicción que utilizan datos históricos como principal base para realizar pronósticos, se vieron afectados por el nuevo comportamiento de los datos, provocando un impacto en la precisión de sus pronósticos.

En base a lo anterior mencionado, en esta tesis se presenta una investigación con el objetivo de mostrar el efecto de este nuevo comportamiento en la eficiencia de los modelos de aprendizaje automático usados para el pronóstico de la demanda energética.

Para esto, se analizó el conjunto de datos de las regiones que conforman el sistema interconectado nacional mexicano, es decir, las regiones: Centro, Noreste, Noroeste, Norte, Occidente, Oriente y Peninsular, donde, los efectos de la pandemia mostraron un impacto diferente sobre cada región.

Por otra parte, se realizaron predicciones de la demanda energética para los años 2019, 2020 y 2021 utilizando los algoritmos: Random Forest, XGBoost, Regresión vectorial de soporte (SVR), K vecinos más cercanos (KNN) y un ensamble de estos, mostrando como resultado una mejoría en la eficiencia de los predictores durante los años de pandemia, ya que, en el análisis de resultados, se observó que los pronósticos de los modelos entrenados con cada uno de estos algoritmos tuvieron una mayor precisión durante los años 2020 y 2021.

Palabras clave— Aprendizaje automático, predicción, demanda energética, análisis estadístico, pandemia

Abstract

As is known, the outbreak of the pandemic brought with it a series of regulations and restrictions that led to a modification in the daily lifestyle that was had previously, resulting in a change in the patterns and levels of electricity consumption. In this way, forecasting systems that use historical data as the main basis for making forecasts were affected by the new behavior of the data, causing an impact on the accuracy of their forecasts.

Based on the aforementioned, in this thesis an investigation is presented with the objective of showing the effect of this new behavior on the efficiency of the machine learning models used for forecasting energy demand.

For this, the data set of the regions that make up the Mexican national interconnected system was analyzed, that is, the regions: Center, Northeast, Northwest, North, West, East and Peninsular, where the effects of the pandemic demonstrate an impact different about each region.

On the other hand, energy demand predictions were made for the years 2019, 2020 and 2021 using the algorithms: Random Forest, XGBoost, Support Vector Regression (SVR), k nearest neighbor (KNN) and an assembly of these, showing as a result an improvement in the efficiency of predictors during the pandemic years, since, in the analysis of results, it is revealed that the forecasts of the models trained with each of these algorithms had a greater precision during the years 2020 and 2021.

Keywords— Machine learning, prediction, energy demand, statistical analysis, pandemic.

Índice de figuras

Índice de cuadros

Glosario de términos

Bias	Incapacidad de un método de machine learning para capturar la relación de los datos
Estacionalidad	Comportamiento o patrón que a veces se observa en las series de tiempo. Consiste en subidas o bajadas periódicas que se presentan en forma regular en la serie de tiempo
Hiperparametro	Valores de las configuraciones utilizadas durante el proceso de entrenamiento, las cuales son determinadas por el usuario
Kurtosis	Grado de concentración de los datos alrededor de su media
Parámetro	Variable que se estiman durante el proceso de entrenamiento de los modelos de ML, y que no son determinas por el usuario
Residual	Diferencia entre una predicción y el valor real
Skewness	Medida de la simetría de una distribución de datos
Tendencia	Dirección a la cual se mueven los datos
Varianza	Cambios en la eficiencia el modelo cuando se utilizan diferentes partes de un mismo conjunto de datos para el entrenamiento
Volatilidad	Variables que se estiman durante el proceso de entrenamiento de los modelos de ML, y que no son determinas por el usuario

Capítulo 1

Introducción

El ser humano siempre ha querido conocer el futuro, con el objetivo de anticipar los eventos y así poder tomar medidas que le otorguen el mayor beneficio posible. En el transcurso del tiempo se han creado diversas formas de predicción, que van desde magos y profetas, hasta técnicas más sofisticadas como la estadística y el Machine Learning, siendo este último, uno de los más populares y con mayor auge gracias a la gran capacidad de procesamiento de las computadoras y al florecimiento del Big Data.

El uso de técnicas de machine learning para la predicción requiere de la implementación de algoritmos cuyo funcionamiento se basa en el análisis de datos históricos los cuales deberán estar relacionados entre sí, esto significa que el conjunto de datos deberá describir algún fenómeno, por ejemplo: Ventas de un cierto producto, Temperatura de alguna región, Precio de alguna moneda, etc. Estos algoritmos tienen como fin la construcción de un modelo que ayude a predecir el comportamiento de los datos.

Sin embargo, la pandemia originada en el año 2020 trajo consigo una serie de cambios en la sociedad y su comportamiento. La declaración de cuarentena, así como el cierre de diversas instituciones, centros comerciales y sitios de trabajo, modifico en gran medida nuestro estilo de vida, y por consecuencia, el comportamiento y relación de los datos que describen ciertos aspectos de ella.

Estos cambios de comportamiento modificaron de manera directa a la demanda energética que se tenía con respecto a años anteriores a la pandemia, teniendo como resultado una modificación en sus patrones de conducta, lo cual tuvo un impacto en la eficiencia con la que se realizaban sus predicciones.

Entonces, esta tesis se centra en evidenciar el efecto que tuvo la pandemia sobre las técnicas de machine learning utilizadas para la predicción y la manera en que el comportamiento de los datos afecto su eficiencia. Para ello, se toma como ejemplo la predicción de la demanda energética en México, donde se realizan comparaciones estadísticas de los datos, y de eficiencia de los predictores en periodos de tiempo anteriores y durante la pandemia.

1.1. Antecedentes

El origen del análisis predictivo se remonta al menos a la estadística moderna a finales del siglo XIX, pero no fue hasta la década de 1950 que varias organizaciones comenzaron a usar modelos basados en computadora para anticipar todo, desde patrones climáticos hasta riesgos crediticios. En la década de 1970, se desarrolló el famoso modelo Black Scholes para predecir los mejores precios de las opciones sobre acciones, y en la década de 1990, la analítica se utilizó ampliamente para todo, incluyendo búsquedas en la web. Este sería el año en el que el campo del Machine Learning comenzaría a florecer, haciéndose presente en muchas de las disciplinas basadas en datos, como lo son; La biología, Marketing, Economía y hasta la justicia penal.

A partir de entonces, empezarían a surgir diversas técnicas que pretendían comprender las relaciones de los datos. Pronto, estas técnicas se empezaron a diferenciar de acuerdo a la forma en que reciben y analizan los datos, generando así, las siguientes categorías: Aprendizaje Supervisado, Aprendizaje no supervisado, Aprendizaje por refuerzo. A su vez cada categoría se subdividiría de acuerdo al tipo de problema que se estuviera intentando resolver, dando como resultado las siguientes clasificaciones de los algoritmos: Regresión, Clasificación, Agrupación (Clustering) y Reducción Dimensional. Estas divisiones y clasificaciones se ven representadas en la Fig.1.1

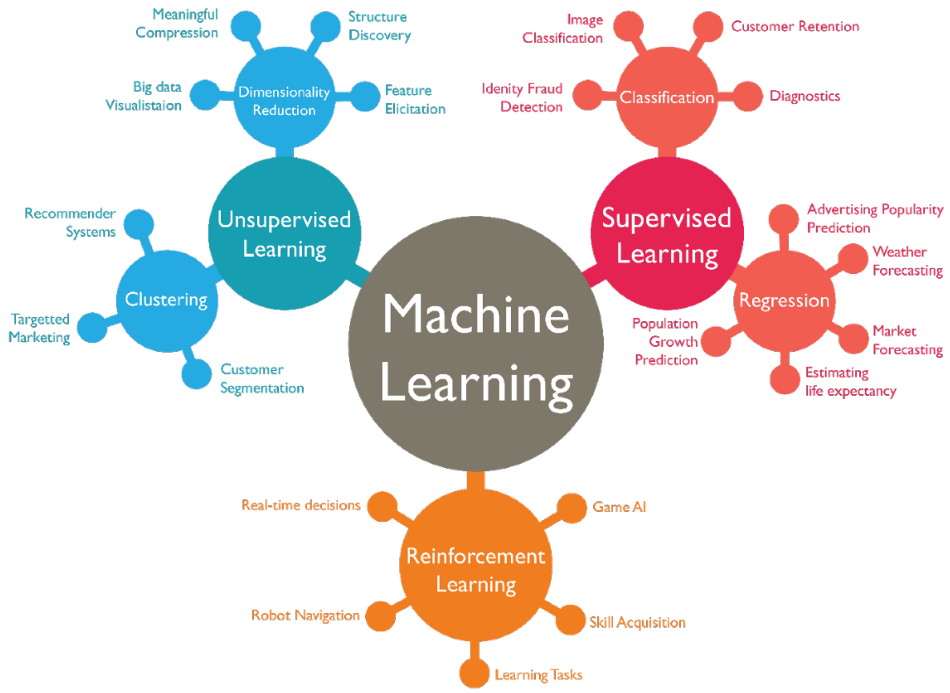


Figura 1.1: Clasificación de las técnicas y algoritmos de Machine Learning [1]

Las técnicas de aprendizaje supervisado nos permiten establecer una relación entre una variable dependiente (Target) y una o varias variables independientes (características). Estas técnicas se utilizan principalmente para la predicción, el modelado de series de tiempo y la determinación de la relación causa-efecto entre variables.

Efecto de la pandemia sobre los datos y su comportamiento

El artículo escrito por Morales et al. (2021)[2]. muestra la problemática que trajo la pandemia respecto a la operación estadística y la recopilación de datos que se realiza en los países alrededor del mundo. Se menciona que la creciente demanda por información para el monitoreo de la propagación del virus y los desafíos que la pandemia trajo consigo interrumpió las operaciones regulares de recolección de datos de muchas oficinas nacionales, regionales y mundiales, lo que provoco el cese de sistemas estadísticos donde la información solía ser recopilada de manera presencial, tales como los censos y encuestas. Esto obligo a las instituciones a buscar alternativas tales como recolección de datos vía telefónica y a través de la web, sin embargo, esto requiere de una modernización en la infraestructura de recopilación, procesamiento y difusión de los datos, así como la capacitación del personal que trabaja en esta área, lo cual es un problema para aquellos países con ingresos bajos, debido a que suelen tener limitantes en su equipo e infraestructura tecnológica. Sin embargo, este cambio traerá consigo una serie de innovaciones que favorecerán a la colecta de datos y a la comunidad estadística.

En el artículo Eftimov et al. (2020)[3]. se realiza una investigación sobre el cambio en los patrones de consumo de alimentos tomando como punto de referencia el periodo de declaración de cuarentena. La investigación se realizó analizando dos conjuntos de datos que consisten en 69,444 y 10,009 recetas de cocina pertenecientes a 24 países publicadas antes y durante el periodo de cuarentena respectivamente. El análisis de estos se llevó a cabo utilizando una metodología de Inteligencia Artificial, la cual analiza el texto de cada receta y proporciona una lista de etiquetas de los ingredientes principales que la conforman. Los resultados obtenidos fueron separados en dos conjuntos, uno donde se grafica la frecuencia de las etiquetas de los ingredientes específicos (queso, mantequilla, leche) y otro donde se grafica la frecuencia de las etiquetas de los grupos alimenticios a los que pertenecen los ingredientes (p.ej. lácteos, frutas, aceites, etc.). Los resultados arrojan un aumento del 300 % en el consumo de legumbres, 280 % en el consumo de Pancake / Tortilla / Oatcake, 180 % en sopa/potaje, 100 % en Pastelería/ Frutas con hueso y carne de res. Por otro lado tenemos a los alimentos que tuvieron un decremento en su consumo: -50 % en peces, -40 % en cereales/ maíz/ grano, -30 % en vino, -10 % en cítricos, cerdo, agua, mariscos, jugo de fruta / calabaza y hierbas.

Carvalho et al. (2021)[4]. analizó el efecto de las medidas de distanciamiento físico (ocasionadas por la pandemia) sobre las tendencias del consumo de energía eléctrica del sistema interconectado nacional brasileño, comprendiendo el periodo del 1 de enero al 27 de mayo del 2020. El estudio realiza las comparaciones de la demanda energética sobre un periodo anterior al inicio de los decretos de aislamiento (1 enero al 14 de marzo) y un periodo posterior a estos (a partir del 15 marzo), tomando los datos de la demanda energética de las regiones: Norte, Noreste, Sur y Sureste-Medio Oeste, las cuales conforman el sistema interconectado nacional. Realizando graficas que describen el comportamiento semanal de la demanda energética para cada región, se muestra en ellas disminuciones de entre el -7 y -20 % dependiendo la región, esta disminución comienza a partir de la semana 11 en la cual se decretó el aislamiento, por lo que queda evidenciado efecto de las restricciones de distanciamiento sobre la demanda energética.

Aplicaciones de técnicas de aprendizaje supervisado para la solución de problemas de predicción durante la pandemia.

Gulati et al.(2021)[5]. se centra en la predicción de la carga eléctrica del estado de Haryana (India) tomando en consideración la reducción de la demanda energética debido a las regulaciones gubernamentales impuestas el 20 de marzo del año 2020 a causa de la pandemia. El objetivo principal de este trabajo es conseguir un modelo de machine learning que prediga la carga eléctrica máxima y mínima una semana por delante, para un conjunto de 7 ciudades dentro del estado de Haryana. La metodología seguida es la implementación de cinco algoritmos de predicción (ANN, Linear Regression, SVR, Decision Tree Regressor y Random Forest) entrenados con datos normalizados pertenecientes a los primeros 4 meses del año 2020 y puestos a prueba, con tal de comparar su eficiencia. Los resultados de la eficiencia de cada algoritmo entrenado fueron medidos usando como referencia el error cuadrático medio entre las predicciones realizadas y los datos de la demanda real, teniendo como resultado una tabla donde se reflejan los errores de cada algoritmo para cada una de las 7 ciudades. Las conclusiones arrojan que el modelo entrenado con el algoritmo ANN demostró hacer mejores predicciones en comparación con los demás, sin embargo, no se descarta la posibilidad de que los resultados cambien si se realiza el mismo análisis en un periodo de tiempo distinto.

Tiwari et al (2021)[6]. presenta un análisis basado en técnicas de aprendizaje automático con el objetivo de crear un modelo que prediga la tendencia mundial de casos confirmados de covid-19. Para ello, se implementaron tres algoritmos de predicción (Naïve bayes, SVM y Regresión Lineal), los cuales fueron entrenados con un conjunto de datos de series de tiempo recopilados del 22 de enero del 2020 al 19 de mayo del 2020, donde algunas de las características clave incluían: Fecha de recopilación, Estado, Región, casos confirmados, casos recuperados y casos de defunción. Una vez entrenados los modelos, fueron puestos a prueba con el fin de comparar su eficiencia utilizando las métricas de error absoluto medio (MAE) y error cuadrático medio (MSE), dando como resultado a Naïve Bayes como la técnica que menor error tuvo al momento de hacer las predicciones, y por lo tanto, mostrándola como la que mejor predice la tendencia de casos confirmados respecto a SVM y Regresión Lineal.

Aljameel et al (2021)[7]. proporciona un método predicción para la identificación temprana de la gravedad de los pacientes enfermos con Covid-19, con la finalidad de que se tomen medidas de precaución para reducir la tasa de mortalidad. La metodología seguida en el desarrollo de la investigación consistió en un preprocesamiento de los datos previo, con el fin de extraer las características más importantes que describen a los pacientes que sobrevivieron, de los fallecidos por covid-19, posteriormente los datos resultantes fueron analizados utilizando tres algoritmos de clasificación (Regresión Logistica, Random Forest y XGBoost), los cuales fueron entrenados utilizando distintos grupos de características (25, 20 15 y 10) y diversos conjuntos de hiperparametros, a fin de optimizar los modelos producidos. Los resultados obtenidos muestran la superioridad del algoritmo Random Forest sobre los demás, teniendo una precisión del 95.2 % respecto a la Regresion Logistica y a XGBosst, los cuales obtuvieron una precisión máxima del 86.3 % y 93.2 % respectivamente.

1.2. Objetivo General

Observar el comportamiento de los datos históricos de la demanda energética debido la pandemia, y mostrar el efecto que esto generó sobre los predictores estadísticos usados para su pronóstico.

1.3. Objetivos Particulares

- Analizar y comparar el comportamiento de la demanda energética a través de métodos estadísticos en los diferentes periodos de tiempo recopilados.
- Mediante una ingeniería de características, identificar las principales características que pueden ser usadas para la predicción de la demanda energética
- Implementar los sistemas de predicción utilizando los algoritmos: XGboost, Random Forest, Regresión vectorial de soporte (SVR), K vecinos más cercanos (KNN) y un ensamble de los tres anteriores.
- Realizar pruebas de predicción en los años 2019, 2020 y 2021, con el fin de comparar la eficiencia de los predictores antes y durante la pandemia
- Analizar los resultados y obtener conclusiones

1.4. Justificación

Las técnicas de aprendizaje supervisado usadas para la predicción, entrenan modelos usando algoritmos, a los cuales se le otorgan variables de entrada (X), denominadas como características, y variables de salida (Y), denominadas como etiquetas o target. Esto se hace con la finalidad de que el algoritmo aprenda una función de mapeo entre cada entrada y salida Y=f(X).

El objetivo es aproximar la función de mapeo Y=f(X), llamada modelo, de tal manera que cuando se obtengan nuevos datos de entrada (X) pueda predecir las variables de salida (Y) para estos datos.

El pronóstico para las series de tiempo utiliza técnicas de machine learning supervisado, cuyo funcionamiento concuerda con el descrito anteriormente, sin embargo, este tipo de problemas tiene la característica de que tanto las variables de entrada (X), como las variables de salida (Y), son datos históricos con marcas en el tiempo, los cuales describen patrones que tienden a repetirse cada cierto periodo de tiempo. Esto quiere decir que para poder entrenar un modelo que prediga correctamente un periodo, los datos de entrenamiento deberán cubrir por lo menos un ciclo de repetición, de esta manera el modelo resultante podrá estar preparado para predecir los resultados del siguiente ciclo.

Uno de los tantos problemas que genero la pandemia está relacionado con el comportamiento periódico de las series de tiempo, ya que debido a la llegada del covid-19 entramos a una "nueva normalidad", en donde el comportamiento de los datos que describen ciertos fenómenos sociales se vieron alterados, modificando de esta manera sus patrones y conducta.

La finalidad de este trabajo de tesis es evidenciar como la llegada de la pandemia afectó el comportamiento de los datos que describen de la demanda energética en México, y la manera en que esto repercutió en la eficiencia de las técnicas de machine learning usadas para su predicción.

El hecho de utilizar el sector energético como ejemplo es debido a que la predicción de la demanda eléctrica además de haber sido afectada por la pandemia también es de gran importancia, porque su pronóstico ayuda a planificar las estrategias a seguir en el proceso de creación energética, lo cual se traduce en un mayor ahorro monetario además de evitar la contaminación necesaria al medio ambiente, ya que, si se sabe con antelación la cantidad de energía eléctrica que se va a consumir, es posible preparar los materiales necesarios para su generación y así tener un desperdicio mínimo.

Al comprar la eficiencia de los predictores frente a conjuntos de datos que describen el mismo fenómeno, pero tienen con comportamientos diferentes, nos permite analizar las principales propiedades de los datos que impactan en la eficiencia de las predicciones, por lo menos en el ámbito de la predicción de la demanda eléctrica. Esto puede sentar las bases para un futuro estudio en el cual se busquen técnicas o estrategias que ayuden a los predictores a ser más robustos y precisos frente a estas propiedades.

1.5. Metodología

La metodología aplicada para el desarrollo de esa investigación consiste en una serie de fases en las cuales se analiza y procesa a los datos de manera que estos puedan ser descritos mediante un conjunto de características, las cuales son utilizadas para entrenar modelos de predicción utilizando distintos algoritmos de Machine Learning en diferentes periodos de tiempo, esto con el objetivo de comparar su eficiencia en cada periodo.

La descripción de lo que se realiza en cada fase es la siguiente:

- Fase de recolección: Se procede a descargar los datos diarios de la demanda energética desde la base de datos del Centro Nacional de Control de Energía (CENACE).
- Fase de preprocesamiento: Se realiza una integración del conjunto de datos descargado, de tal manera que estos estén ordenados por fecha y hora en un solo archivo, lo que facilita su análisis y manipulación.
- Fase de análisis de datos: Se dividen los datos en periodos correspondientes a un año, posteriormente se realiza un análisis exploratorio de los datos, mostrando sus siguientes características:
    - Normalidad
    - Volatilidad

    - Estacionalidad
    - Tendencia
- Fase de procesamiento: Utilizando los datos de la fase anterior se realiza una ingeniería de características sobre los datos, la cual tiene como objetivo definir un conjunto de características relevantes que ayude a los predictores a describir mejor el comportamiento de los datos y a elevar su eficiencia.
- Fase de modelado: Para cada periodo de tiempo, se implementan y entrenan los sistemas de predicción; XGBoost, Random Forest, SVR y KNN. Cada sistema requerirá un conjunto de hiperparametros los cuales serán aplicados al momento de su entrenamiento, por lo que se diseñó una función para cada sistema la cual tendrá como objetivo entrenar al predictor varias veces usando un conjunto de hiperparametros diferente, obteniendo así, el modelo que mejor describa los datos.
- Fase de pruebas: Se realiza un ensamble de los predictores creados, con el fin de mejorar la eficiencia de las predicciones individuales de cada uno de ellos.
- Fase de análisis de resultados: Se realiza una comparación de eficiencia de los predictores para cada periodo de tiempo, permitiéndonos ver el periodo en el cual se realizaron las predicciones más acertadas y obtener conclusiones con base en ello.

1.6. Descripción de los capítulos

En el capítulo 1, se describe la investigación que se llevara a cabo, así como la metodología a seguir. También se explica la decisión de trabajar sobre la predicción de la demanda energética.

En el capítulo 2, se obtiene, expone y compara el comportamiento de los datos de la demanda energética en México antes y durante de la pandemia, así como el preprocesamiento e ingeniería de características aplicada a ellos, con tal de extraer las características más importantes con las cuales se entrenan los predictores.

En el capítulo 3, se explica la metodología para la construcción y entrenamiento de cada uno de los modelos que describen el comportamiento de la demanda de energética en México, resaltando el periodo de tiempo en el que se realizaron mejores predicciones.

En el capítulo 4, se realizan pruebas haciendo un ensamble de los predictores obtenidos y se analizan los resultados de la eficiencia grupal (ensamble) e individual para la predicción de la demanda energética en los años 2019, 2020 y 2021.

En el capítulo 5, se presentan las conclusiones sobre este trabajo y algunos de los posibles trabajos futuros que pueden continuar desarrollándose como resultado de la investigación.

Capítulo 2

Análisis y preprocesamiento de datos

El presente capítulo presenta una descripción de la demanda energética en México, y a su vez expone la recolección, análisis y procesamiento de estos datos, con el fin de prepararlos para su uso en el entrenamiento de los modelos de predicción.

En la primera parte de este capítulo se da una ligera descripción de la forma en que está compuesto el sistema eléctrico en México, y la manera en la que se realizan las mediciones que la demanda energética.

En la segunda parte se muestra la forma en la que el conjunto de datos fue extraído y manipulado, con el objetivo de organizar los datos para facilitar su análisis.

En la tercera parte se realiza un análisis exploratorio de los datos, lo cual nos permite conocer sus características y comportamiento en diferentes periodos de tiempo.

En la cuarta parte se realizará una ingeniería de características, en la cual se añaden características que describen propiedades adicionales del conjunto de datos, lo que facilita a los predictores encontrar relaciones en su comportamiento y mejorar la precisión de sus predicciones.

2.1. Descripción de la demanda energética en México

El Sistema Eléctrico Nacional (SEN) es uno de los mayores y más complejos del mundo. Es un sistema integrado que da servicio a 128 millones de mexicanos, que habitan en dos millones de kilómetros cuadrados, y que ha alcanzado el 98.7 % de cobertura del servicio. [8]

Este sistema está compuesto por cuatro subsistemas eléctricos: el Sistema Interconectado Nacional (SIN), el Sistema eléctrico de baja california, el Sistema eléctrico Mulegé y el Sistema eléctrico de baja california sur, donde estos últimos tres sistemas se encuentran aislados del SIN. Fig.2.1

Figura 2.1: Subsistemas y regiones del Sistema Eléctrico Nacional [9]

El Sistema Interconectado Nacional es donde se concentra el mayor conjunto de infraestructura del país, ya que contempla siete regiones de la república: Central, Oriental, Occidental, Noroeste, Norte, Noreste y Peninsular, las cuales engloban a 30 de los 32 estados de la república. Fig.2.1

Cada región del SIN genera y demanda una cierta cantidad de energía, por lo que cuando una de ellas no produce lo suficiente, otra región tiene que suministrarle, con tal de mantener un equilibrio y evitar apagones. Por este motivo y debido al aislamiento de los otros tres sistemas, el enfoque de esta tesis está centrado únicamente en el SIN.

2.2. Datos de la demanda energética en México

Los datos de la demanda energética del SEN, utilizados para desarrollar la metodología para el pronóstico de la demanda energética, pueden ser encontrados en la página oficial del CENACE [10]. La recopilación y utilización de estos datos no requiere de ningún permiso, por lo que están a disposición del público.

La medición de la demanda real del SEN es estimada para cada hora del día por el CENACE, utilizando dos metodologías: por balance y por retiros.

- La estimación de la Demanda Real del Sistema por Balance se obtiene con base en la generación neta inyectada al sistema en cada hora, menos la energía de exportación. Se incluyen las pérdidas técnicas y no técnicas.

- La estimación de la Demanda Real del Sistema por Retiros se obtiene agregando todas las compras de energía que se realizan por las Entidades Responsables de Carga, incluyendo las exportaciones. Se excluyen las pérdidas técnicas y no técnicas de la red que corresponde al Mercado Eléctrico Mayorista.

Las estimaciones resultantes son agrupadas, ordenadas y puestas a disposición del público en archivos individuales que reflejan la demanda energética diaria para cada uno de los subsistemas y regiones que conforman el SEN.

Para este trabajo de tesis, los datos de la demanda energética utilizados fueron tomados desde el 4 de mayo 2016 al 31 de julio 2021, y separados en seis periodos de tiempo que corresponden a los años 2016, 2017, 2018, 2019, 2020 y 2021, con el fin de estudiar el comportamiento de los datos y mostrar su influencia en la eficiencia de los modelos de predicción.

2.3. Preprocesamiento de datos

El preprocesamiento de datos es el primer paso en la creación de un modelo de machine learning, ya que este consiste en pasar los datos brutos a un formato que sea más fácil de trabajar.

Para esta tarea y las siguientes se utiliza el lenguaje de programación Python, ya que cuenta con una gran cantidad de recursos y librerías disponibles que facilitan el análisis y tratamiento de los datos.

2.3.1. Extracción del conjunto de datos

Como se mencionó anteriormente, los datos de la demanda energética están separados por día de operación, lo que significa que la información de los periodos a analizar esta separada en distintos archivos, los cuales contienen información de todos los subsistemas y regiones del SEN. La primera tares es entonces, extraer los datos de cada región del SIN, con el fin de agruparlos en documentos separados de tal forma que estos contengan información exclusiva de cada región.

Al descargar los reportes de la demanda energética para un periodo de tiempo específico, lo que obtenemos es un conjunto archivos organizados por fecha, lo cual otorga una manera de identificar el día, mes y año al que corresponden los datos extraídos.

La extracción de la información se realizó de manera iterativa, lo que significa que cada archivo fue analizado de la misma forma. Durante el análisis de cada archivo se extrajeron las estimaciones de la demanda energética de cada una de las siete regiones que conforman en SIN, las cuales están ordenadas de manera horaria en un formato de 24hrs. Esto nos permite agrupar la información extraída en siete tablas correspondientes cada región del SIN, las cuales contienen un registro de la fecha y hora de las estimaciones de la demanda energética. En la Fig.2.2 se puede ver la tabla resultante para la región central.

	energy
2016-05-04 00:00:00	3921.623
2016-05-04 01:00:00	3867.698
2016-05-04 02:00:00	3759.255
2016-05-04 03:00:00	3774.048
2016-05-04 04:00:00	3796.157
...	...
2021-08-24 19:00:00	7206.89878
2021-08-24 20:00:00	7732.98357
2021-08-24 21:00:00	7576.25221
2021-08-24 22:00:00	7129.08996
2021-08-24 23:00:00	6507.23337

Figura 2.2: Demanda energética de la Región Central

2.3.2. Tratamiento de datos faltantes

Durante la recolección de datos existen ocasiones en las que por descuido humano se omite el registro de algunas variables después de realizar una medición, a este conjunto de variables no registradas se les conoce como datos faltantes. La razón por la que debemos tratar estos datos es debido a los errores que pueden provocar, ya que los algoritmos de ML no están diseñados para lidiar con ellos.

Existen dos maneras de lidiar con esta falta de datos, la primera es borrar todo el registro de la medición, lo cual puede reduce nuestro conjunto de datos, y la segunda es sustituir los datos faltantes.

Durante la extracción del conjunto de datos de la demanda energética, se observaron cinco datos faltantes para cada una de las siete regiones que conforman el SIN. Para lidiar con estos datos se aplicó una técnica de interpolación, la cual estima el dato desconocido calculando el promedio de dos puntos de datos adyacentes conocidos.

2.3.3. Escalamiento de los datos

El escalado es un proceso que se utiliza para cambiar el rango de valores de los datos sin afectar su distribución. Una de las razones por las cuales se realiza el escalado de dato es debido a que los algoritmos de ML son entrenados mediante un conjunto de características, las cuales toman distintos significados, sin embargo, estos algoritmos solo ven números, por lo que, si hay una gran diferencia en el rango, se asume que los más altos tienen algún tipo de superioridad, entonces estos números comenzaran a jugar un papel más decisivo durante el entrenamiento del modelo.

Otra razón para realizar el escalado, es que algunos algoritmos pueden mejorar su rendimiento, tales como las redes neuronales, máquinas de soporte vectorial y k-vecinos, esto debido a que su funcionamiento se basa en el cálculo de distancias entre puntos de datos.

Existe una variedad de formas en las que se pueden escalar los datos, sin embargo, las técnicas más populares son la normalización, la cual, cambia la distribución de los datos, y la estandarización, siendo esta última la empleada sobre los datos de la demanda energética.

La estandarización, también conocida como *"Z-Score Normalization"* es proceso de escalado donde los valores se centran alrededor de la media con una desviación estándar unitaria. Esto significa que la media del atributo se vuelve cero y la distribución resultante tiene una desviación estándar unitaria, tal como se muestra en la Fig.2.3. Al realizar la estandarización se asume que los datos tienen una distribución gaussiana, mejor conocida como distribución normal, sin embargo, esto no es un requisito.

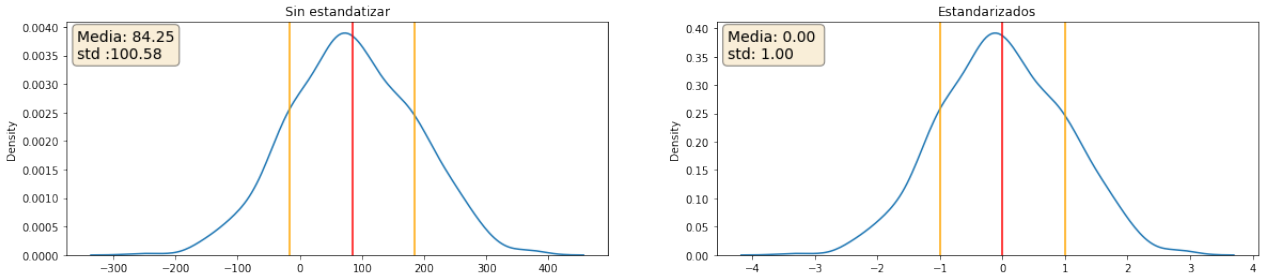


Figura 2.3: Datos estandarizados y no estandarizados

Para la estandarización de los datos de la demanda energética, se aplicó la fórmula 2.1 la cual consiste en restar cada dato entre la media y dividirlo entre su desviación estándar.

$$Z = \frac{X - \mu}{\sigma} \tag{2.1}$$

2.4. Análisis exploratorio de los datos mediante métodos estadísticos

El análisis exploratorio de datos (EDA, por sus siglas en inglés) es utilizado para investigar conjuntos de datos, resumir y presentar sus características principales. Permite una mejor comprensión de los datos mediante la identificación de características como: anomalías (outliers), saltos o discontinuidades, concentración de los valores, forma de la distribución, patrones de comportamiento, etc.

A menudo son empleados métodos de visualización tales como gráficas y diagramas, las cuales son complementadas con mediciones estadísticas tales como la media, desviación estándar y varianza.

Para realizar este análisis sobre el conjunto de datos de la demanda energética, los datos correspondientes al SIN y a cada una de las regiones que lo conforman fueron divididos por año (2016-2021), no solo con el fin de analizarlos individualmente, sino también de compararlos entre si con tal de ver su comportamiento en diferentes periodos de tiempo.

Los conjuntos de datos analizados ascienden a un total de 48, (7 regiones + SIN) * 6 periodos, debido a esto se generaron varias gráficas y estadísticas, las cuales por motivos de espacio fueron agrupadas en el apéndice de gráficas.

2.4.1. Análisis de normalidad

Dado un conjunto de datos, es posible generar un gráfico que indique la frecuencia con la que cada valor aparece en dicho conjunto, esta gráfica es conocida como histograma y a partir de ella podemos obtener la distribución de probabilidad de los datos. Una de las distribuciones más famosas y conocidas es la distribución de probabilidad normal o distribución Gaussiana Fig.2.4, ya que una gran cantidad de procesos en la naturaleza y en la sociedad siguen este tipo de distribución, algunos ejemplos de ello son: la altura, peso, coeficiente intelectual, puntuaciones de los exámenes, entre otros.

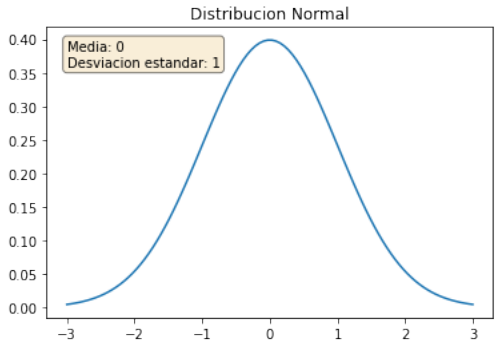


Figura 2.4: Distribución Normal o Gaussiana

En esta sección se realiza un contraste de la distribución de los datos de la demanda energética contra la distribución Gaussiana, para analizar cuanto difiere una distribución respecto de la otra, y descubrir los años y regiones del SIN donde los datos se asemejan más a este tipo de distribución.

El primer análisis realizado es una representación gráfica de los datos mediante un histograma, donde se superpone una curva de densidad de probabilidad continúa. Adicionalmente, se muestran estadísticas como la media y desviación estándar, las cuales están representadas por una línea roja y naranja respectivamente.

En la parte superior izquierda de cada gráfica se muestra el skewness, kurtosis y Pval de la distribución de probabilidad, los cuales representan las siguientes características:

- Skewness: Es la simetría de la distribución de probabilidad, un valor negativo indica que la curva es más larga por el lado izquierdo, mientras que su valor positivo indica que la curva es más larga por el lado derecho. El skewness para una distribución normal es 0.

- Kurtosis: Muestra la concentración de valores alrededor de la media, de manera gráfica se ve representado en cuan puntiaguda o cuan chata es la punta de la distribución. Valores negativos indican una concentración menor de los datos (punta menos picuda), mientras que valores positivos indican una mayor concentración (punta más picuda). La kurtosis para una distribución normal es 0.
- Pval: Obtenido a partir de la prueba de Shapiro Wilk, la cual se usa para contrastar la normalidad de un conjunto de datos. La prueba rechaza la hipótesis de normalidad cuando el valor p es menor o igual a 0.05. No aprobar la prueba de normalidad le permite afirmar con un 95 % de confianza que los datos no se ajustan a la distribución normal. Pasar la prueba de normalidad solo le permite afirmar que no se encontró una desviación significativa de la normalidad.

A continuación, en la Fig.2.5 se muestra el análisis de normalidad correspondiente a la región Noroeste. Esta región fue seleccionada específicamente debido a que mostró una de las mayores afectaciones respecto a la distribución de los datos durante los años de la pandemia. Cómo se puede observar, en años previos a la pandemia (en especial el 2016) la distribución de los datos tiende a la de una distribución normal, que, si bien no es del todo exacta, su forma y simetría no difieren en gran medida. En cambio, para los años 2020 y 2021 podemos afirmar tanto por la forma como simetría, que no hay relación alguna con una distribución normal.

Una posible razón para este comportamiento se debe a que esta es una de las principales regiones industriales en México, la cual se vio afectada por el cierre de actividades no esenciales declarada el 30 de marzo del 2020 en el diario oficial de la federación [11]. De igual forma, en la región central, otra de las principales regiones industriales del país, las gráficas de distribución mostraron una alteración en la normalidad de los datos durante los años 2020 y 2021.

Otra representación utilizada con frecuencia para analizar la normalidad de un conjunto de datos son los gráficos de cuantiles teóricos (Graficos Q-Q). Estos gráficos comparan los cuantiles de la distribución observada con los cuantiles teóricos de una distribución normal con la misma media y desviación estándar que los datos. Cuanto más se aproximen los datos a una normal, más alineados están los puntos en torno a la recta.

Para contrastar con el análisis anterior, esta vez se muestra el gráfico Q-Q de la región Oriental Fig.2.6, cuya distribución fue la menos afectada en los años de pandemia.

A diferencia de la región central y noroeste que son grandes zonas industriales, la región Oriental no presenta esta característica, pudiendo ser una razón por la cual su normalidad no se vio afectada al mismo grado que las otras.

En conclusión, los histogramas y gráficos Q-Q de las regiones en cada uno de los periodos de tiempo (2016-2021), muestran que la distribución de los datos en años anteriores a la pandemia tienen una mayor similitud a una distribución normal que en los años 2020 y 2021, por lo que podemos decir que la pandemia altero en mayor o menor medida la normalidad de los datos en todas las regiones del SIN.

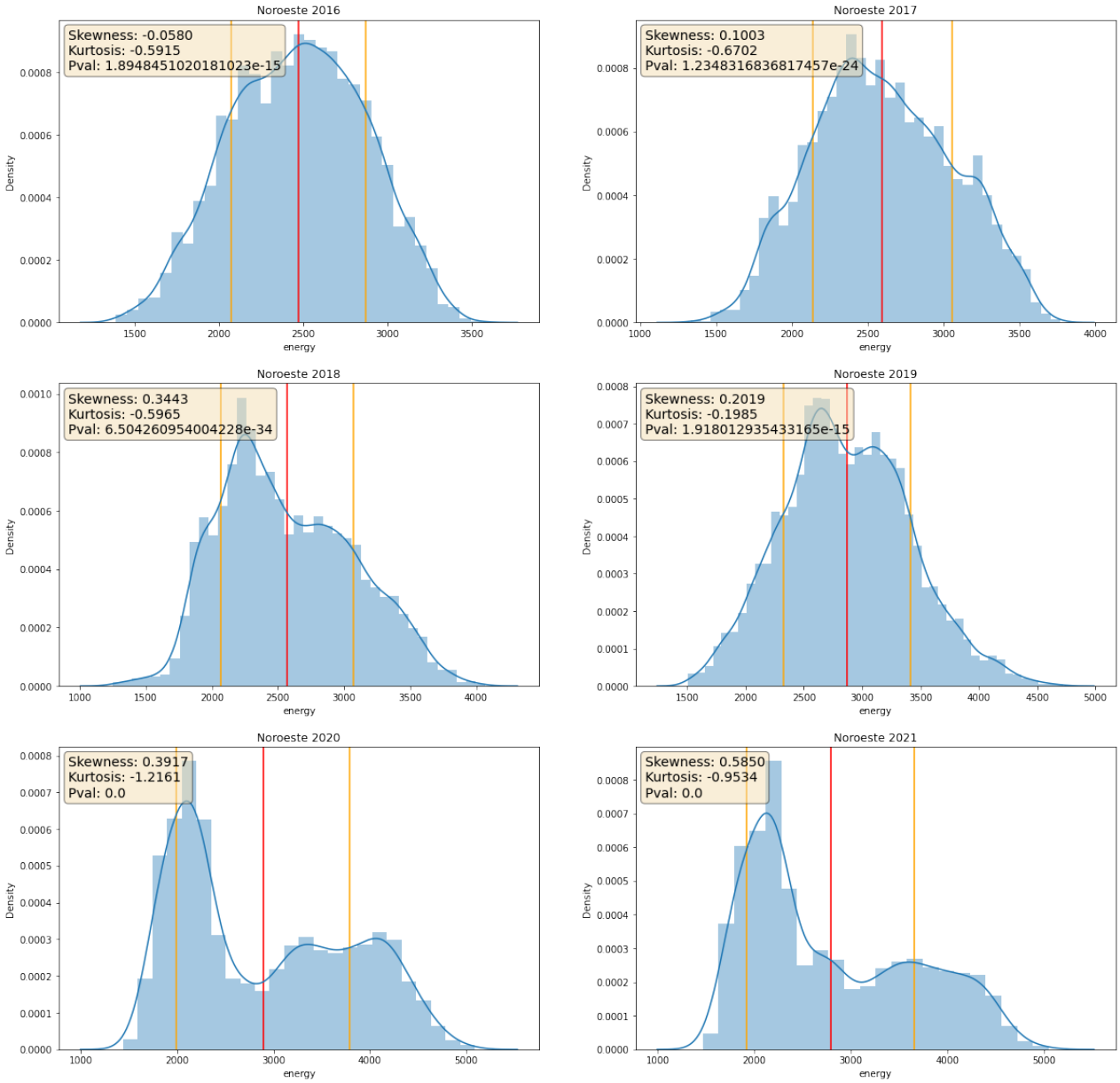


Figura 2.5: Análisis de Normalidad en la región Noroeste

Los gráficos del análisis de normalidad que muestran los histogramas para los demás años y regiones se encuentran en el Apéndice A, mientras que los gráficos Q-Q se ubican en el Apéndice B.

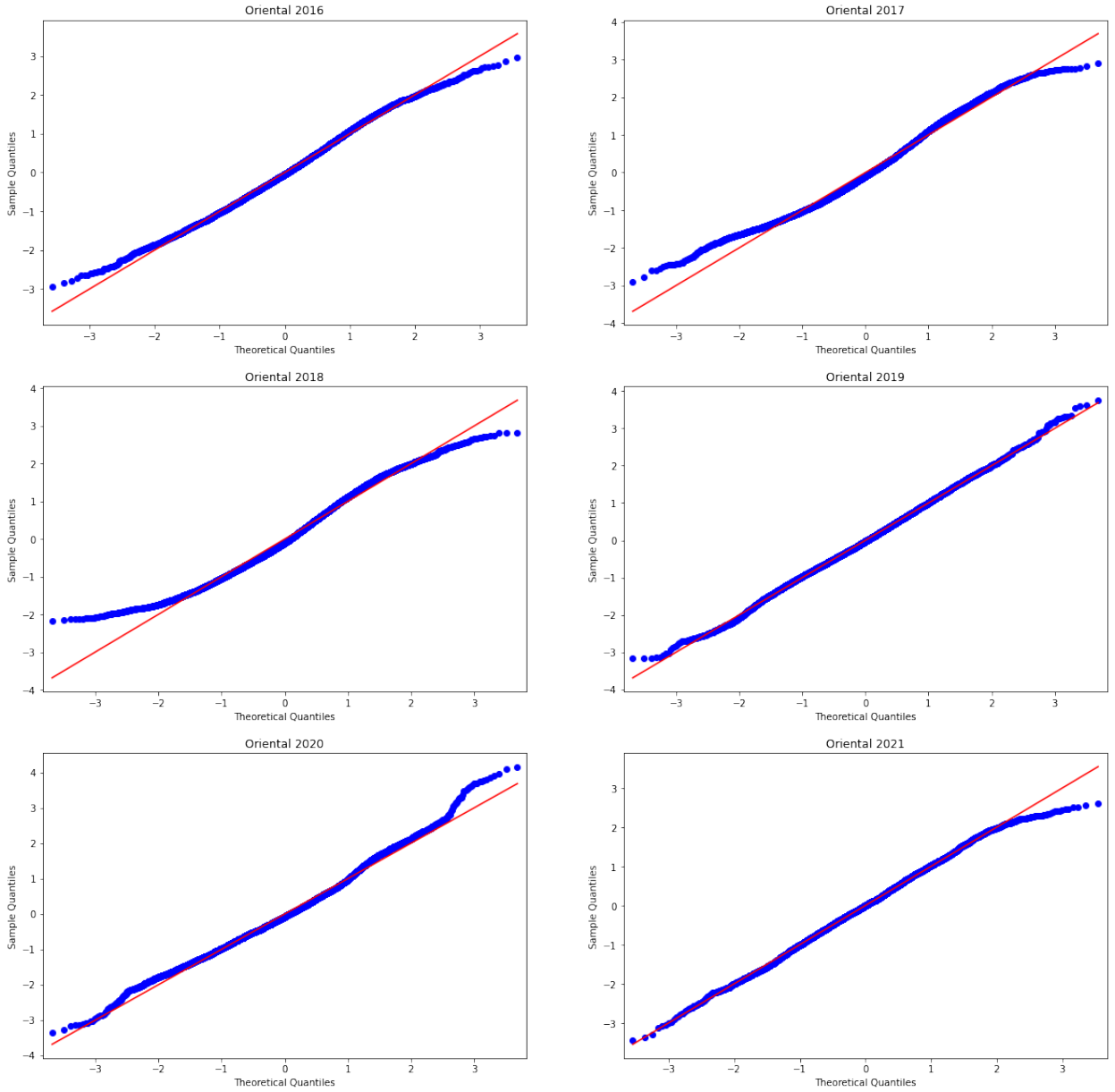


Figura 2.6: Gráficos QQ para el análisis de normalidad de la región Oriental

En Machine Learning, los datos que satisfacen la distribución normal son beneficiosos para la creación de modelos, ya que facilita las matemáticas para algoritmos como Gaussian Naive Bayes, Regresión logística y Regresión lineal, debido a que parten del supuesto de que la distribución de los datos es normal. Sin embargo, la normalidad es solo una suposición para los algoritmos de ML, es decir, no es obligatorio que los datos sigan siempre esta normalidad, ya que, modelos como árboles de decisión o XGboost, no consideran la distribución de los datos.

2.4.2. Análisis de Volatilidad

La volatilidad es una medida de la frecuencia e intensidad de los cambios en los valores de una serie de datos a lo largo del tiempo. Matemáticamente, es la desviación estándar calculada durante un período de tiempo; una medida de cuánto se distribuyen los números alrededor de la media. Una volatilidad alta significa que los valores, o en este caso las magnitudes de la demanda energética tienen el potencial de distribuirse en un rango más amplio de valores, significando que su magnitud puede cambiar drásticamente en un corto periodo de tiempo en cualquier dirección (positiva o negativa). Una volatilidad baja significa que las fluctuaciones de los valores no cambian drásticamente y tienden a ser más estables.

Una manera de visualizar la volatilidad es por medio de los cuantiles, ya que con ellos podemos conocer la concentración de los datos y su variación a través del tiempo.

Los cuantiles son puntos de corte que dividen el rango de una distribución de probabilidad en intervalos continuos, o que dividen a las observaciones en una muestra de la misma forma. Los deciles son el nombre que reciben los cuantiles que crean un grupo de diez divisiones en la distribución de probabilidad, los cuales van desde D_1 a D_9.

Cada decil divide a la distribución de probabilidad en dos partes, por ejemplo: para el decil D1 todo el conjunto de valores que están a su derecha (valores $>$ D1) corresponden al 90 % de todo el conjunto de datos, mientras que los de su izquierda (valores $<$ D1) corresponden al 10 % de los datos. Esto mismo ocurre para los demás deciles, valores menores al D_N-esimo decil corresponden al N*10 % de los datos, mientras que los valores mayores al D_N-esimo decil corresponden al (1-N)*10 %.

En la Fig.2.7, se muestra la gráfica de los deciles D_1, D_5, y D_9 correspondientes a cada región para a los últimos 3 meses de cada paso en el tiempo, lo cual nos permite ver su evolución a través de los años.

La volatilidad de cada región se ve representada tanto por las variaciones como la distancia entre las gráficas de los deciles. La distancia entre D1 (azul) y D9 (verde) representa el rango de valores en los que oscila el 80 % de los datos de la demanda energética en un periodo de 3 meses, siendo que, a mayor distancia, mayor intensidad de los cambios.

Como se puede observar en la Fig.2.7, todas las regiones presentaron un cambio brusco poco tiempo después de entrado en año 2020, y algunas de ellas como la región noreste, noroeste y norte, presentan una mayor dispersión de los datos. Sin embargo, regiones como la central, occidental, oriental y peninsular presentan una disminución en dicha dispersión.

El análisis de los deciles proporciona una manera gráfica de visualizar la volatilidad de un conjunto de datos, sin embargo, es difícil hacer comparaciones precisas a través del tiempo, de modo que esta no es una forma apropiada de medir con certeza las regiones y años en donde se presentó una mayor dispersión de los datos.

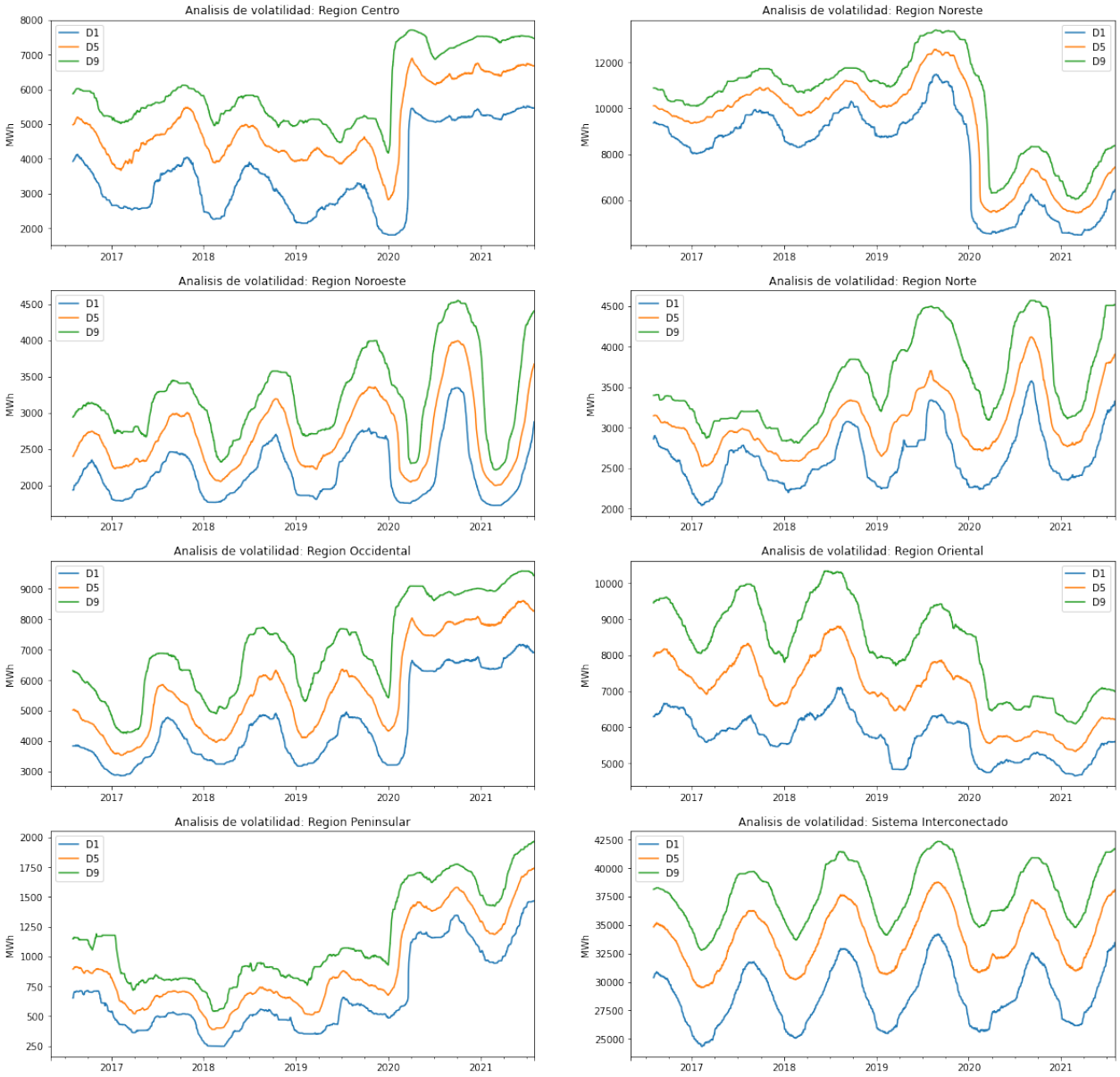


Figura 2.7: Análisis de volatilidad mediante la graficación de Deciles

Una manera cuantitativa de medir la volatilidad de la demanda energética es a través del coeficiente de variación (CV), el cual se expresa siguiendo la ecuación 2.2. Este coeficiente es una medida estadística (expresada en porcentaje) que nos informa la dispersión del conjunto de datos alrededor de la media. Esta métrica es comúnmente usada para comparar la dispersión de los datos entre distintas series de datos. A diferencia de la desviación estándar que siempre debe considerarse en el contexto de la media de los datos, el coeficiente de variación proporciona una forma relativamente simple para comparar series de datos con diferentes características.

$$CV = \frac{\sigma}{\mid \mu \mid} \tag{2.2}$$

Este coeficiente fue utilizado como una medida para comparar la volatilidad de cada una de las regiones en diferentes contextos, siendo el primero una comparativa en los diferentes periodos de tiempo, y el segundo una comparación en varias divisiones de estos periodos, como son: trimestral, mensual, y por días de la semana.

En la Fig.2.8 se muestra el ejemplo para la región centro, donde la gráfica (1) muestra el CV de manera anual, e indica que durante los primeros cuatro años (2016 a 2019), este coeficiente fue en aumento, sin embargo, con la llegada del 2020 y 2021, se observó una disminución de este en aproximadamente un 10 %, lo cual es un indicativo de que la volatilidad de los datos de la demanda energética fue menor durante los años de pandemia.

Las gráficas siguientes (2, 3 y 4) muestran el coeficiente de variación de manera trimestral, mensual y semanal, con los cuales se realiza una comparación tomando en cuenta dos años anteriores a la pandemia (2018-2019) y los dos años que llevamos de ella (2020 y 2021). Al igual que en la gráfica (1), estos resultados muestran que la llegada de la pandemia disminuyo la variación de la demanda energética en estos periodos de tiempo, para esta región.

Asimismo, este comportamiento se vio replicado en las regiones occidental, oriental y peninsular, sin embargo, las regiones noreste, noroeste y norte, demostraron el comportamiento contrario, en donde la dispersión de los datos de la demanda energética era menor antes de la pandemia. Las gráficas de las demás regiones pueden ser encontradas en el apéndice C.

Medir la volatilidad nos da una idea de la frecuencia e intensidad de los cambios que puede experimentar la demanda energética, lo cual otorga una noción del grado de libertad o conservacionismo que los modelos deben de tener para poder ajustarse a los cambios en el comportamiento de los datos.

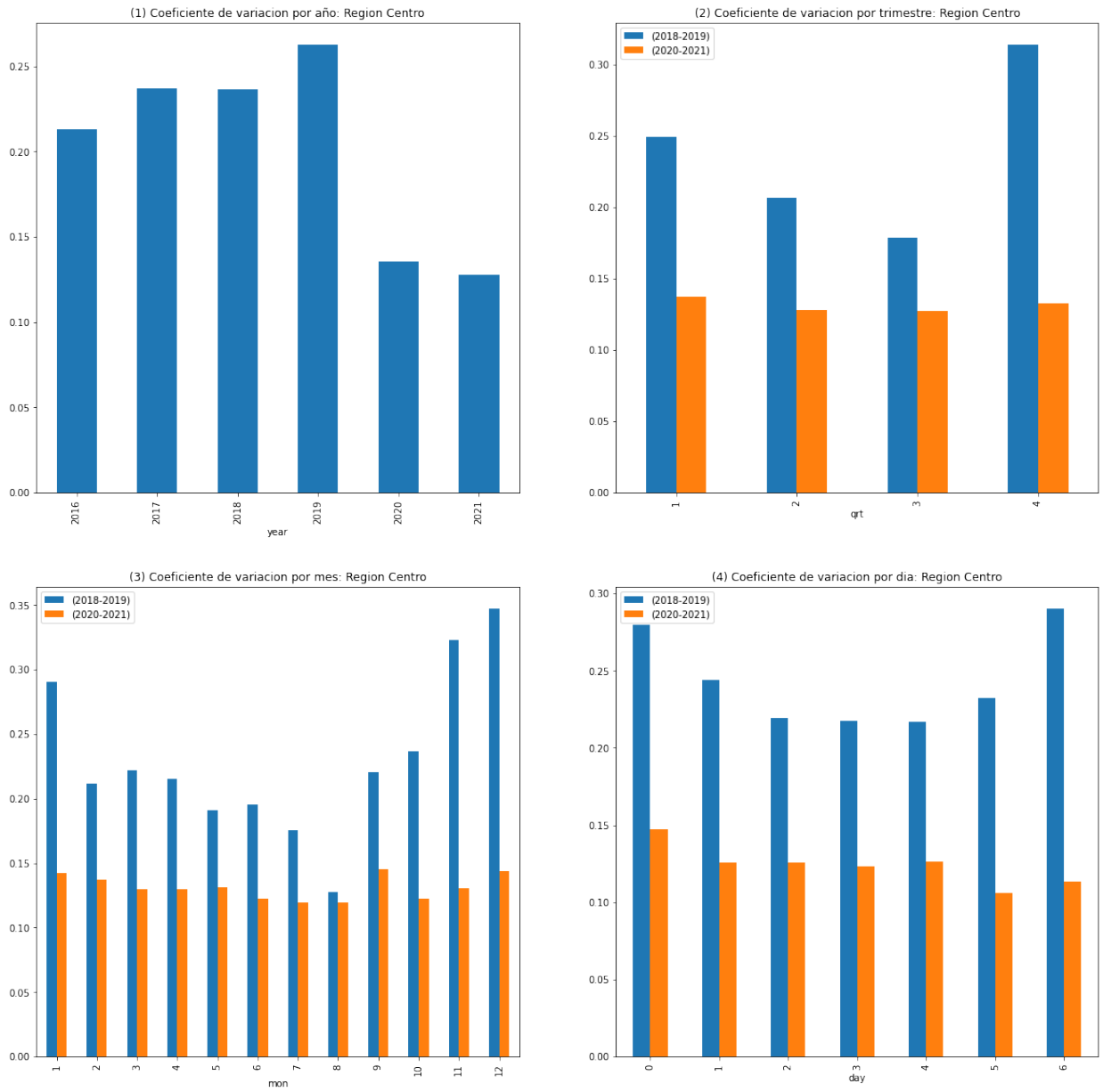


Figura 2.8: Coeficientes de variación de la región central

2.4.3. Estacionalidad

La estacionalidad es un comportamiento o patrón que a veces se observa en una serie de tiempo. Consiste en subidas y bajadas periódicas que se presentan en forma regular, las cuales permiten conocer el comportamiento de los datos y facilitar su predicción.

Una forma sencilla de visualizar la estacionalidad en una serie de tiempo es graficar la relación del tiempo y las variables involucradas, que en este caso es la demanda energética. Por lo tanto, el primer análisis de estacionalidad es realizado mediante la graficación de los promedios mensuales y trimestrales de cada paso en el tiempo en el SIN y sus regiones, Fig.2.9.

Figura 2.9: Promedios móviles de la demanda energética en las regiones del SIN

Al analizar las gráficas de la Fig.2.9 y centrándonos en los años anteriores a la pandemia, podemos apreciar un patrón de crecimiento y decrecimiento a principios y finales de cada año en los promedios de la demanda energética de cada región, los cuales difieren en magnitud a través de los años, pero logran ser recurrentes. Este comportamiento puede apreciarse más claramente en las regiones Noroeste, Occidental, y Oriental.

Con la llegada del 2020 la estacionalidad en las regiones se vio afectada de diferentes formas. En las regiones Central, Occidental, Oriental y Peninsular la estacionalidad de la demanda se vio bastante alterada ya que su comportamiento difiere en gran medida con respecto a los años anteriores, e incluso no es posible distinguir un patrón a simple vista.

En la región noreste, a principios del 2020 se observó un retraso en el patrón que se venía viendo en años anteriores, pero se observa que la estacionalidad sigue siendo la misma, teniendo como única diferencia una disminución en la magnitud de los promedios de la demanda energética.

Por último, en las regiones Noroeste y Norte el patrón de la demanda solo se vio afectado en amplitud, oscilando en un rango más amplio de valores, sin embargo, la estacionalidad siguió siendo la misma.

Para complementar este análisis, además de las gráficas de los promedios móviles, también se muestran los diagramas de caja-bigotes de cada región para los lapsos de tiempo 2016-2019 y 2020-2021.

Estos diagramas representan los tres cuartiles (Q1, Q2 y Q3) y los valores mínimo y máximo de un conjunto de datos, sobre un rectángulo alineado verticalmente Fig.2.10. El Extremo superior e inferior indican el valor mínimo y máximo del conjunto de datos. Los cuartiles son el nombre que reciben los cuantiles (explicados en el subtema 2.4.2), por lo que su significado es el mismo. El cuartil inferior (Q1) indica que el 25 % de los datos se concentra en valores más bajos que él, mientras que el 75 % restante se encuentra en valores más altos. La mediana (Q2) es el cuartil que divide el 50 % de los datos y el cuartil superior (Q3) aquel que divide al 75 % de los datos. Por último, los valores atípicos representan puntos individuales fuera del rango de 1.5*(Q3 – Q1), los cuales no son considerados parte del conjunto de valores.

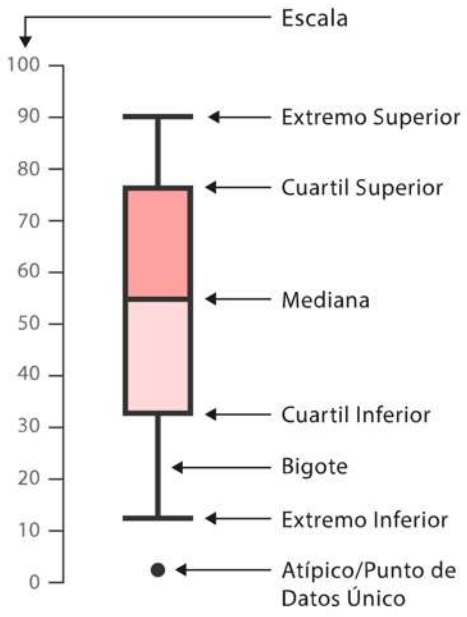


Figura 2.10: Diagrama de caja-bigote [12]

La finalidad de mostrar los diagramas de caja-bigote es exponer a través de ellos la estacionalidad de cada región en diferentes periodos de tiempo: mensual, trimestral y semanal, con tal de comparar su comportamiento antes y durante la pandemia.

En la Fig.2.11 se muestra el diagrama de caja-bigote para la región occidental, donde, al observar la estacionalidad mensual de la demanda energética para los periodos previos y durante la pandemia, podemos ver el comportamiento antes explicado (Fig.2.9) para esa misma región, en donde se observa un patrón estacional de incremento y decremento en años anteriores a la pandemia, pero durante ella, este comportamiento cambia.

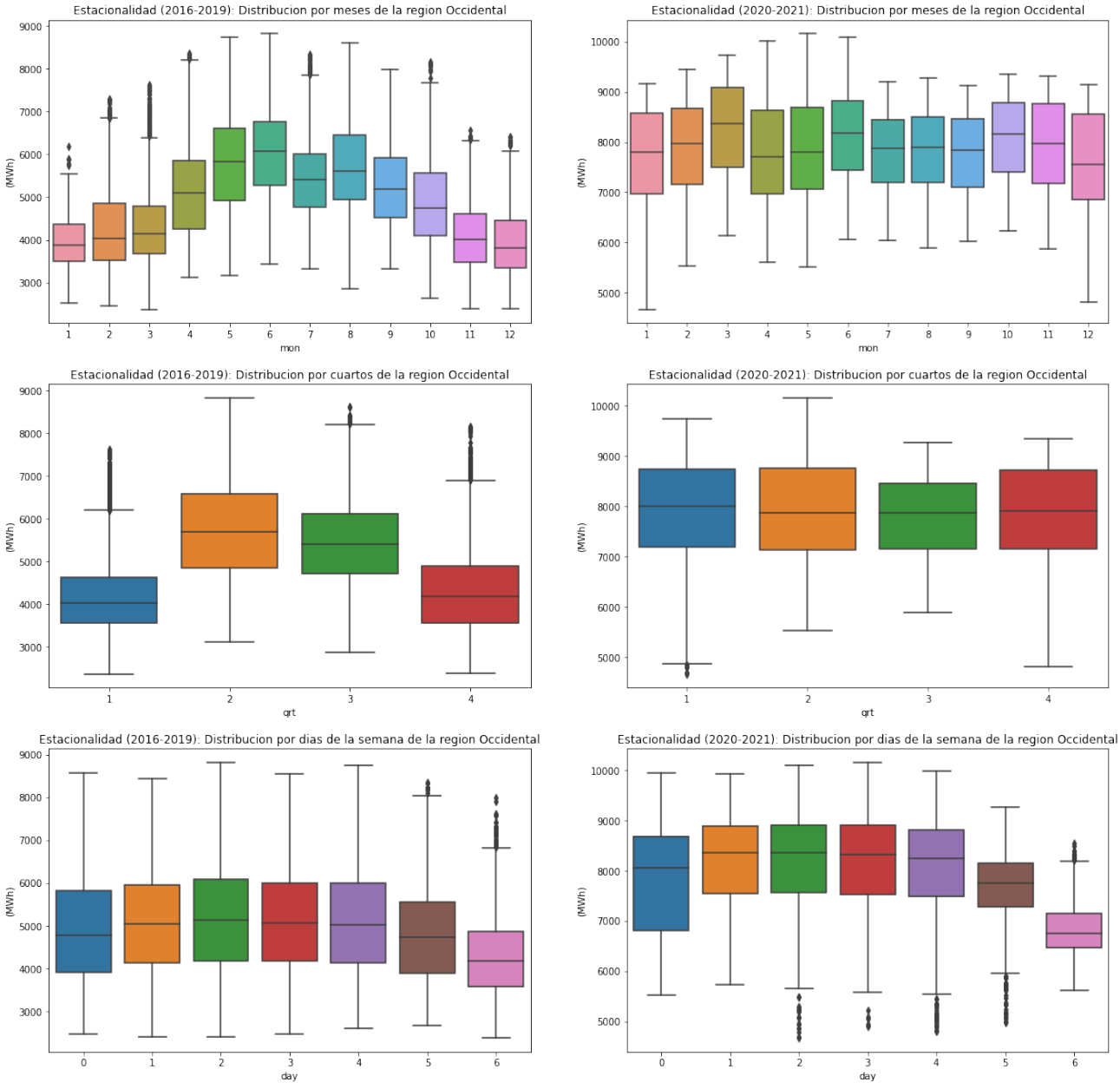


Figura 2.11: Diagramas de caja para el análisis de estacionalidad en la región Occidental

De la misma forma, podemos ver a nivel trimestral la existencia de un patrón estacional diferente en tiempos anteriores y durante la pandemia, donde en los años 2016 a 2019 se observa que en el primer y cuarto trimestre se genera una menor demanda energética, pero en los años 2020 a 2021 este comportamiento cambia y la demanda parece ser la misma en todos los trimestres. Sin embargo, la demanda estacional a nivel diario continúa siendo el mismo antes y durante la pandemia, teniendo que en los días lunes a viernes (0 a 4) experimenta una mayor demanda energética que los fines de semana (5 y 6).

En conclusión, para las regiones en donde la estacionalidad siguió siendo la misma antes y durante la pandemia, los predictores solo tendrán que ajustarse al nuevo rango de valores de la demanda energética, sin embargo, para las regiones donde la estacionalidad se vio alterada, los predictores tendrán que lidiar con el nuevo patrón de comportamiento de los datos.

Para una mejor claridad en la explicación del análisis de estacionalidad mediante diagramas de caja-bigote, solo se expusieron los diagramas de la región Occidental, sin embargo, los gráficos correspondientes a las demás regiones se encuentran agrupados en el apéndice D.

2.4.4. Tendencia

La tendencia se refiere al patrón de movimiento que sigue una secuencia de observaciones de una serie temporal en un momento determinado. Visto de otro modo, la tendencia es la dirección que sigue una serie de valores. Esto, teniendo en cuenta que los datos de una serie, usualmente, fluctúan en un intervalo. Es decir, se mantienen entre un máximo y un mínimo, dibujando en una representación gráfica un zigzag.

El análisis de la tendencia proporciona información para detectar un patrón de comportamiento subyacente en los elementos de una serie temporal.

Una forma de analizar la tendencia en los datos de la demanda energética es visualizando los diagramas de caja-bigotes de manera anual, lo cual nos muestra un resumen de la concentración de los datos y su variación a través de los años. Otra manera de analizar la tendencia, es haciendo una regresión lineal sobre el conjunto de datos, la cual consiste en trazar una línea recta que mejor se ajuste a ellos, siendo que, la inclinación (pendiente) de la recta es usada como indicador de la dirección hacia la cual crece o decrece la demanda energética.

En la Fig.2.12 se muestra el análisis de la tendencia para la región centro, donde se muestran los diagramas de caja-bigote de cada año de estudio, y un análisis por regresión lineal, en donde cada punto en la gráfica representa la suma mensual de la demanda energética a lo largo de cada año recopilado. Cade decir, que el número total de meses que hay entre los periodos de tiempo recopilados es de 63, siendo este el número de muestras con el cual se realizó la regresión lineal para cada región.

La gráfica naranja representa la tendencia pre-pandemia, es decir, la tendencia comprendida entre los años 2016 a 2019, ignorando los años de pandemia. Por otra parte, la gráfica azul representa la tendencia actual, que incluye los años 2020 y 2021.

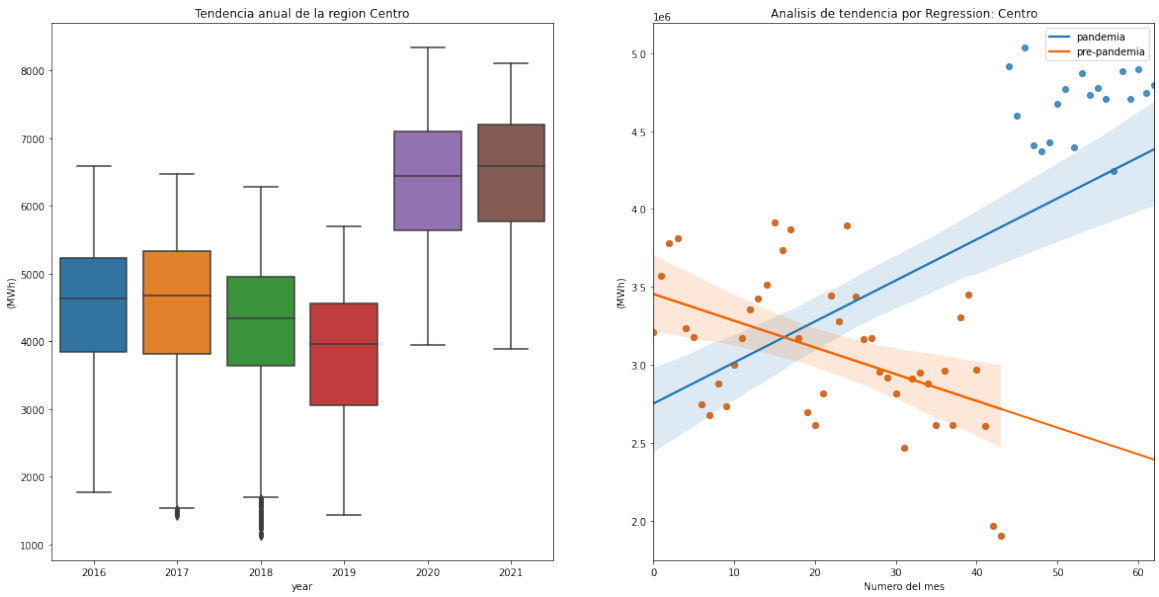


Figura 2.12: Análisis de tendencia para la región Centro

Como podemos observar, en el diagrama de caja para la región centro, en los años previos a la pandemia se observó una tendencia descendente en la demanda energética, sin embargo, para la llegada del 2020 se vio un cambio abrupto en el incremento de la demanda, lo cual ocasiono que la línea trazada mediante la regresión lineal incrementara su pendiente y describiera una tendencia ascendente.

Contrario a este comportamiento, podemos ver que en el análisis de tendencia para la región noreste Fig.2.13, donde inicialmente la demanda energética mostraba un comportamiento ascendente, sin embargo, con la llegada de la pandemia hubo un gran decremento en el consumo energético, lo cual provoco que la pendiente de la recta trazada fuera negativa y describiera una tendencia decreciente, sin embargo, el comportamiento de los diagramas de caja para los años 2020 y 2021 dejan ver que la tendencia de la demanda energética va nuevamente en aumento.

De manera general, y al analizar la tendencia del sistema interconectado, podemos apreciar que año con año existe un incremento en la demanda energética, la cual experimento una disminución en el 2020, pero se está recuperando en este 2021. Las gráficas para este análisis y los correspondientes a las demás regiones, pueden ser encontrados en el apéndice E.

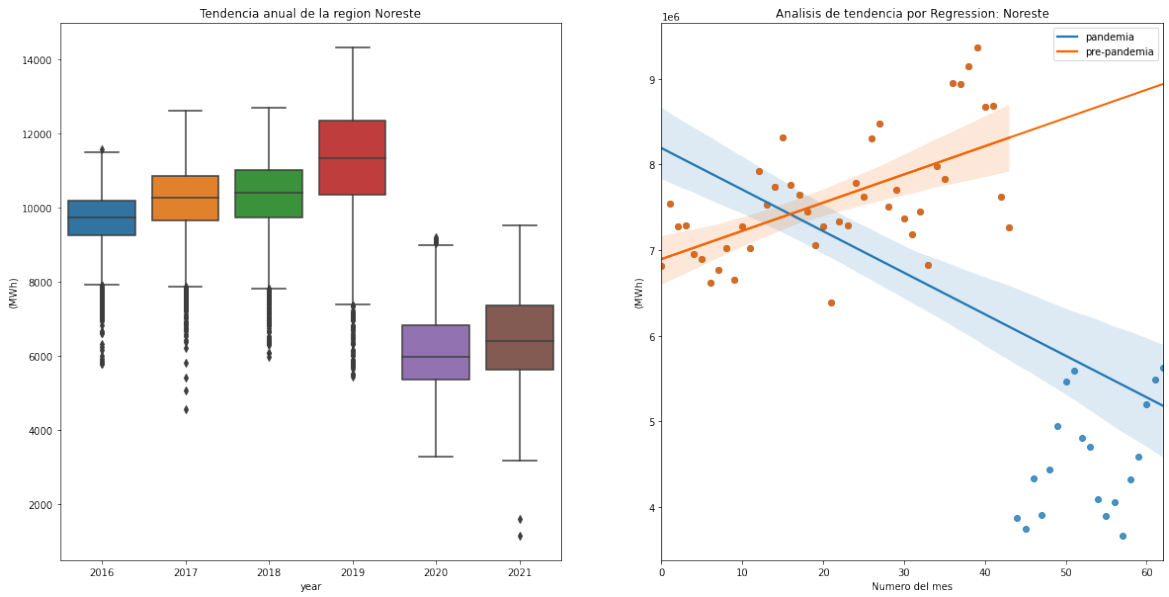


Figura 2.13: Análisis de tendencia para la región Noreste

2.5. Ingeniería de características

La ingeniería de características es el proceso de creación y selección de variables explicativas (características) a partir de un conjunto de datos, las cuales pueden ser usadas para entrenar un modelo de predicción mediante algoritmos de ML. Cuanto mejores sean las características, mejores serán los resultados obtenidos.

El objetivo de la ingeniería de características es preparar un conjunto de datos de entrada conformado por características, las cuales proporcionan información adicional acerca del fenómeno que representan los datos, lo que permite una mejor descripción de su comportamiento y mejora la eficiencia de los modelos entrenados.

Los modelos predictivos constan de una variable objetivo y variables predictoras, y es durante el proceso de ingeniería de características que se crean y seleccionan las variables predictoras más útiles para el modelo predictivo. Por ejemplo, en un modelo que predice los precios de las propiedades, los datos que muestran los precios reales son la variable objetivo, mientras que los datos que muestran cosas como el tamaño de la casa, el número de dormitorios y la ubicación, son las variables predictoras que determinan el valor de la propiedad.

2.5.1. Creación de características

Originalmente, el conjunto de datos de la demanda energética consta de dos partes, la primera es la magnitud de la demanda expresada en mega watts (MWh) y la segunda es un índice del tiempo (fecha y hora) en la cual se presentó dicha demanda. Tabla 2.1

Con un total de 45960 datos registrados y ninguna característica de utilidad para la predicción de la demanda energética, se procedió a crear características relacionadas con el tiempo y datos pasados.

Time	energy
2016-05-04 00:00:00	3921.62300
2016-05-04 01:00:00	3867.69800
2016-05-04 02:00:00	3759.25500
2016-05-04 03:00:00	3774.04800
2016-05-04 04:00:00	3796.15700
...	...
2021-07-31 19:00:00	6635.60310
2021-07-31 20:00:00	7001.43186
2021-07-31 21:00:00	7011.49156
2021-07-31 22:00:00	6686.36267
2021-07-31 23:00:00	6237.14921

Cuadro 2.1: Datos originales de la demanda energética de la región centro

Las características temporales son de ayuda para identificar la época en la que nos encontramos, lo cual permite reconocer patrones en el comportamiento de los datos, tal como se vio en la sección 2.4.3, donde la demanda energética exhibe un patrón de conducta a nivel trimestral, mensual y diario.

Al considerar características de datos pasados, como el valor de la demanda del día anterior o del último mes, podemos identificar la tendencia de los datos, y con ayuda de características estadísticas tales como el mínimo, máximo, media y desviación estándar, incluso es posible definir el rango y dispersión de los mismos, lo que permite determinar su magnitud de manera más precisa.

2.5.1.1. Características temporales

A partir del índice del "Tiempo" mostrado en la tabla 2.1 y con la ayuda de Python, se extrajeron las siguientes características: Trimestre, mes, semana, día y hora. Las primeras tres correspondientes al año en cuestión, mientras que las últimas dos corresponden al día de la semana y a la hora del día. Cada una estas características pretenden describir el comportamiento de la demanda energética en diferentes agrupaciones de tiempo. Por ejemplo, ya hemos visto que existe un patrón a nivel trimestral, mensual y diario, sin embargo, no está de más tratar de encontrar un patrón a nivel semanal y horario.

Las características resultantes se pueden observar en la tabla 2.2, donde se muestra una configuración numérica para cada una de las etiquetas resultantes, por ejemplo: Los días lunes, martes, miércoles, jueves, viernes, sábado y domingo, están etiquetados de manera numérica, siendo que el número 0 corresponde al lunes y número 6 al domingo. De la misma forma ocurre para los trimestres (1 - 4), meses (1 - 12) y días de la semana (1 - 53).

Esta etiquetación numérica ocurre de manera predeterminada ya que muchos algoritmos de ML no pueden trabajar directamente con datos categóricos, por lo que las categorías deben convertirse en números de tal forma que sean admitidas para entrenar modelos.

Time	energy	qtr	mon	week	day	hour
2016-05-04 00:00:00	3921.623	2	5	18	2	0
2016-05-04 01:00:00	3867.698	2	5	18	2	1
2016-05-04 02:00:00	3759.255	2	5	18	2	2
2016-05-04 03:00:00	3774.048	2	5	18	2	3
2016-05-04 04:00:00	3796.157	2	5	18	2	4

Cuadro 2.2: Características temporales

El tipo de relación que expresan los valores numéricos que toman las distintas divisiones de los años, días y horas, representa el orden en el que aparecen, sin embargo, esta numeración puede provocar que al entrenar el modelo este capture algún tipo relación diferente al orden, como lo puede ser lunes

En la tabla 2.3 se muestra la codificación One-Hot para los trimestres del año, donde está representada por las columnas qtr_1, qtr_2, qtr_3 y qtr_4, las cuales son variables binarias que indican la presencia del trimestre del año en cuestión.

Time	qtr	qtr_1	qtr_2	qtr_3	qtr_4
2016-06-30 22:00:00	2	0	1	0	0
2016-06-30 23:00:00	2	0	1	0	0
2016-07-01 00:00:00	3	0	0	1	0
2016-07-01 01:00:00	3	0	0	1	0
2016-07-01 02:00:00	3	0	0	1	0

Cuadro 2.3: Codificación One-Hot para los trimestres anuales

Esta codificación también fue aplicada para los meses, semanas, días y horas, sin embargo, por motivos de espacio no es posible mostrar su representación al igual que en la tabla anterior.

2.5.1.2. Características estadísticas

Este tipo de características se obtiene a partir del uso de valores anteriores con el cual se realizan operaciones matemáticas tales como la suma, resta, multiplicación y división. Para su creación se emplea un método denominado como "ventana rodante", el cual itera sobre los datos, tomando un conjunto de N datos anteriores entre cada iteración, con los cuales se realizan operaciones matemáticas para generar los valores de la nueva característica.

Las características creadas a partir de este método son: la media, desviación estándar, mínimo y máximo, tomando en cuenta los datos de cada día, semana, quincena y mes anteriores. Esto significa que se consideraron cuatro grupos de datos, o cuatro ventanas rodantes, con las cuales se generaron las cuatro características mencionadas anteriormente, obteniendo como resultado un total de 16 características, las cuales tienen como objetivo ayudar a identificar el rango y dispersión de los datos considerando distintos los intervalos de tiempo.

En la tabla 2.4 se muestra el resultado de las características correspondientes al minino, máximo, media y desviación estándar de las últimas 24 hrs anteriores.

Time	energy	movmin1	movmax1	movave1	movstd1
2016-06-30 22:00:00	5578.952	5750.494	4160.806	5169.05483	523.887746
2016-06-30 23:00:00	5196.676	5716.074	4160.806	5161.90725	516.731391
2016-07-01 00:00:00	4834.18	5716.074	4160.806	5160.14854	516.529795
2016-07-01 01:00:00	4765.986	5716.074	4160.806	5161.44279	515.63444
2016-07-01 02:00:00	4756.977	5716.074	4160.806	5175.58467	498.98935

Cuadro 2.4: Características estadísticas del día anterior

2.5.1.3. Características de retardo

A menudo, los valores pasados de una variable en un momento determinado puede estar relacionados con el valor de la misma, indicando una correlación entre ellos, la cual puede ser de utilidad para la predicción de sus valores futuros.

Al igual que las características estadísticas, las características de retraso son valores creados a partir de datos anteriores, con la diferencia que estos se obtienen individualmente, y no se le es aplicada ninguna operación matemática.

Para determinar la cantidad de datos pasados a tener en cuenta, se hizo una gráfica de autocorrelación sobre los datos de la demanda energética en el SIN. La autocorrelación consiste en una representación matemática del grado de similitud entre una serie de tiempo dada y una versión retrasada de la misma en diferentes intervalos de tiempo sucesivos, siendo que un valor de 1 representa una correlación perfecta, y un valor 0 indica una ausencia de correlación.

En la Fig.2.14 se muestra la gráfica de autocorrelación para las últimas 24 hrs del SIN, en la cual se puede apreciar que existe una correlación mayor a 0.7 para las primeras y últimas 4 horas de atraso del día en cuestión, lo cual indica que los datos en dichas horas tienen una alta correlación con la demanda futura, por lo que pueden ser usados para su predicción. Esta autocorrelación se ve replicada en cada una de las regiones del SIN, en donde algunas de ellas como la región Noroeste y Peninsular, es factible considerar todos los datos de las últimas 24 hrs para realizar predicciones. Las gráficas de autocorrelación de las demás regiones se encuentran en el apéndice F.

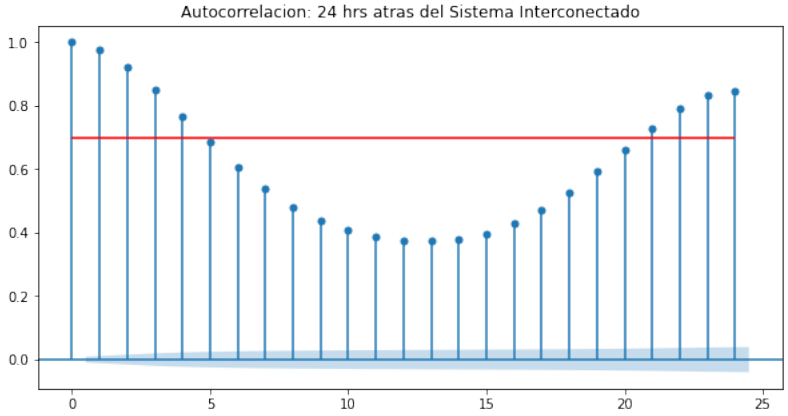


Figura 2.14: Gráfica de autocorrelación del SIN con 24 retrasos

Por lo tanto, las características obtenidas consisten en traer los últimos 24 datos anteriores, obteniendo así una colección de 24 características nuevas que representan la demanda energética en cada una de las últimas 24 hrs.

En la tabla 2.5 se muestran las características de retraso para las últimas 5 horas, las cuales provienen de la columna target que representa los datos estandarizados de la demanda energética.

Time	**target**	**feat_ar1**	**feat_ar2**	**feat_ar3**	**feat_ar4**	**feat_ar5**
2016-06-30 22:00:00	0.487823	0.471926	0.448526	0.454704	0.538669	0.584851
2016-06-30 23:00:00	0.217320	0.487823	0.471926	0.448526	0.454704	0.538669
2016-07-01 00:00:00	-0.039185	0.217320	0.487823	0.471926	0.448526	0.454704
2016-07-01 01:00:00	-0.087440	-0.039185	0.217320	0.487823	0.471926	0.448526
2016-07-01 02:00:00	-0.093815	-0.087440	-0.039185	0.217320	0.487823	0.471926

Cuadro 2.5: Primeras cinco características de retraso de la demanda energética estandarizada

2.6. Selección de características

La selección de características se refiere al proceso de elegir el conjunto de variables predictoras más importantes que serán utilizadas para entrenar a los modelos de predicción. Esto permite la eliminación de aquellas características que son redundantes o irrelevantes, lo cual reduce la dimensionalidad del problema de entrenamiento, crea modelos más simples de explicar, reduce el tiempo de entrenamiento, aumenta la precisión de las estimaciones y evita la maldición de la alta dimensionalidad [13], la cual señala que el error aumenta con el incremento del número de características.

Con un total de 144 características, la selección de las más relevantes se llevó a cabo empleando el método de correlación de Pearson, el cual mide la correlación de cada una de las variables predictoras contra la variable objetivo, puntuando de esta manera las características de mayor influencia sobre la demanda energética.

Este método consiste en el cálculo de coeficientes que miden la fuerza con la que dos variables se correlacionan. Para realizar este cálculo es necesario obtener la covarianza entre las variables, que es un valor que indica su grado de variación conjunta respecto de sus medias, el cual se calcula como el promedio del producto entre los valores de cada variable, donde los valores se les ha restado la media (ecuación 2.3).

$$cov(x, y) = \frac{1}{n} * \sum_{i=1}^{n} \frac{(x_i - \mu_x) * (y_i - \mu_y)}{n} \tag{2.3}$$

El coeficiente de Pearson es calculado como la covarianza dividida por el producto de la desviación estándar de cada muestra de datos (ecuación 2.4). Consiste en la normalización de la covarianza entre las dos variables para dar una puntuación interpretable.

$$r_{x,y} = \frac{cov(x,y)}{\sigma_x * \sigma_y} \tag{2.4}$$

El coeficiente puede devolver un valor entre -1 y 1 que representa los límites de una correlación completamente negativa o positiva respectivamente. Cuando el valor del coeficiente es 0, significa que no hay correlación. Una correlación positiva (

Para obtener mejores resultados, la selección de características se llevó a cabo de manera individual para cada una de las regiones del SIN, por lo que las características usadas en el entrenamiento de los modelos pueden variar entre regiones.

En la Fig.2.15 se tiene una matriz de correlación expresada con un mapa de calor (heatmap en inglés), en el cual se muestra la correlación de las 10 características más relevantes para la predicción de la demanda energética en la región central. Cada celda contiene el coeficiente de correlación de Pearson y una tonalidad de color asociado, los cuales expresan el grado de correlación que existe entre cada par de variables.

Como se puede observar, las primeras cinco variables más relevantes pertenecen a las características de retraso, que indican el valor de la demanda energética de las últimas 1 ,2, 3, 23 y 22 horas, sin embargo, también aparecen algunas características estadísticas que indican el promedio, mínimo y máximo de la demanda energética en el último día de operación.

El coeficiente de correlación otorga una medida para identificar a las mejores características que podrían ser usadas al momento de entrenar los modelos de predicción, sin embargo, su resultado no es absoluto. Por esta razón, las 144 características creadas fueron ordenadas en una lista de acuerdo a su correlación con variable de la demanda energética estandarizada (target). Esta lista se utiliza para entrenar diversos modelos de predicción tomando en cuenta un número diferente de características cada vez, permitiendo seleccionar el modelo de mayor precisión.

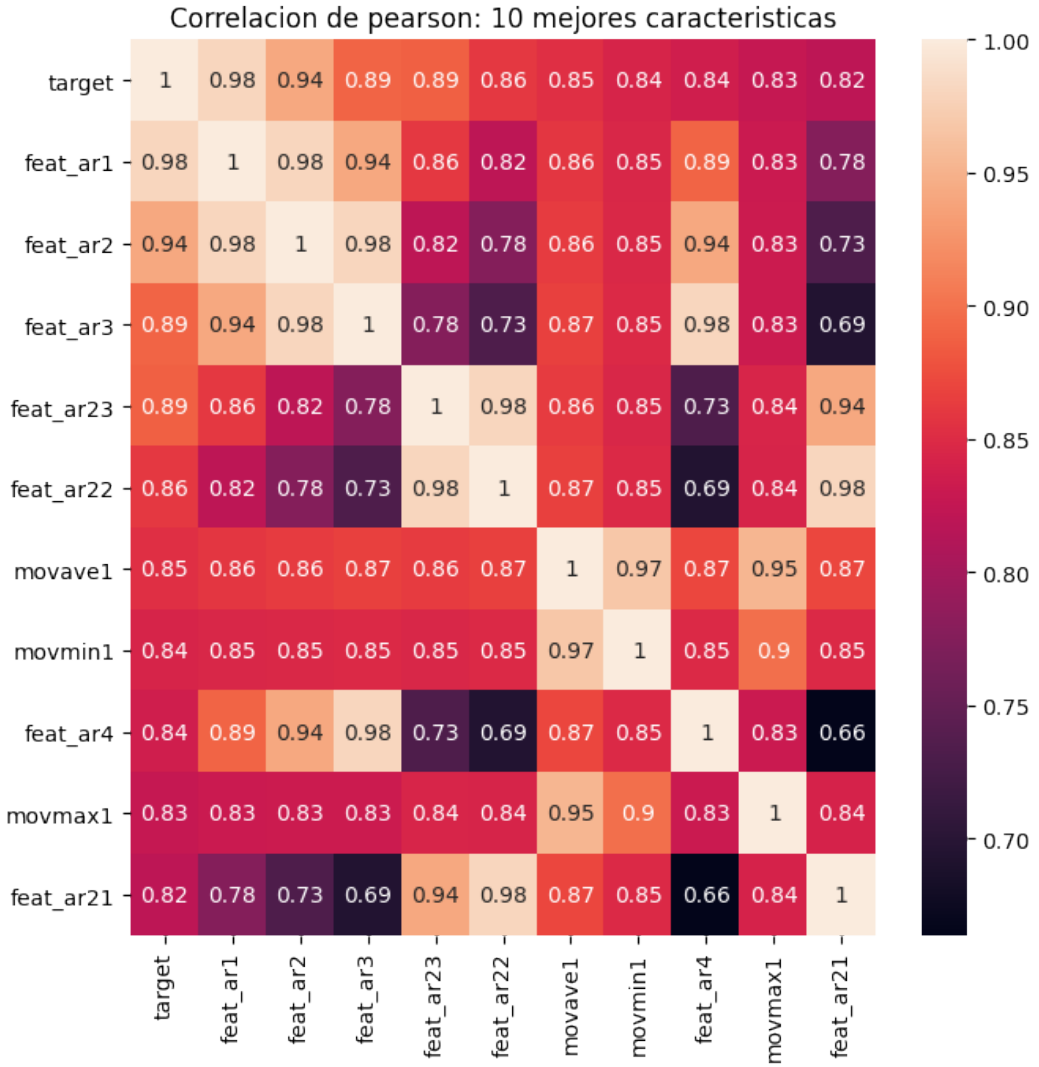


Figura 2.15: Matriz de correlación con las 10 características más relevantes de la región Centro

Comentarios finales

En este capítulo se describió la composición del SEN y la forma en la que los datos de la demanda energética fueron extraídos y procesados para realizar un análisis estadístico sobre ellos, el cual mostró un cambio en su comportamiento con la llegada de la pandemia.

Finalmente se aplicó una ingeniería de características sobre los datos procesados, con la finalidad de ser aprovechados en el entrenamiento de diferentes modelos de ML, los cuales son implementados en el capítulo 3.

Capítulo 3

Implementación de modelos de predicción

El presente capítulo expone la implementación de los modelos de predicción utilizando los siguientes algoritmos: Random Forest, XGboost, Regresión vectorial de soporte (SVR) y K vecinos más cercanos (KNN), los cuales son entrenados utilizando los datos procesados en el capítulo 3.

En la primera parte de este capítulo se explica de manera general el funcionamiento del algoritmo Random Forest, así como los hiperparametros utilizados para su implementación y los resultados de su eficiencia al predecir un conjunto de datos de prueba. De la misma forma, la segunda, tercera y cuarta parte de este capítulo explican lo anterior mencionado para los algoritmos XGBoost, SVR y KNN.

3.1. Evaluación de los modelos

Para mostrar la eficiencia de las técnicas de predicción antes y durante la pandemia, se implementó un modelo de predicción para cada una de las siete regiones del SIN y para cada año de predicción, 2019, 2020 y 2021, usando cada uno de los algoritmos anteriormente mencionados.

Estas predicciones fueron realizadas para el mes de abril en cada uno de los años mencionados. La razón es que en dicho mes del año 2020 es cuando inicia el cierre de actividades no esenciales en México, dando por terminando, la época de "normalidad".

Tanto las características como los hiperparametros utilizados para la construcción de cada modelo fueron obtenidos de manera experimental, esto quiere decir que se probaron distintas combinaciones de ellos hasta encontrar aquella con la que se obtuvieron mejores resultados.

La selección de los hiperparametros se llevó a cabo a través de la experimentación, utilizando la librería de ParameterGrid de Python, la cual permite establecer una red de hiperparametros e iterar a través de ellos, dando la oportunidad de entrenar los modelos utilizando todas las combinaciones posibles

Adicionalmente, la selección del conjunto de características se llevó a cabo realizando el coeficiente de correlación de Pearson en cada una de las regiones y organizando las características de acuerdo con los resultados obtenidos, con el fin de entrenar a los modelos utilizando un conjunto incremental de 40 características cada vez, esto quiere decir, que, por cada combinación de hiperparametros, los modelos son entrenados utilizando diferentes cantidades de características.

La evaluación de los modelos se llevó a cabo mediante la estimación de sus errores de predicción, los cuales fueron obtenidos a través de las siguientes métricas de error:

- Error cuadrático medio (RMSE, por sus siglas en inglés), obtenido como la raíz cuadrada del promedio de los errores al cuadrado.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{3.1}$$

  Dado que los errores se elevan al cuadrado antes de promediarlos, RMSE otorga un peso relativamente alto a los errores grandes. Esto significa que RMSE es más útil cuando los errores grandes son particularmente indeseables.

- Error absoluto medio (MAE, por sus siglas en inglés), obtenido como el promedio del error absoluto entre los valores predichos y los valores reales

$$MAE = \frac{1}{n} * \sum_{i=1}^{n} |Y_i - \hat{Y}_i| \tag{3.2}$$

  A diferencia del error RMSE, MAE es una puntuación lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio.

- Error porcentual absoluto medio (MAPE, por sus siglas en inglés), obtenido como el promedio del porcentaje de error entre los valores predichos y los valores reales

$$MAPE = (\frac{1}{n} * \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i}) * 100 \tag{3.3}$$

  El error MAPE, es una métrica, que a diferencia del MSE y MAE, mide el tamaño del error en términos porcentuales que hacen más fácil su interpretación, sin embargo, presenta algunas limitantes debido a la forma en la que se calcula [14], ya que, mientras más cercanos a cero estén los valores reales, los correspondientes errores porcentuales serán más altos, sesgando la información que proporciona. Debido a esta limitante, el error MAPE solo es usado como métrica auxiliar, no jugando un papel decisivo para comparar la eficiencia entre los modelos de predicción.

3.2. Random Forest Regressor

El concepto de Random Forest fue introducido por Leo Breiman en 2001 [15]. Para explicarlo, partiremos de los árboles de decisión. Los árboles de decisión son métodos comúnmente utilizados en tareas de clasificación y regresión. La predicción se realiza sobre el modelo (árbol) construido, el cual está formado por nodos, ramas y hojas, Fig.3.1. Cada nodo representa una variable (característica) que divide al conjunto de datos en dos partes, de tal manera que ambos conjuntos contienen instancias de valores similares (homogéneos), esto hace que en cada nivel de profundidad del árbol se creen subconjuntos con valores cada vez más homogéneos, permitiendo una mejor predicción/clasificación de la variable objetivo de acuerdo a sus características.

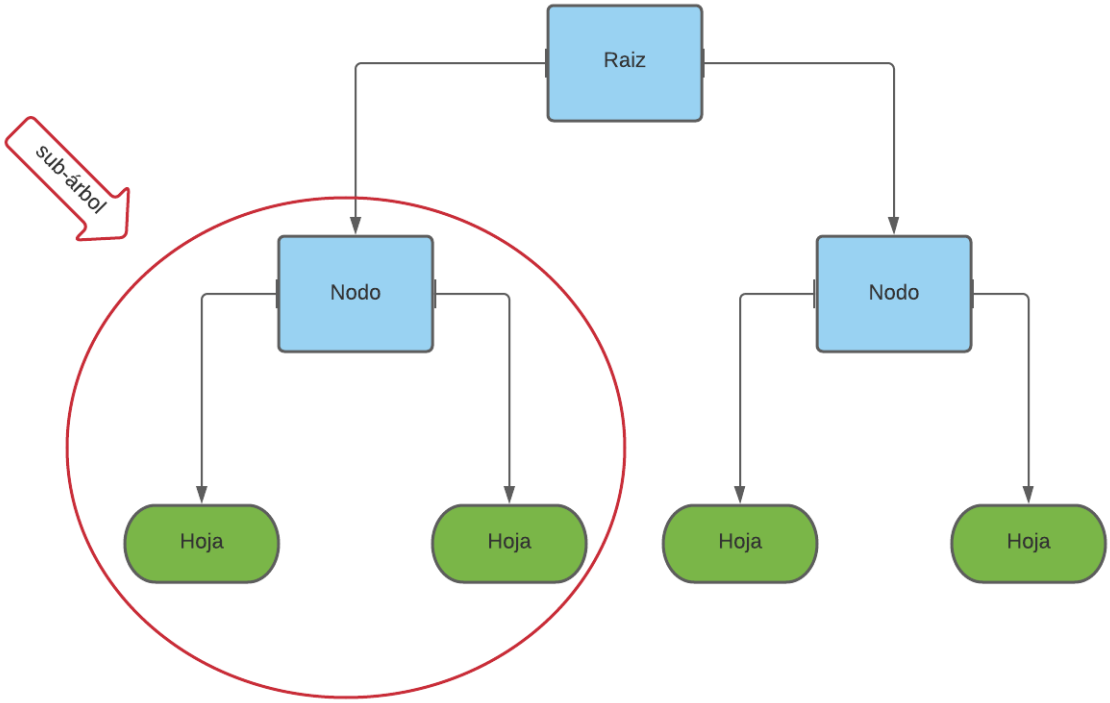


Figura 3.1: Representación de un árbol de decisión

El posicionamiento de los nodos en cada nivel del árbol está determinado por el grado de impureza de la división resultante de los datos, el cual indica la homogeneidad de los mismos al ser divididos por la característica X. Existen diversas medidas para determinar la impureza y seleccionar la característica que mejor divida a los datos: para problemas de clasificación la más común es el coeficiente de Gini, mientras que para la regresión es el error cuadrático medio (MSE).

Normalmente el árbol generado está sobre ajustado (bias bajo), lo que implica una alta varianza en el modelo y reduce la precisión para la predicción o clasificación de nuevos datos, es por esto que los árboles tienen que ser "podados", es decir, convertir ciertos sub-árboles en hojas, con tal de mitigar el sobreajuste.

Algunas de las limitantes más importantes que conllevan los árboles de decisión son su inestabilidad ante el ruido y el aumento en su complejidad frente a grandes conjuntos de datos, lo que implica una alta varianza y bias bajo. Es por esto que un solo árbol no es lo suficientemente bueno, lo que conlleva a la creación de múltiples de ellos con el fin de formar un conjunto de modelos llamado ensamble.

Los métodos más utilizados para el ensamble de modelos son el boosting y el bagging:

- Boosting: Es una forma de técnica de aprendizaje secuencial. El algoritmo funciona entrenando un modelo con todo el conjunto de entrenamiento, y los modelos posteriores se construyen ajustando los valores de error residual del modelo inicial. De esta manera, Boosting intenta dar mayor peso a aquellas observaciones que el modelo anterior estimó pobremente. Una vez que se crea la secuencia de los modelos, las predicciones hechas por los modelos son ponderadas por sus puntuaciones de precisión y los resultados finales se obtienen mediante un promedio ponderado de acuerdo a las puntuaciones de cada modelo.
- Bagging (Boostrap Agregating): Consiste en el entrenamiento de un conjunto de modelos de manera independiente, de tal manera que cada uno de ellos aporta un resultado diferente. El resultado final es estimado mediante una votación (para clasificadores) o un promedio (para regresores) de los resultados de cada modelo.

Boostrap Agregating entrena a cada modelo con un conjunto de datos diferente. Los conjuntos de datos son generados mediante bootstrapping, el cual genera nuevos conjuntos de datos del mismo tamaño usando muestreo con reposición. El muestreo con reposición toma un punto de datos y lo devuelve, de tal manera que este pueda ser reelegido por el resto de muestras.

Random forest es un algoritmo que utiliza el método de ensamble Bagging para la construcción un conjunto de árboles de decisión que ayudan a reducir las limitaciones que tiene un solo árbol. Los pasos involucrados en este algoritmo son:

1. Crear un conjunto de datos A mediante bootstrapping
2. Seleccionar aleatoriamente un subconjunto B de características, en las cuales basar la decisión del particionamiento de los datos.
3. Calcular el nodo D del árbol, utilizando un método de división (Gini, MSE, etc.) para obtener la característica que mejor divida el conjunto de datos A.
4. Repetir los pasos 2 y 3 para crear un árbol de decisiones individual.
5. Evaluar la precisión del árbol creado utilizando los datos que quedaron fuera del conjunto A.
6. Construir un bosque aleatorio repitiendo todos los pasos N número de veces, para crear N número de árboles.

En Random Forest, la predicción resultante se calcula como el promedio de las predicciones individuales de cada árbol creado.

En la Fig.3.2 se observa una representación de un bosque aleatorio, donde los círculos verdes representan el camino de las decisiones tomadas por los diferentes árboles ante los mismos datos de entrada.

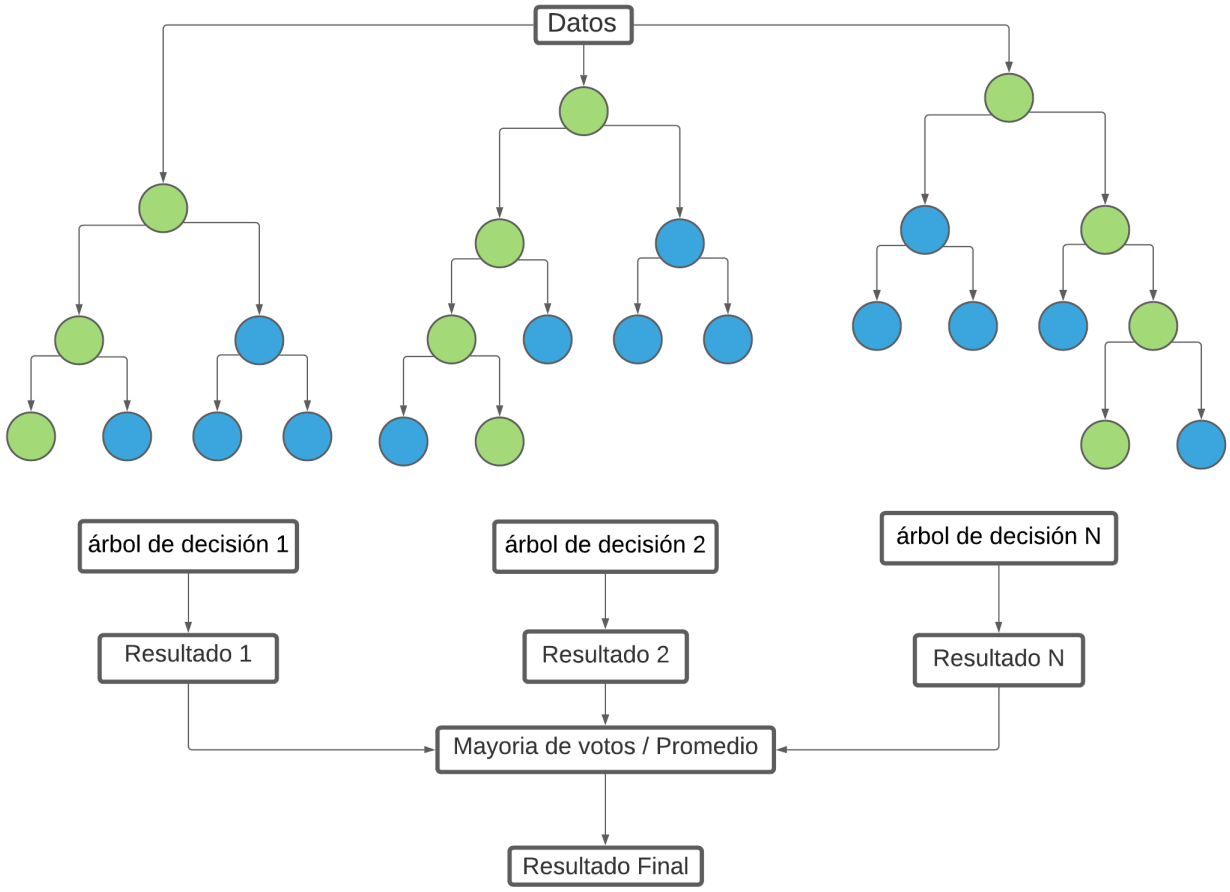


Figura 3.2: Representación de un bosque aleatorio

Los bosques aleatorios o Random Forest combinan varios árboles de decisión con bagging. Pueden ser lentos cuando hay muchos datos o características, sin embargo, tienen un menor sobreajuste y mayor precisión a comparación de solo un árbol de decisión. Sus hiperparametros más importantes son el número y profundidad de los árboles.

3.2.1. Implementación y resultados

La implementación de Random Forest se realizó utilizando la librería *sklearn.ensemble.RandomForestRegress* de Python. Para la implementación de los modelos se experimentaron con diferentes valores del siguiente conjunto de hiperparametros:

- n_estimators: 200, 300, 500, 600, 800
- max_depth: 20, 25 50, 100
- min_samples_split: 3, 6, 10

El hiperparametro *n_estimators* decide la cantidad de árboles de regresión que serán creados en el bosque aleatorio. Un número muy grande de árboles puede generar sobreajuste. *max_depth* regula la profundidad que puede tener el árbol, valores muy pequeños pueden derivar la incapacidad para realizar predicciones acertadas. *min_samples_split* indica el número mínimo de observaciones en cualquier nodo para que este pueda ser dividido, valores altos impiden que el modelo aprenda relaciones muy específicas entre los datos.

En la tabla 3.1 se muestra el resultado de las predicciones realizadas en el mes de abril para los años 2019, 2020 y 2021, en el SIN y sus regiones.

Región	**2019**			**2020**			**2021**		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
Centro	611.79	455.84	14.45	**362.53**	**276.81**	**4.58**	526.00	397.92	6.40
Noreste	1350.60	1180.67	10.54	**585.30**	**466.72**	**8.70**	1321.21	1160.72	17.66
Noroeste	**216.29**	**170.10**	**6.70**	302.18	195.85	8.04	236.05	236.05	7.38
Norte	381.18	303.13	10.48	**350.29**	**279.47**	**9.15**	436.185	369.46	10.70
Occidente	1250.24	1009.22	15.59	**571.03**	**481.34**	**6.83**	1084.18	993.39	11.92
Oriente	848.54	681.94	10.41	**496.85**	**379.44**	**6.71**	478.05	408.03	6.57
Peninsular	182.49	156.97	19.16	**169.44**	**127.45**	**9.22**	212.56	180.91	11.76
SIN	2812.67	2469.27	7.23	**1841.73**	**1448.48**	**4.71**	3811.13	3422.37	9.78

Cuadro 3.1: Errores de predicción usando Random Forest

En su mayoría, las respuestas obtenidas en cada uno de los tres años, muestran que durante la pandemia se realizaron mejores predicciones, más específicamente en el 2020. Una excepción se encuentra en la región Noroeste, donde las predicciones más acertadas fueron realizadas en el año 2019. Los resultados muestran que Random Forest se adecuó mejor al comportamiento de la demanda eléctrica durante los años de pandemia.

3.3. XGBoost Regressor

El algoritmo XGBoost se desarrolló como un proyecto de investigación en la Universidad de Washington, el cual fue presentado por Tianqi Chen y Carlos Guestrin en la Conferencia SIGKDD en 2016 [16]. Al igual que Random forest, XGBoost es un algoritmo de ML que emplea un método de ensamble utilizando árboles de decisión. XGboost utiliza el método de ensamble de potenciación del gradiente o Gradient Boosting, cuyo objetivo es minimizar la función de pérdida del modelo utilizando el descenso del gradiente, mediante la agregación de árboles básicos.

Un modelo XGBoost está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial. Cada nuevo árbol emplea información del árbol anterior para aprender de sus errores, entre cada iteración. En cada árbol individual, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal, al igual que en los árboles de decisión. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

La flexibilidad de este algoritmo ha hecho posible aplicar boosting a multitud de problemas (regresión, clasificación múltiple...) convirtiéndolo en uno de los métodos de machine learning de mayor éxito. Si bien existen varias adaptaciones, la idea general de todas ellas es la misma: entrenar modelos de forma secuencial, de forma que cada modelo ajusta los residuos (errores) de los modelos anteriores. Los pasos involucrados en este algoritmo son:

1. Se ajusta un modelo inicial F_1 con el que se predice la variable Y, a menudo F1 es el promedio de Y.

2. Se calculan los pseudo-residuales Y-F_1. Es la diferencia entre las observaciones Y y la estimación inicial F_1.
$$F_1 \approx Y$$

3. Se crea un modelo F_2, a partir de un árbol de regresión, que intenta predecir los pseudo-residuales del modelo anterior, en otras palabras, trata de corregir los errores que ha hecho el modelo F_1.

$$F_2 \approx Y - F_1$$

4. Se calculan los pseudo-residuales de los dos modelos en forma conjunta, los errores cometidos por F_1 y que F_2 no ha sido capaz de corregir, y se ajusta un tercer modelo F_3 para tratar de corregirlos.
$$F_3 \approx Y - F_2 - F_1$$

5. Este proceso se repite N veces, de forma que cada nuevo modelo minimiza los residuos (errores) del anterior. La creación de modelos se detiene hasta los pseudo-residuales sean muy pequeños o se haya alcanzado el máximo de árboles establecidos.

Dado que el objetivo de Gradient Boosting es ir minimizando los residuos iteración a iteración, es susceptible al sobreajuste. Una forma de evitar este problema es empleando un valor de regularización, también conocido como tasa de aprendizaje o learning rate (α), que limite la influencia de cada modelo en el conjunto de ensamble. Como consecuencia de esta regularización, se necesitan más modelos para formar el ensamble final, pero se consiguen mejores resultados. Los valores de α pueden ir en un rango de 0 a 1.

$$F_1 \approx Y$$
$$F_2 \approx Y - F_1$$
$$F_3 \approx Y - F_2 - F_1$$
$$Y \approx \alpha F_1 + \alpha F_2 + \alpha F_3 + ... + \alpha F_N$$

En la Fig.3.3 se observa una representación de la estructura de árboles usando aumento del gradiente. Cada árbol fue construido secuencialmente, y la predicción final es obtenida de la sumatoria de los resultados obtenidos en cada árbol, multiplicado por una tasa de aprendizaje α.

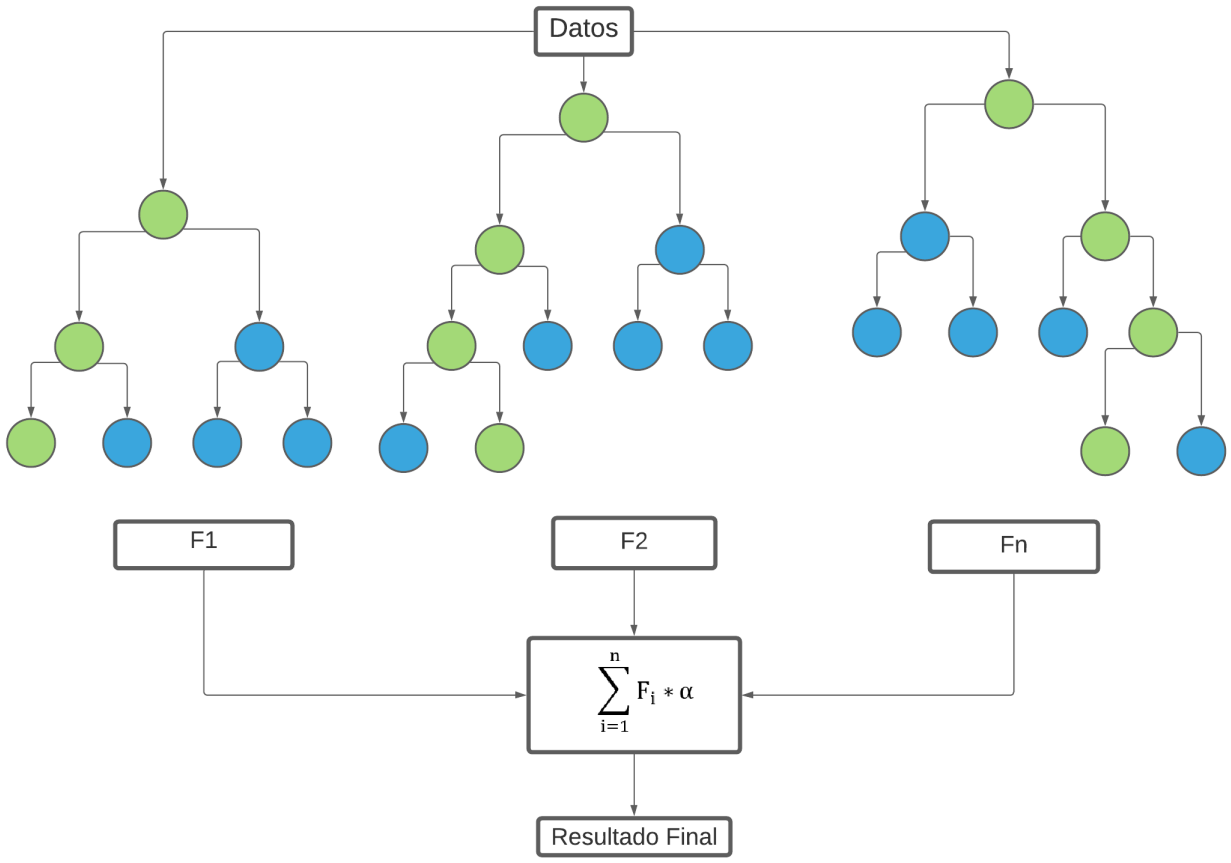


Figura 3.3: Representación del algoritmo XGBoost

Los árboles construidos tanto con Random Forest como en XGBoost se basan en la creación de nodos a partir de las mejores características que dividan al conjunto de datos, sin embargo, XGBoost no usa el cálculo del MSE como criterio de selección, sino que se utiliza una medición de Ganancia, obtenida a partir de la puntuación de similitud (similarity score) de la partición (nodo izquierdo y derecho) creada por la característica X (ecuación 3.4), donde la ganancia mayor indica la mejor partición.

$$Gain = Izquierdo_{\text{similitud}} + Derecho_{\text{similitud}} - Raiz_{\text{similitud}} \tag{3.4}$$

La puntuación de similitud es calculada como la suma de los residuales al cuadrado, dividido entre el número de residuales más lambda "λ" (ecuación 3.5). Lambda es un hiperparametro de regularización que reduce la sensibilidad de las predicciones a las observaciones individuales, resultando en la poda de más nodos en un árbol.

$$Similarity\ score = \frac{\sum_{i=1}^{n} Residual_i}{N + \lambda} \tag{3.5}$$

3.3.1. Implementación y resultados

La implementación de XGBoost se realizó utilizando la librería xgboost. XGBRegressor de Python. Para la implementación de los modelos se experimentaron con diferentes valores del siguiente conjunto de hiperparametros:

- n_estimators: 300, 600, 800, 1000
- learning_rate: 0.01, 0.1
- min_child_weight: 1, 6, 10
- max_depth: 6, 20

Debido a que este algoritmo parte de la creación de árboles de decisión al igual que Random Forest, muchos de sus hiperparametros comparten el mismo significado, tales como n_estimators, max_depth y min_child_weight, que controla las divisiones de los nodos, sin embargo, XGBoost implementa parámetros adicionales tales como: learning_rate, que controla la influencia de las predicciones realizadas por cada árbol, donde, valores pequeños aumentan el tiempo de entrenamiento, así como la cantidad de árboles necesarios, pero el modelo resultante es más robusto y eficiente.

En la tabla 3.2 se muestra el resultado de las predicciones realizadas en el mes de abril para los años 2019, 2020 y 2021, en el SIN y sus regiones, utilizando el algoritmo XGBoost.

Región	**2019**			**2020**			**2021**		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
Centro	552.99	424.12	13.07	**356.83**	**253.12**	**4.23**	387.25	291.71	4.67
Noreste	1105.20	968.91	8.75	**525.59**	**423.91**	**8.27**	997.59	806.74	13.91
Noroeste	159.70	124.05	5.04	181.33	141.00	6.36	**147.66**	**122.25**	**4.84**
Norte	313.05	259.02	8.74	**220.93**	**183.10**	**6.12**	315.99	251.38	7.50
Occidente	949.07	786.68	14.52	**564.89**	**458.99**	**6.41**	731.62	620.58	7.58
Oriente	796.69	629.59	9.44	**432.48**	**336.30**	**5.79**	458.91	383.28	6.19
Peninsular	**109.41**	**89.94**	**11.72**	144.58	121.21	8.47	168.10	145.48	9.38
SIN	1935.60	1478.77	4.50	**1660.21**	**1305.86**	**4.26**	2209.10	1569.73	4.74

Cuadro 3.2: Errores de predicción usando XGBoost

Los pronósticos obtenidos en cada una de las regiones muestran en su mayoría que, en el año 2020 se obtuvieron mejores predicciones. Una excepción para esto, se encuentran en la región Peninsular, donde las predicciones más acertadas fueron realizadas en el año 2019. Sin embargo, los resultados muestran que XGBoost se adecuó mejor al comportamiento de la demanda eléctrica durante los años de pandemia.

3.4. Support Vector Machine

Las máquinas de vectores soporte (SVM, del inglés Support Vector Machines) tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico introducidos en los años 90 por Vapnik y sus colaboradores [17]. Aunque originariamente las SVM fueron pensadas para resolver problemas de clasificación binaria, se pueden utilizar para resolver diversos tipos de problemas, por ejemplo, la regresión.

El objetivo de las SVM es encontrar un hiperplano en un espacio de n dimensiones que separe de forma equidistante a los puntos de datos más cercanos de cada clase, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano solo se consideran los datos que están en la frontera del margen. Estos ejemplos reciben el nombre de vectores de soporte. De esta manera los puntos del vector que están de un lado del hiperplano se etiquetan con una categoría y los que se encuentran del otro lado, se etiquetan con otra categoría.

Entonces, el problema a resolver es la optimización del margen geométrico alrededor del hiperplano separador, el cual es un problema de optimización geométrica que se puede escribir como un problema de optimización cuadrático convexo con restricciones lineales, cuya resolución nos garantiza una solución única.

Para comprender mejor este concepto se expone un ejemplo sencillo de clasificación binaria, para después introducir la regresión.

3.4.1. SVM para Clasificación

SVM al ser un clasificador linear implica que el hiperplano solución se exprese de la siguiente manera:

$$0 = W^T X_i + b$$

Donde:

W es el vector ortogonal al hiperplano

b es el coeficiente de intercepción

En la Fig.3.4 inciso (A) se observa un ejemplo de un hiperplano que separa en dos clases al conjunto de datos, sin embargo, este no suele ser único. En el inciso (B) se observa que existe más de un plano separador, y la cuestión es ¿Qué plano es el que mejor separa los datos?

La selección de un hiperplano de entre todos los posibles hiperplanos de separación se realizará a partir del concepto de margen, que se define como la distancia mínima entre dicho hiperplano y el ejemplo más cercano a cada clase, denotado por d.

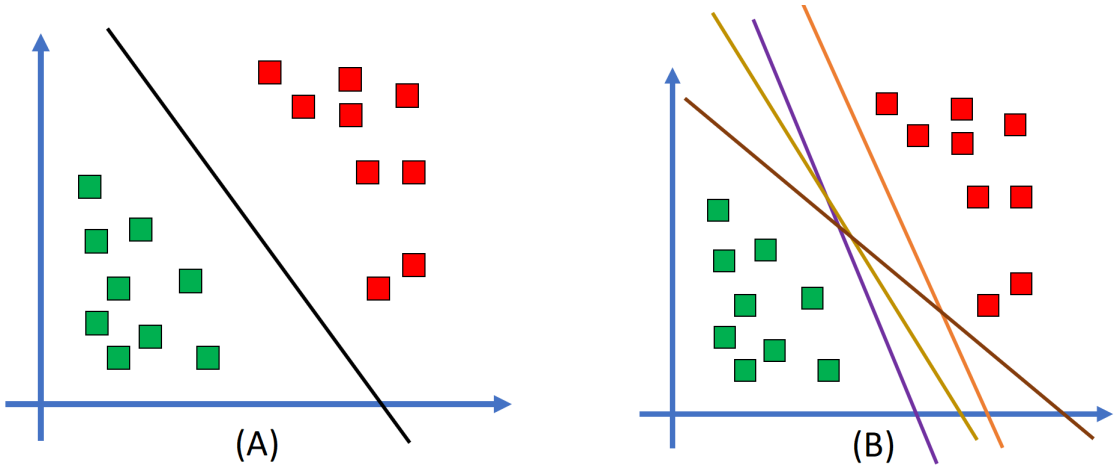


Figura 3.4: Hiperplanos de clasificación binaria

En la Fig.3.5 inciso (a) se observa un margen de separación no óptimo, frente al inciso (b) el cual es un margen de separación máximo. Los puntos de datos sobre los que atraviesa el margen superior e inferior son considerados vectores de soporte, siendo que, el hiperplano de separación se construye como combinación lineal de los dos vectores de soporte del conjunto de ejemplos (uno de cada clase).

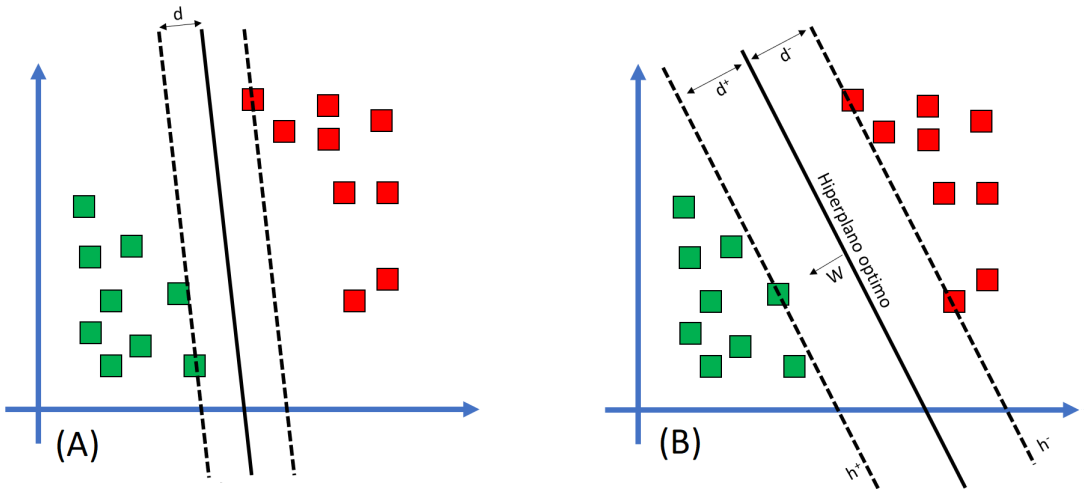


Figura 3.5: Selección de los márgenes e hiperplano de separación optimo

Este margen puede ser calculado como la diferencia entre los dos hiperplanos que contienen los vectores de soporte (h^+, h^-), denotado por la siguiente expresión:

$$d^+ = d^- = \frac{|WX + b|}{||W||} = \frac{1}{||W||}$$

$$Margen = d^+ + d^- = 2\frac{1}{||W||}$$

De la expresión anterior se deduce que encontrar el hiperplano óptimo, es equivalente a encontrar el valor de w que maximiza el margen, sin embargo, al reescribir esta expresión como el problema inverso obtenemos.

$$Margen = Minimizar\ \Phi(W) = \frac{1}{2}||W||^2$$

Como se puede ver, el margen únicamente depende de W y entonces la solución es minimizar Φ(w), sujeto a.

$$h^+ \rightarrow W^T X_i + b = +1$$
$$h^- \rightarrow W^T X_i + b = -1$$

Lo cual puede ser reescrito como:

$$Y_i(W^T X_i + b) \geq 1$$

Este problema de optimización se resuelve mediante multiplicadores de Lagrange, como esto no es nuestro objetivo y es una explicación un tanto tediosa, se omite.

El caso anterior es un caso ideal, en el que los datos pueden ser perfectamente separables, sin embargo, existen ocasiones donde no existe un hiperplano que pueda separar linealmente el conjunto de datos, tal como se muestra en la Fig.3.6 inciso (a).

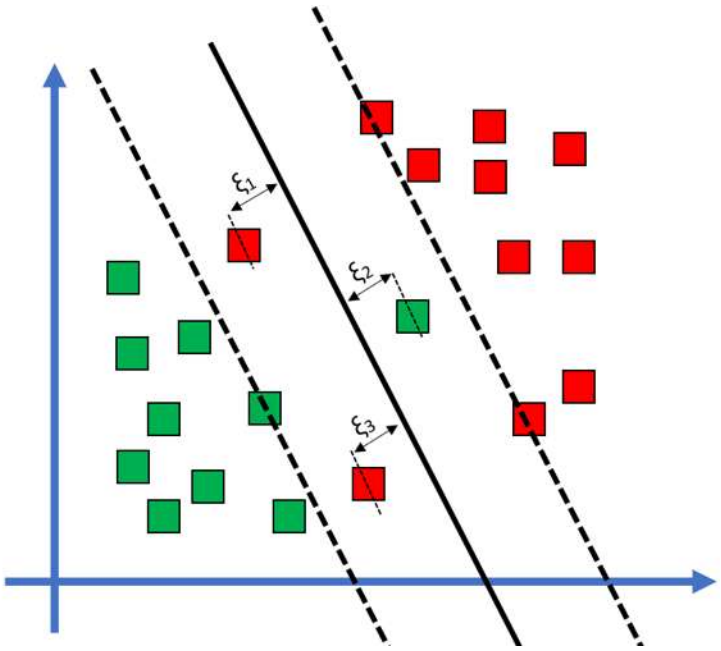


Figura 3.6: Caso Lineal no separable

La idea para tratar con este tipo de casos con ruido es introducir un conjunto de variables de holgura ξ, que controla el error permitido y que penalizan las muestras mal clasificadas, dicho de otra manera, permiten que algunos vectores se encuentren dentro del margen.

En este caso, el margen es calculado como:

$$Minimizar\ \Phi(W) = \frac{1}{2}||W||^2 + C\sum_{i}^{n}\xi_i$$

Donde: C es una constante elegida por el usuario, que permite cambiar la influencia del coste de los términos no separables.

Sujeto a:

$$Y_i(W^T X_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Este hiperplano recibe el nombre de hiperplanos de separación de "margen blando", mientras que el obtenido en el caso perfectamente separable se denomina de "margen duro".

Finalmente existen ocasiones donde no es posible encontrar un hiperplano que pueda separar linealmente el conjunto de datos, tal como se muestra en la Fig.3.7 inciso (A).

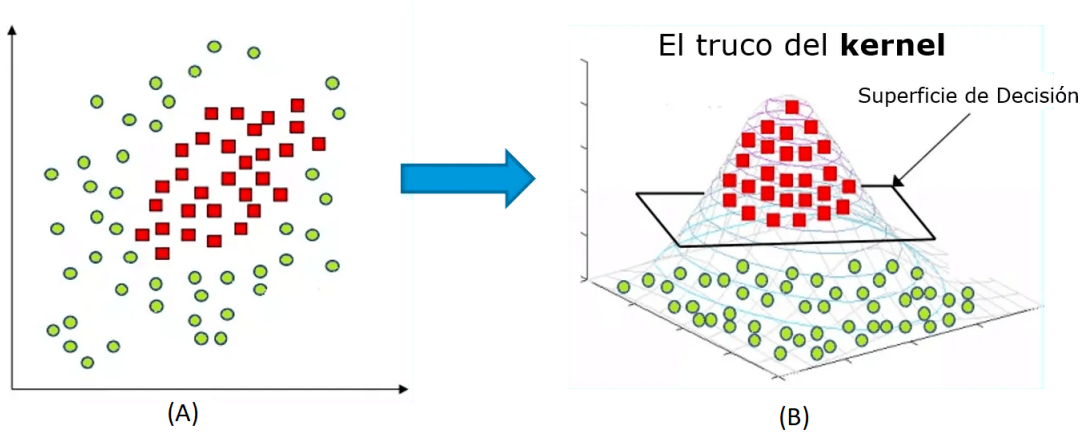


Figura 3.7: Aplicación de una función Kernel SVM [18]

Para resolver este problema, existe la posibilidad de transformar los datos a un espacio de mayor dimensión utilizando una función de transformación X $\rightarrow$ Φ(X), donde se encontrará un hiperplano que pueda separar el conjunto de datos. A esta función de transformación se denomina Kernel (K), la cual emplea "el truco del kernel". Este consiste calcular la relación de cada par de puntos en una dimensión más alta, sin transformarlos directamente, lo cual reduce la cantidad de cálculos requeridos, evitando las transformaciones matemáticas de los datos a mayores dimensiones.

En la Fig.3.7 inciso (B) se observa una representación gráfica del aumento en una dimensión del conjunto de datos (3D), lo cual permite encontrar una superficie (2D) que separa las dos clases.

Algunos de los kernels más usados son:

- Kernel Lineal:

$$K(X_i, X_j) = X_i \cdot X_j$$

- Kernel polinomial:

$$K(X_i, X_j) = (X_i \cdot X_j + 1)^d$$

- Kernel RBF:

$$K(X_i, X_j) = exp(-\gamma||X_i - X_j||^2),\ \gamma 0$$

- Kernel Sigmoide:

$$K(X_i, X_j) = tanh(b(X_i, X_j) + c)$$

3.4.2. SVM para regresión

Como se mencionó en un principio, SVM también se puede utilizar como método de regresión, manteniendo todas las características principales que caracterizan al algoritmo (margen máximo). La regresión de vectores de soporte (SVR) utiliza los mismos principios que SVM para la clasificación, con solo algunas diferencias menores.

En este caso, la idea es seleccionar el hiperplano regresor que mejor se ajuste a nuestro conjunto de datos de entrenamiento. Ahora no disponemos de clases para separar. La idea se basa en considerar una distancia margen ε, de modo que esperamos que todos los datos se encuentren en una banda o tubo en torno a nuestro hiperplano, es decir, que los datos estén una cantidad menor de ε del hiperplano Fig.3.8. A la hora de definir el hiperplano sólo se consideran los datos que estén fuera del tubo, ósea, a más de ε. En este caso esos datos serán los considerados como vectores soporte.

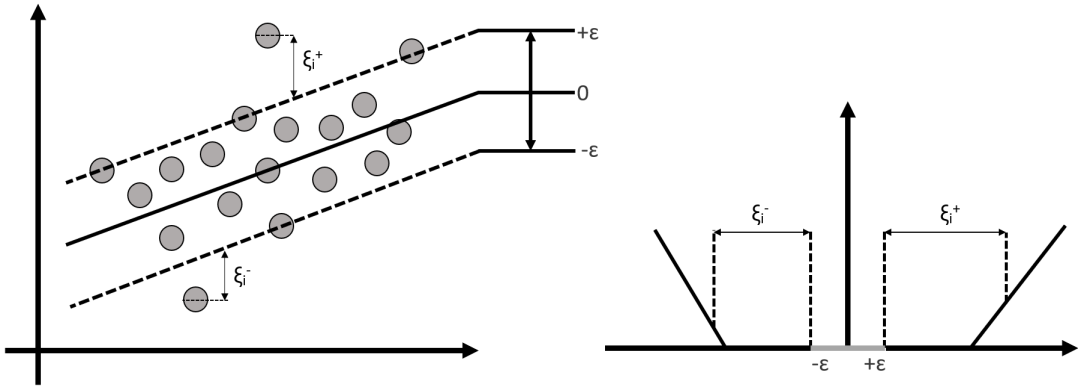


Figura 3.8: Datos englobados alrededor del tuvo

Lo que estamos tratando de hacer aquí es básicamente tratar de determinar un límite de decisión a una distancia ε del hiperplano original, de modo que los puntos de datos que disten exactamente ε de nuestro hiperplano se denominaran vectores de soporte.

Por lo tanto, el límite de decisión es nuestro margen de tolerancia, es decir, vamos a tomar solo aquellos puntos que están dentro de este límite. O en términos simples, vamos a tomar solo aquellos puntos que tienen la menor tasa de error. Dándonos así un modelo de mejor ajuste.

Para encontrar los límites de tolerancia, el problema de optimización que hay que resolver, es el siguiente:

$$Minimizar\ \Phi(W) = \frac{1}{2}||W||^2 + C\sum_{i}^{n}(\xi_i^+ + \xi_i^-)$$

Sujeto a:

$$Y_i - W^T X_i - b \leq \varepsilon + \xi_i^+$$
$$W^T X_i + b - Y_i \leq \varepsilon + \xi_i^-$$
$$\xi_i^-, \xi_i^+ \geq 0$$

Los datos que están fuera del límite de decisión (ξ) son considerados errores y tienen una penalización adicional C, la cual nosotros la determinamos.

Al igual que para la clasificación, en regresión, cuando un problema no es separable linealmente aplicamos un kernel, que mapear los puntos a una mayor dimensionalidad en la que si se pueda hacer el ajuste lineal y luego la solución dada se mapea de regreso al espacio original, tal como se muestra en la Fig.3.9

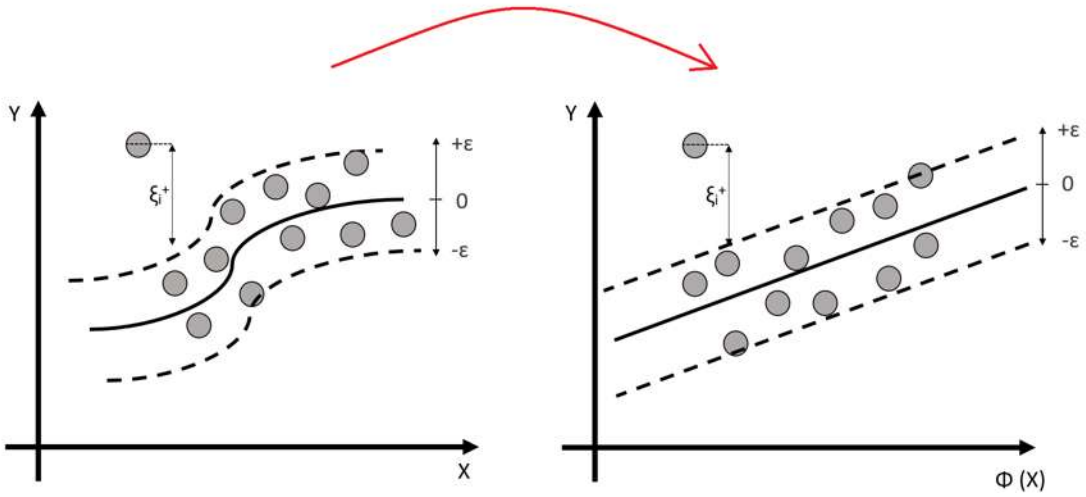


Figura 3.9: Aplicación de una función Kernel en SVR

3.4.3. Implementación y resultados

La implementación de SVR se realizó utilizando la librería sklearn.svm.SVR de Python. Para la implementación de los modelos se experimentaron con diferentes valores del siguiente conjunto de hiperparametros:

- kernel: lineal, rbf
- epsilon: 0.1, 0.001
- C: 0.1, 1

El hiperparametro Kernel especifica el tipo de kernel que se utilizara el algoritmo para transformar los datos a un espacio de mayor dimensión. Épsilon define él tuvo ε dentro del cual no se asocia ninguna penalización. C es un hiperparametro de regularización, que define la tolerancia de los errores fuera del tubo ε, valores pequeños permiten un número elevado errores, mientras que valores altos toleran menos errores.

En la tabla 3.3 se muestra el resultado de las predicciones realizadas en el mes de abril para los años 2019, 2020 y 2021, en el SIN y sus regiones, utilizando el algoritmo SVR.

Región	**2019**			**2020**			**2021**		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
Centro	586.58	429.95	14.28	**360.95**	**247.04**	**4.18**	523.09	447.37	6.74
Noreste	1106.56	987.26	9.07	**700.64**	**565.68**	**11.10**	839.02	723.71	11.80
Noroeste	**170.90**	**134.28**	**5.49**	259.28	212.52	9.97	187.24	147.76	5.71
Norte	333.08	270.21	8.72	**290.99**	**231.30**	**7.48**	379.00	300.64	9.22
Occidente	900.72	716.28	12.04	**658.79**	**493.75**	**7.22**	1422.89	1341.69	16.62
Oriente	767.50	607.36	8.99	**345.33**	**263.92**	**4.50**	464.90	388.51	6.23
Peninsular	146.20	121.19	15.79	**144.08**	**110.64**	**7.49**	202.26	160.93	10.64
SIN	2348.92	1811.58	5.42	**1742.12**	**1366.97**	**4.53**	2291.26	1647.89	5.07

Cuadro 3.3: Errores de predicción usando SVR

Los errores obtenidos en cada una de las regiones muestran en su mayoría que, durante la pandemia se realizaron mejores predicciones, más específicamente en el 2020. La excepción se encuentra en la región Noroeste, donde las predicciones más acertadas fueron realizadas en el año 2019.

De manera general, los resultados muestran que SVR se adecuó mejor al comportamiento de la demanda eléctrica durante los años de pandemia.

3.5. K Nearest Neighbor

El algoritmo de K vecinos más cercanos (KNN, por sus siglas en ingles), fue desarrollado por Evelyn Fix y Joseph Hodges en 1951 [19] a partir de la necesidad de realizar un análisis discriminante cuando la distribución de probabilidad de los datos es desconocida o difícil de determinar. Esto indica que KNN es un algoritmo no paramétrico, es decir, no presupone ninguna distribución de probabilidad en los datos.

Originalmente, KNN fue pensado para la clasificación, pero, también puede ser utilizado para resolver problemas de regresión. En ambos casos, dado un nuevo ejemplo, el objetivo de KNN es encontrar los K ejemplos más similares, llamados vecinos más cercanos, de acuerdo con una métrica de distancia, como la euclidiana. Luego, para el caso de la clasificación, el nuevo ejemplo es clasificado con base en la clase más frecuente de los K ejemplos seleccionados, mientras que, para la regresión, el valor final se obtiene como una agregación de los valores objetivo de sus vecinos más cercanos.

Como un ejemplo ilustrativo, consideremos el caso más simple de utilizar un modelo KNN como un regresor, en el que, los datos de entrenamiento constan de la variable objetivo y solamente una característica, de manera que pueden verse representados en un plano en dos dimensiones, tal como se muestra en la figura 3.10, donde el eje vertical Y representa el valor objetivo, y el eje horizontal X representa el valor de la característica.

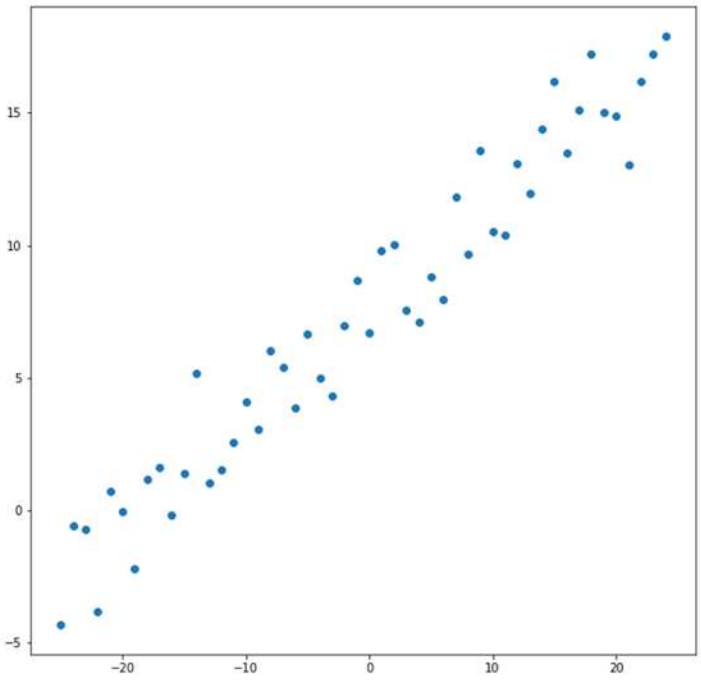


Figura 3.10: Datos de entrenamiento

Considerando la introducción de un nuevo ejemplo, cuya característica tiene un valor de 6.5. Al tener un valor de K igual a cuatro, KNN selecciona los cuatro ejemplos con características más similares a 6.5, y posteriormente, el punto resultante se obtiene como el promedio de los ejemplos seleccionados, 3.11.

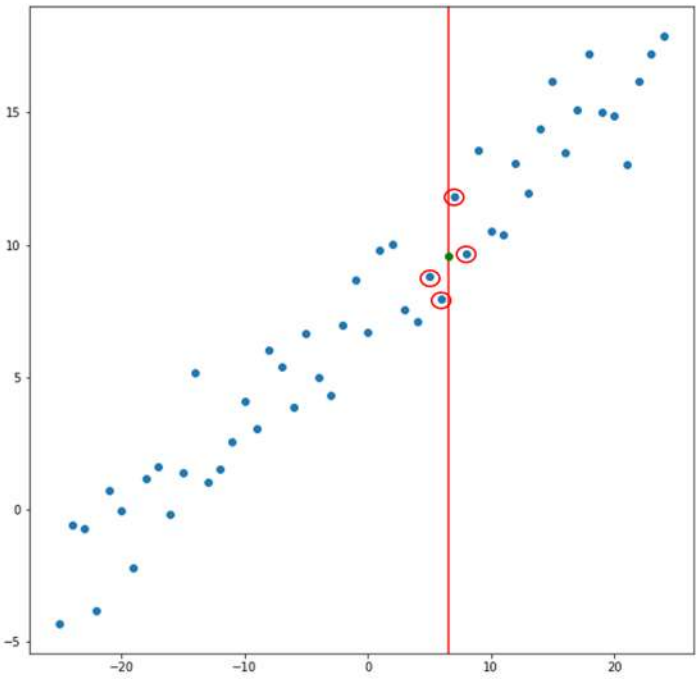


Figura 3.11: Selección de los vecinos más cercanos

Para casos más complejos, donde el número de características es mayor a uno, cada uno de los i-esimos ejemplos de entrenamiento contienen una colección de N características ($C_1{}^i$, $C_2{}^i$, $C_3{}^i$, ... , $C_n{}^i$), las cuales, están asociadas a un valor objetivo ti. Entonces, dado un nuevo ejemplo, cuyas características son conocidas (P_1, P_2, P_3, ... , P_n) pero el objetivo es desconocido, las características del nuevo ejemplo son utilizadas para encontrar los K ejemplos más similares de acuerdo con los vectores de características y una métrica de similitud o distancia. Por ejemplo, suponiendo que la métrica de similitud es la distancia euclidiana, la distancia entre el nuevo ejemplo y el i-esimo ejemplo de entrenamiento se calcula de la siguiente manera:

$$\sqrt{\sum_{x=1}^{n}(C_x^i - P_x)^2}$$

De esta forma, los K ejemplos de entrenamiento que están más cerca del nuevo ejemplo son considerados sus K vecinos más cercanos, y suponiendo que sus objetivos son los vectores t_1, t_2, t_3, ... , t_k, se pueden promediar para obtener el valor del nuevo ejemplo.

En contraste con otros algoritmos de ML, KNN no genera un modelo fruto del aprendizaje con datos de entrenamiento, sino que, simplemente almacena una colección de todos los ejemplos de entrenamiento, y el aprendizaje sucede en el mismo momento en el que se predicen nuevos datos.

3.5.1. Implementación y resultados

La implementación de KNN se realizó utilizando la librería *sklearn.neighbors.KNeighborsRegressor* de Python. Para la implementación de los modelos se experimentaron con diferentes valores del siguiente conjunto de hiperparametros:

- n_neighbors: 3, 5, 6, 10, 20, 40, 80, 160, 200
- weights: uniform, distance

El hiperparametro n_neighbors especifica el número de vecinos más cercanos a considerar. Un número pequeño de K pude llevar a un sobreajuste, mientras que un número elevado puede llevar a predicciones erróneas. weights establece una función de peso en cada uno de los K vecinos, para que estos sean tomados en consideración por igual, u otorgar una mayor influencia a los vecinos más cercanos.

En la tabla 3.4 se muestra el resultado de las predicciones realizadas en el mes de abril para los años 2019, 2020 y 2021, en el SIN y sus regiones, utilizando el algoritmo KNN.

Región	**2019**			**2020**			**2021**		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
Centro	585.80	411.57	13.43	670.52	556.12	9.19	**404.30**	**263.35**	**4.26**
Noreste	1132.00	999.50	9.05	**748.37**	**604.61**	**11.35**	1189.58	1070.27	16.51
Noroeste	226.42	178.69	7.07	278.25	192.62	8.21	**200.08**	**156.53**	**6.44**
Norte	342.21	269.65	9.17	**296.78**	**230.49**	**7.63**	350.08	299.69	8.80
Occidente	1326.72	1033.48	15.93	973.74	796.81	11.45	**858.99**	**669.93**	**8.50**
Oriente	828.10	650.54	9.86	**382.88**	**284.02**	**5.07**	517.32	453.61	7.21
Peninsular	183.02	147.55	18.08	**107.319**	**83.36**	**5.90**	184.77	141.99	9.49
SIN	2636.65	2229.85	6.51	**2286.87**	**1836.98**	**6.05**	2973.22	2584.65	7.52

Cuadro 3.4: Errores de predicción usando KNN

Los errores de predicción obtenidos en cada una de las regiones mostraron que las mejores predicciones fueron obtenidas durante los años 2020 y 2021, mostrando que KNN se adecuó mejor al comportamiento de la demanda eléctrica durante estos años.

Comentarios finales

En este capítulo se mostró la implementación entrenamiento de cada uno de los sistemas de predicción, así como también se dio una descripción del funcionamiento general de los algoritmos usados para su implementación.

De manera general, las predicciones obtenidas para el mes de abril en los años 2019, 2020 y 2021 mostraron que durante la pandemia se realizaron pronósticos más acertados de la demanda energética, esto de acuerdo con los errores RMSE, MAE y MAPE obtenidos al evaluar la eficiencia de cada uno de las modelos de predicción.

Capítulo 4

Pruebas y análisis de resultados

El presente capítulo se encuentra dividido en tres partes. En la primera parte se realiza un ensamble de los modelos de predicción obtenidos en el capítulo anterior, con el objetivo de mejorar las predicciones individuales de estos modelos. En la segunda parte se realiza una comparación de eficiencia entre los modelos obtenidos en el capítulo 3 y el ensamble de estos. En la tercera parte se analizan los resultados de las predicciones realizadas en cada región para los años 2019, 2020 y 2021, tomando en cuenta los modelos de mayor precisión.

4.1. Ensamble de los predictores

Como se mencionó anteriormente, el ensamble es una técnica que combina varios modelos con tal de producir un modelo de predicción óptimo, tal como se vio para los algoritmos de Random Forest y XGBoost en donde las predicciones finales son realizadas tomando en cuenta la predicción individual de cada árbol de regresión.

La ventaja que tienen estos métodos de ensamble es que, al combinar las predicciones individuales de varios modelos estas tienden a generalizarse, de manera que los pronósticos erróneos realizados por un solo modelo pueden ser suavizados debido al peso que ejercen los demás pronósticos.

Por lo tanto, en esta sección se implementa un ensamble de los modelos evaluados y seleccionados en el capítulo 3. La implementación se lleva a cabo utilizando el método de Bagging, mismo que es utilizado por Random Forest. De esta forma, las predicciones finales realizadas por este ensamble son obtenidas a partir del promedio de las predicciones individuales de cada uno de los modelos entrenados con los algoritmos Random Forest, XGBoost, SVR y KNN.

En la tabla 4.1 se muestra el resultado de las predicciones realizadas en el mes de abril para los años 2019, 2020 y 2021, en el SIN y sus regiones, utilizando el ensamble de predictores.

De acuerdo con la tabla de errores, los mejores pronósticos se obtuvieron durante los años de pandemia, más específicamente en el 2020. La excepción se encuentra en la región Noroeste, donde las predicciones más acertadas fueron realizadas en el año 2019.

Región	2019			2020			2021		
	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
Centro	538.73	390.14	12.67	**370.62**	**265.17**	**4.46**	370.09	294.00	4.64
Noreste	1139.52	1016.47	9.18	**554.91**	**450.28**	**8.66**	820.26	715.76	11.44
Noroeste	**165.87**	**131.18**	**5.26**	218.88	152.11	6.61	166.62	136.01	5.34
Norte	303.66	251.21	8.52	**289.66**	**238.59**	**7.60**	361.89	311.50	8.95
Occidente	1004.40	802.44	12.94	**712.08**	**634.69**	**8.74**	908.91	842.20	9.79
Oriente	774.84	622.39	9.39	**363.79**	**277.68**	**4.89**	458.97	392.81	6.29
Peninsular	135.09	114.55	14.29	**114.26**	**92.75**	**6.54**	181.88	147.70	9.68
SIN	2292.70	1911.41	5.64	**1460.02**	**1157.19**	**3.72**	2894.48	2563.52	7.30

Cuadro 4.1: Errores de predicción usando el ensamble de predictores

Sin embargo, los resultados muestran que el ensamble de predictores se adecuó mejor al comportamiento de la demanda eléctrica durante los años de pandemia.

4.2. Comparación de la eficiencia entre algoritmos

A pesar de que los datos en cada una de las regiones describen el mismo fenómeno que es la demanda eléctrica, su comportamiento y en especial, la forma en la que la pandemia afectó a cada una de estas regiones no fue idéntica, por lo que no siempre la misma técnica es la que mejor se adecua al comportamiento de los datos en todas las ocasiones.

Es por esto que se hizo una comparación entre las técnicas usadas para la implementación de los predictores. En la tabla 4.2, se muestra, para cada año y región, el algoritmo con el que se construyó el mejor predictor.

Región	**Año**	**Algoritmo**
Centro	2019	Ensamble
	2020	XGBoost
	2021	Ensamble
Noreste	2019	XGBoost
	2020	XGBoost
	2021	Ensamble
Noroeste	2019	XGBoost
	2020	XGBoost
	2021	XGBoost
Norte	2019	Ensamble
	2020	XGBoost
	2021	XGBoost
Occidental	2019	SVM
	2020	XGBoost
	2021	XGBoost
Oriental	2019	SVM
	2020	SVM
	2021	XGBoost
Peninsular	2019	XGBoost
	2020	KNN
	2021	XGBoost
Sistema Interconectado	2019	XGBoost
	2020	Ensamble
	2021	XGBoost

Cuadro 4.2: Mejores Algoritmos

Estas comparaciones se basaron respecto de las métricas RMSE y MAE. De acuerdo con los resultados, se observa que tanto XGBoost como el Ensamble de predictores tuvieron un mejor desempeño que los algoritmos SVR, KNN y Random Forest, sin embargo, en algunos casos SVR y KNN superaron tanto a XGBoost como el Ensamble.

4.3. Análisis de los pronósticos

A continuación, se realiza un análisis de los resultados obtenidos, tomando en cuenta los modelos que mejores pronósticos realizaron en cada región, tabla 4.2. Para ello se muestran las gráficas del comportamiento real y pronosticado en cada uno de los años de predicción, así como también las métricas del error correspondientes. Esto con el objetivo de exponer las características de la demanda real y pronosticada, antes y durante la pandemia en cada una de las regiones del sistema interconectado.

El hecho de considerar los mejores modelos implica que los predictores seleccionados pueden no estar implementados con el mismo algoritmo. Así evitamos seleccionar un solo tipo de modelo que favorezca las predicciones para alguna región o año en particular. De esta manera las comparaciones se realizan en igualdad de condiciones.

4.3.1. Región Central

De acuerdo con cada una de las métricas del error, en la región centro los mejores pronósticos se obtuvieron en el año 2020, seguido por el 2021 y, dejando al 2019, como el año donde se obtuvieron los pronósticos menos acertados.

Año	**Error**			**Algoritmo**
	RMSE	MAE	MAPE (%)	
2019	538.73	390.14	12.67	Ensamble
2020	**312.90**	**215.46**	**3.63**	**XGBoost**
2021	370.09	294.00	4.64	Ensamble

Cuadro 4.3: Errores de predicción en la región Central

Esto se ve representado en la Figura 4.1, en la que se muestran las gráficas de los valores reales (azul) contra los predichos (naranja) de la demanda energética del mes de abril, en cada uno de los años de predicción.

Idealmente ambas gráficas deberían estar superpuestas, esto para que el error de predicción fuera cero, sin embargo, es algo imposible de lograr en la vida real, por lo que siempre habrá algún margen de error, el cual, se ve representado por las diferencias en las gráficas la demanda real y pronosticada.

La principal característica al comparar la demanda real en cada uno de los años, es un evidente cambio en su comportamiento, que es muy similar durante el 2020 y 2021, lo cual, es un indicativo de que los efectos de la pandemia continuaron estando presentes durante este último año.

Otra de las características que se presenta en esta región, es la estabilidad en los patrones de comportamiento de la demanda durante los años de pandemia. Y es que, el comportamiento de la demanda durante el 2019 tiene una mayor variación a comparación del 2020 y 2021, años en los cuales, se mantiene un patrón de comportamiento constante a lo largo de todo el mes.

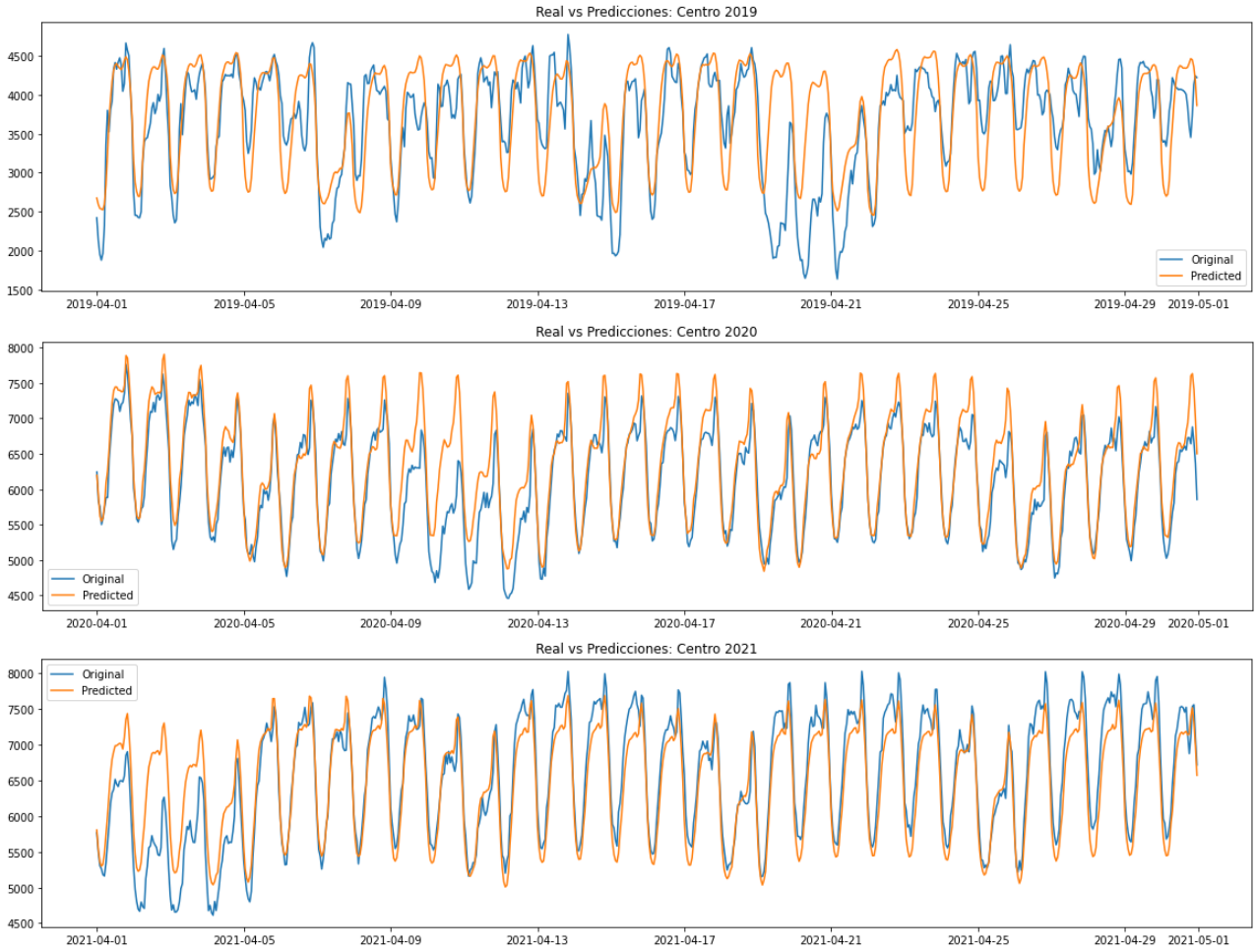


Figura 4.1: Errores de predicción en la región Central

Cabe mencionar que esta última característica no es una sorpresa, ya que el análisis de volatilidad realizado en el capítulo 2, se mencionó que el coeficiente de variación para esta región, se vio reducido con la llegada de la pandemia.

4.3.2. Región Noreste

De acuerdo con las métricas del error, fue durante la pandemia donde se realizaron los mejores pronósticos de la demanda, especialmente en el año 2020, seguido por el 2021 y finalmente el 2019. Esto de acuerdo a las métricas RMSE y MAE.

La razón por la que la métrica MAPE difiere de acuerdo a lo anterior mencionado, se debe a que sus resultados dependen de la magnitud en la que se encuentren los valores reales, tal como se mencionó en la sección 4.4, y es que, esta región presento una disminución en la demanda durante los años de pandemia.

Año	Error			Algoritmo
	RMSE	MAE	MAPE (%)	
2019	1105.20	968.91	8.75	XGboost
2020	**525.59**	**423.91**	**8.27**	**XGboost**
2021	820.26	715.76	11.44	Ensamble

Cuadro 4.4: Errores de predicción en la región Noreste

Figura 4.2: Errores de predicción en la región Noreste

Al ver la figura 4.2, Se hace evidente que los predictores tuvieron dificultad para pronosticar con exactitud la demanda real, principalmente en los años 2019 y 2021, dando como resultado predicciones por debajo de los valores reales, especialmente en el 2019. Y es que, una de las principales características que presento esta región, es la inexistencia de un patrón de comportamiento constante, haciendo que los predictores sean incapaces de encontrar relaciones que describan con exactitud el comportamiento de los datos.

Otra característica es que, a diferencia de la región centro, el comportamiento real de la demanda en los años 2020 y 2021 no son muy similares, mostrando que, el efecto de la pandemia sobre este último año, fue diferente al presentado en el 2020.

4.3.3. Región Noroeste

De acuerdo con cada una de las métricas del error mostradas para esta región, fue en el año 2021 donde se obtuvieron los mejores pronósticos, sin embargo, al compararlos con los errores obtenidos en el 2019, se observa que estos son muy similares entre sí, por lo que prácticamente, la precisión de los pronósticos fue la misma durante estos años.

Año	**Error**			**Algoritmo**
	RMSE	MAE	MAPE (%)	
2019	159.70	124.05	5.04	XGBoost
2020	181.33	141.00	6.36	XGBoost
2021	**147.66**	**122.25**	**4.84**	**XGBoost**

Cuadro 4.5: Errores de predicción en la región Noroeste

Este no fue el mismo caso para el 2020, donde se muestra una diferencia más notoria en la magnitud de estos errores, por lo que, este fue el año con los peores pronósticos.

En el año 2019, el comportamiento de la demanda presenta una especie de ruido, sin embargo, la razón de ello es que, a nivel horario, existe una alta variación en la demanda energética, la cual afecta los pronósticos realizados, ya que estos no tuvieron la capacidad de predecir, a ese nivel de detalle, las variaciones presentadas.

En el caso de los años 2020 y 2021, este no es un problema, ya que las gráficas muestran una mayor suavidad y estabilidad en el comportamiento de los datos, sin embargo, un punto en contra, fue el incremento en la variación de los valores de la demanda, la cual, al igual que la estacionalidad, juega un papel importante en la precisión de los pronósticos realizados.

Y es que, la razón por la cual se realizaron los peores pronósticos durante el 2020, se debió al inesperado incremento tanto en rango como magnitud, que presentaron los datos en la última semana del mes, los cuales, claramente no pudieron ser pronosticados.

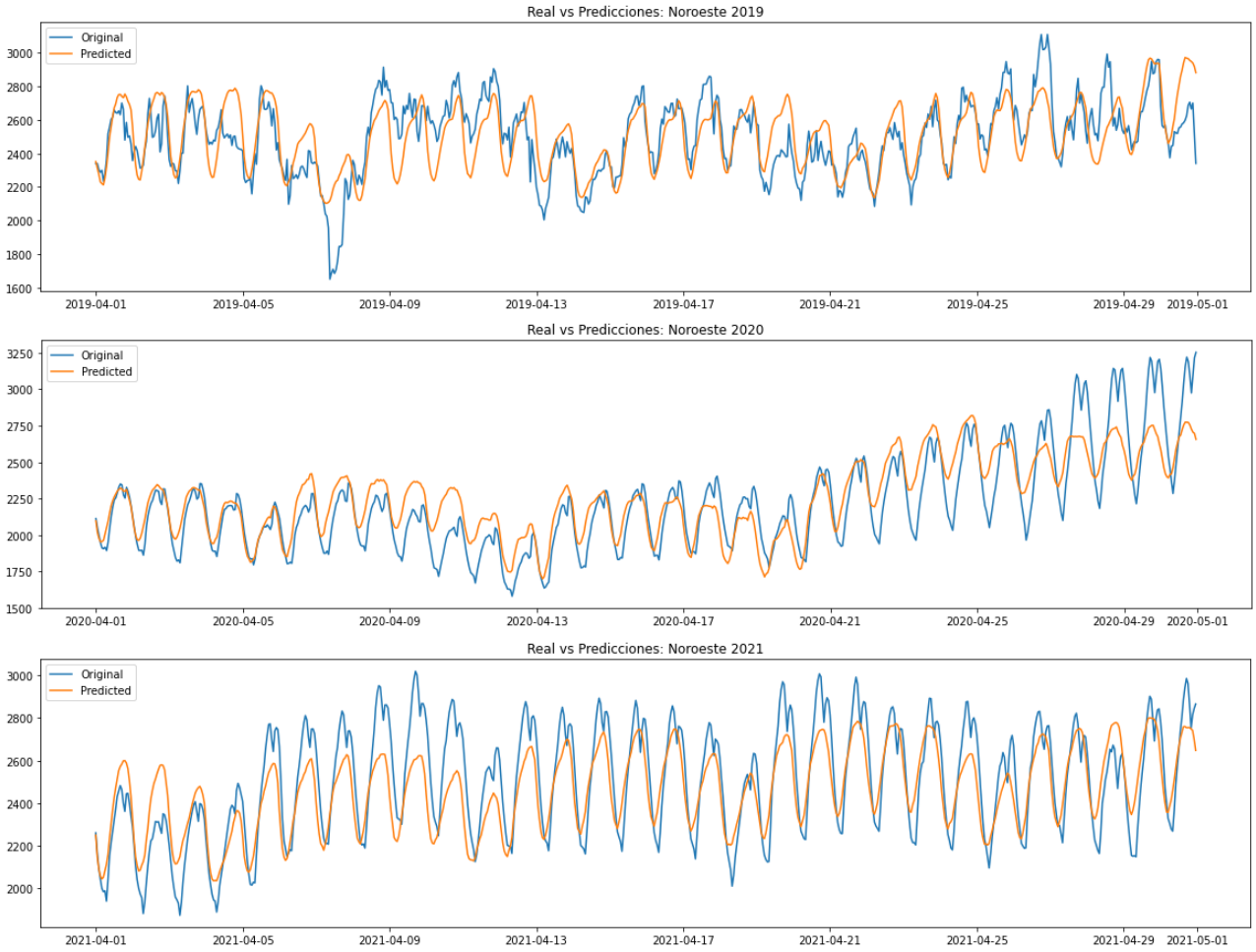


Figura 4.3: Errores de predicción en la región Noroeste

4.3.4. Región Norte

De acuerdo con las métricas del error, los mejores pronósticos fueron realizados en el 2020, seguido por el 2021, superando los pronósticos del 2019 por un uno por ciento de error, de acorde a la métrica MAPE.

Año	**Error**			**Algoritmo**
	RMSE	MAE	MAPE (%)	
2019	303.66	251.21	8.52	Ensamble
2020	**220.93**	**183.10**	**6.12**	**XGboost**
2021	315.99	251.38	7.50	XGboost

Cuadro 4.6: Errores de predicción en la región Norte

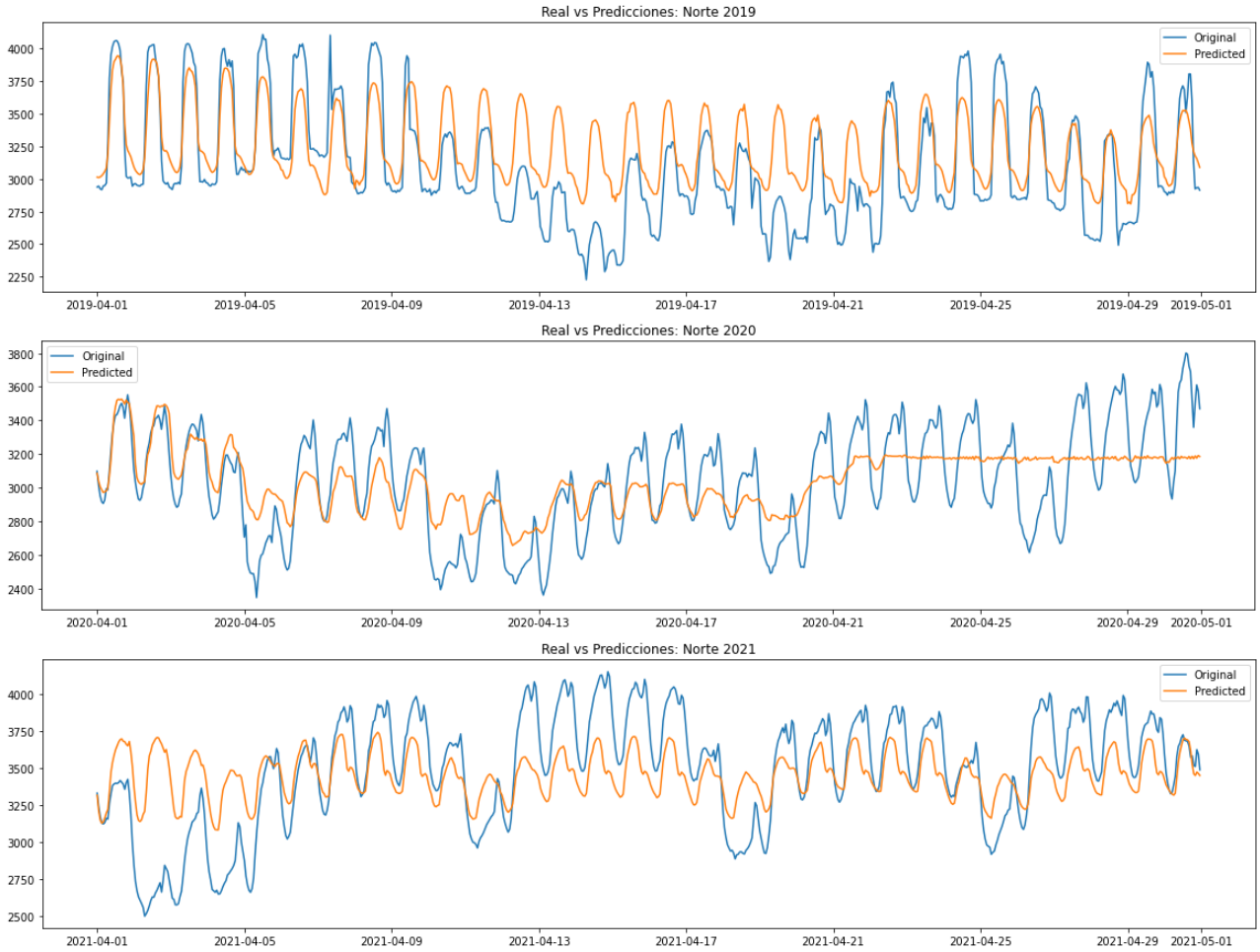


Figura 4.4: Errores de predicción en la región Norte

Similar a lo ocurrido en la región central y noroeste, los patrones de comportamiento de la demanda real en el 2019 no fueron constantes durante todo el mes, de manera que, los pronósticos realizados durante este año, fueron capaces de ajustarse al verdadero comportamiento de los datos debido a estas variaciones.

Sin embargo, a pesar de que el comportamiento de la demanda fue más constante durante la pandemia, los pronósticos no fueron capaces de ajustarse al verdadero patrón de los datos, pero a pesar de ello, los resultados mostraron un menor error durante estos años.

Cabe mencionar que tanto en el 2020 como 2021, se presentan comportamientos muy similares de la demanda real, por lo tanto, al igual que ocurrió en regiones como la centro, oriente y occidente, el efecto de la pandemia todavía continuó presente.

4.3.5. Región Occidental

Al evaluar la tabla de errores para la región occidental, se hace evidente que el año donde se realizaron los peores pronósticos de la demanda fue en el 2019, esto de acuerdo a cada una de las métricas. Sin embargo, al comparar los errores obtenidos en los años de pandemia, se observa que el 2020 presento menor error RMSE, pero mayor error MAE, respecto del 2021.

Año	**Error**			**Algoritmo**
	RMSE	MAE	MAPE (%)	
2019	900.72	716.28	12.04	SVR
2020	**564.89**	**458.99**	**6.41**	**XGBoost**
2021	731.62	620.58	7.58	XGBoost

Cuadro 4.7: Errores de predicción en la región Occidental

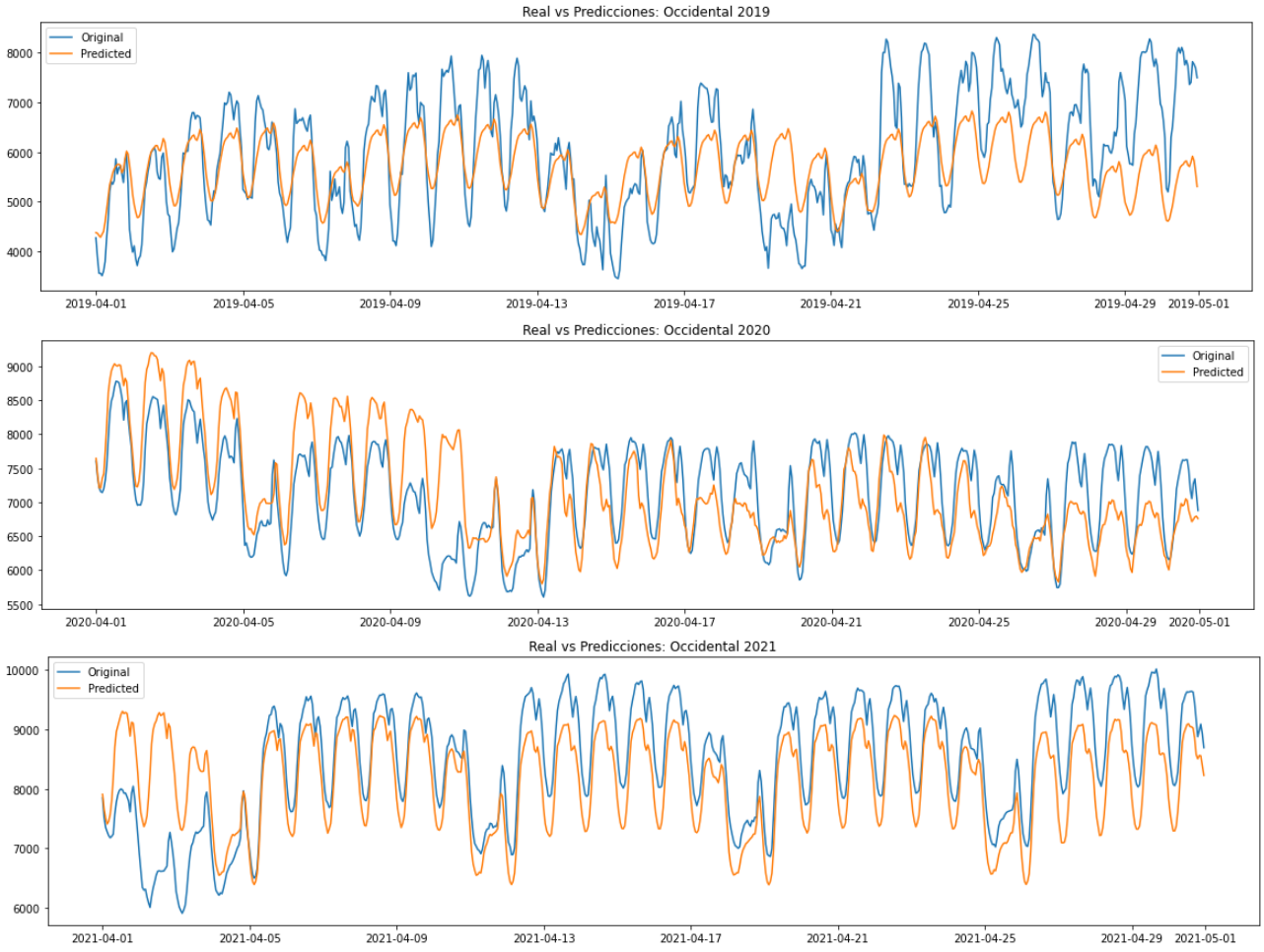


Figura 4.5: Errores de predicción en la región Occidental

De manera similar que la región centro, esta región presento características semejantes que influenciaron la precisión de los pronósticos en cada uno de los años presentados. Estas características se resumen de la siguiente manera:

- Mayor variación de la demanda energética en el 2019, respecto del 2020 y 2021.
- Patrones de comportamiento constantes durante los años de pandemia Cabe mencionar que el comportamiento de la demanda es similar en el 2020 y 2021, a excepción de la primera semana de este último año, sin embargo, es posible afirmar que los efectos de la pandemia todavía siguen presentes.

4.3.6. Región Oriental

De acuerdo con la tabla de errores, los mejores pronósticos de la demanda fueron realizados durante los años de pandemia, siendo el 2020, el año con las mejores predicciones, seguido por el 2021 y por último el 2019.

Año	**Error**			**Algoritmo**
	RMSE	MAE	MAPE (%)	
2019	767.50	607.36	8.99	SVR
2020	**345.33**	**263.92**	**4.50**	**SVR**
2021	458.91	383.28	6.19	XGBoost

Cuadro 4.8: Errores de predicción en la región Oriental

En la figura 4.6, además del evidente cambio en el comportamiento de los datos, se puede observar que una de las principales diferencias al comparar la demanda real el año 2019 frente al 2020 y 2021, es que, en estos dos últimos años, el patrón de comportamiento de los datos fueron constantes y muy similares entre sí, indicando que los efectos de la pandemia siguieron estando vigentes en este último año.

Cabe mencionar que estas características igualmente estuvieron presentes en las regiones centro y occidente, de manera que, de manera que, los efectos de la pandemia repercutieron de maneras específicas sobre algunas de estas regiones.

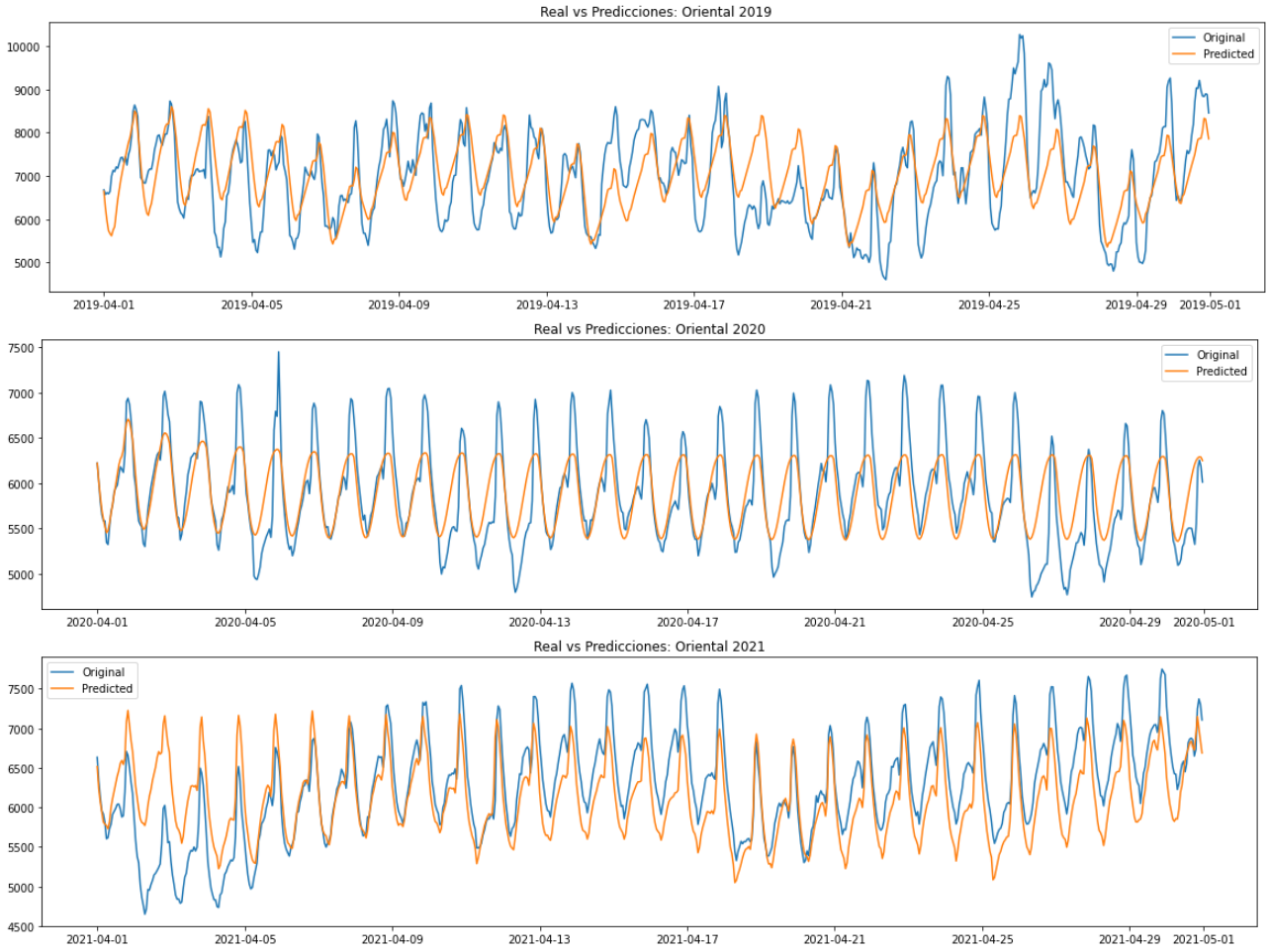


Figura 4.6: Errores de predicción en la región Oriental

4.3.7. Región Peninsular

De la misma forma que ocurrió con las demás regiones, en esta, los mejores pronósticos fueron realizados durante el año 2020, seguido por el 2019 y finalmente el 2021, esto de acuerdo con las métricas RMSE y MAE, sin embargo, al igual que ocurrido en la región Noreste, la métrica MAPE demuestra lo contrario, pero esto es debido a la dependencia que tiene MAPE respecto de la magnitud de los valores, y es que, la región peninsular presento un aumento en la demanda energética, por lo que sus valores, antes y durante la pandemia, se encuentran en diferentes magnitudes.

En la figura 4.7, al comparar la demanda real en cada uno de los años, es notable que, durante la pandemia, los patrones de comportamiento son más constantes, a diferencia del 2019. Sin embargo, de manera contraria a lo ocurrido en las regiones centro, oriental y occidental, que también presentaron esta característica, los peores pronósticos no se realizaron en el 2019, sino que, estos fueron en el 2021.

Año	Error			Algoritmo
	RMSE	MAE	MAPE (%)	
2019	109.41	89.94	11.72	XGboost
2020	**107.319**	**83.36**	**5.90**	**KNN**
2021	168.10	145.48	9.38	XGboost

Cuadro 4.9: Errores de predicción en la región Peninsular

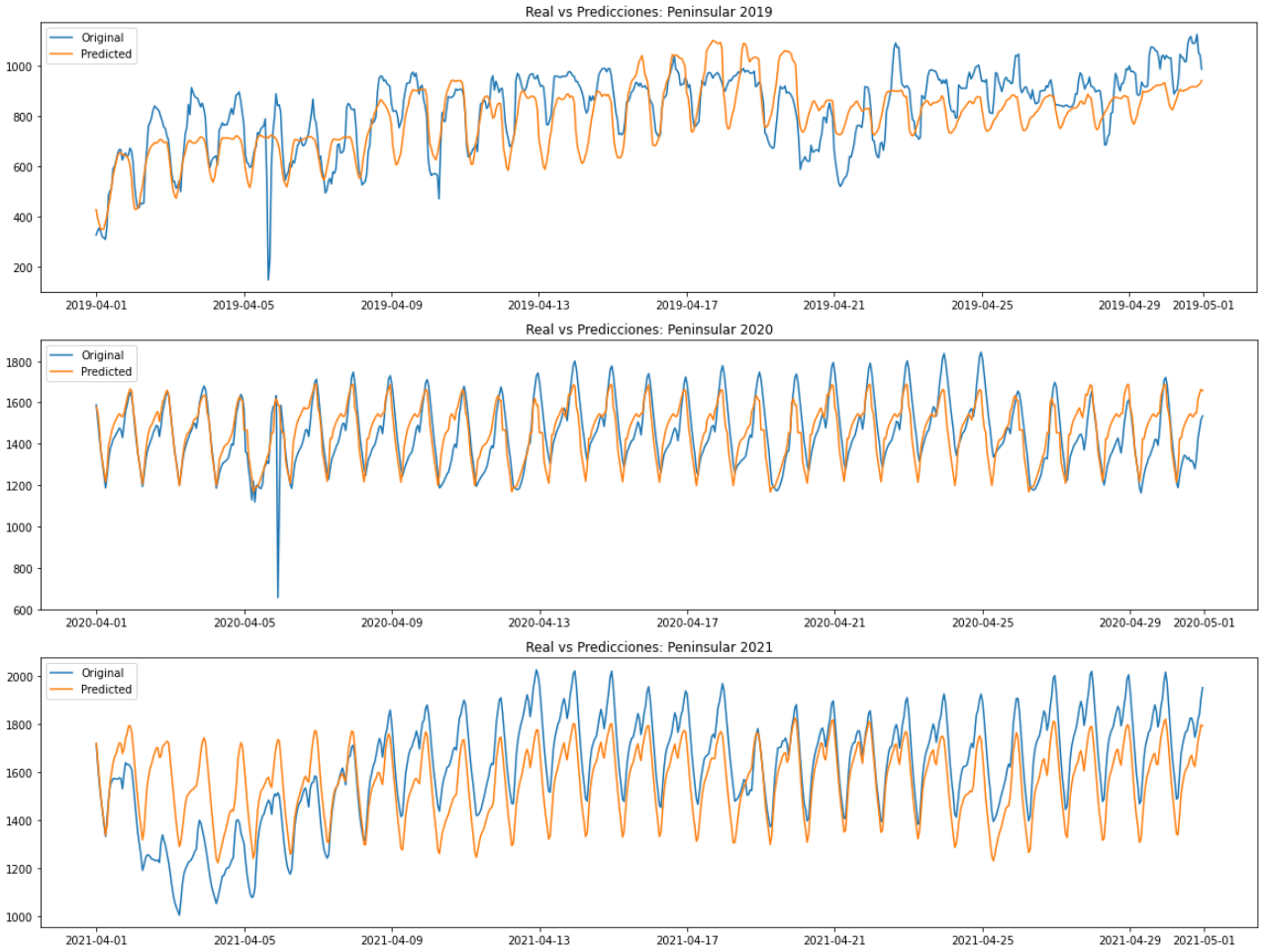


Figura 4.7: Errores de predicción en la región Peninsular

Y es que, a pesar de que la forma de los pronósticos realizados en el 2021 describió muy bien la verdadera forma de la demanda real, no pudo predecir con exactitud la magnitud, ya que la mayoría de los pronósticos estuvieron por debajo de la demanda real, con excepción de los primeros días del mes, donde se sobreestimó la demanda real.

4.3.8. Sistema Interconectado

A diferencia de las otras regiones, en las cuales se implementó un predictor específicamente para cada región, las predicciones sobre el sistema interconectado surgieron como la suma de las predicciones individuales de cada una de las regiones que lo conforman, por lo cual, no hubo la necesidad de implementar un predictor específicamente para este sistema.

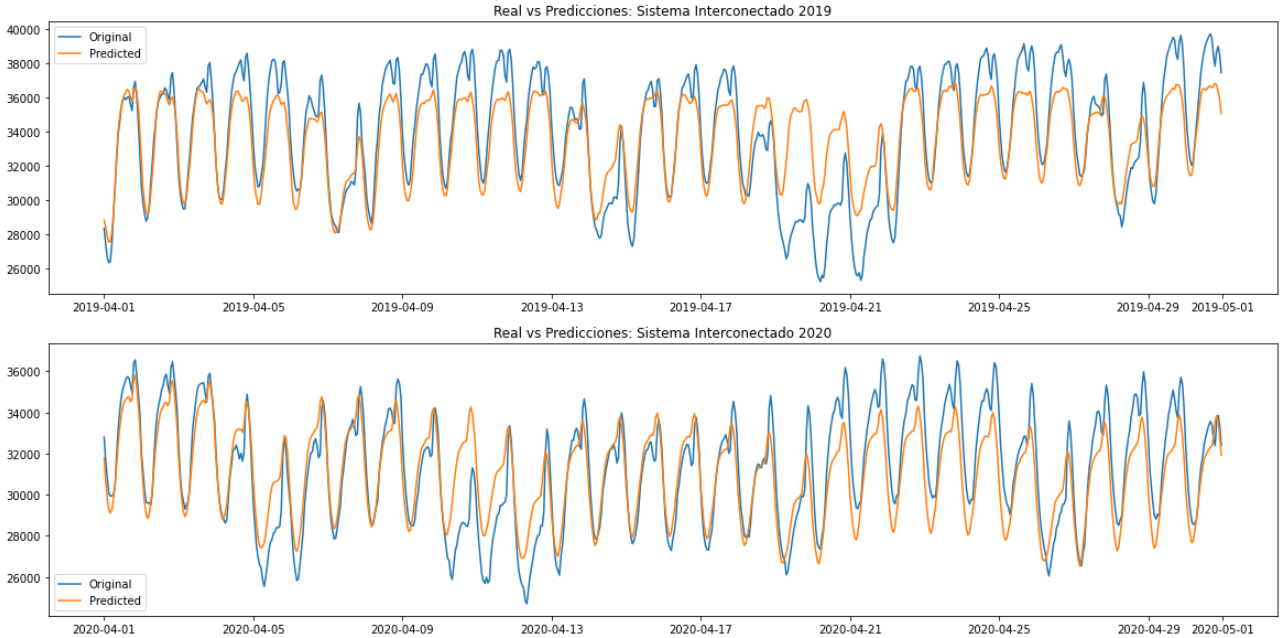
Una cuestión que aclarar sobre el sistema interconectado es que, al estar compuesto por la suma de las demandas en cada una de sus regiones, significa que sus resultados son una generalización de todas las regiones, y no una representación exacta de lo sucedido en cada una de ellas.

Año	**Error**			**Algoritmo**
	RMSE	MAE	MAPE (%)	
2019	1935.60	1478.77	4.50	XGboost
2020	**1460.02**	**1157.19**	**3.72**	**Ensamble**
2021	2209.10	1569.73	4.74	XGBoost

Cuadro 4.10: Errores de predicción en el Sistema Interconectado Nacional

De acuerdo con cada una de las métricas del error, en el sistema interconectado, los mejores pronósticos de la demanda energética se obtuvieron en el año 2020, seguido por el 2019 y, dejando al 2021, como el año donde se obtuvieron los pronósticos menos acertados.

Esto significa que, en el año 2020, la suma de los errores de predicción obtenidos en cada una de las regiones fue la menor, caso contrario del 2021, año en el que los errores de predicción fueron mayores.



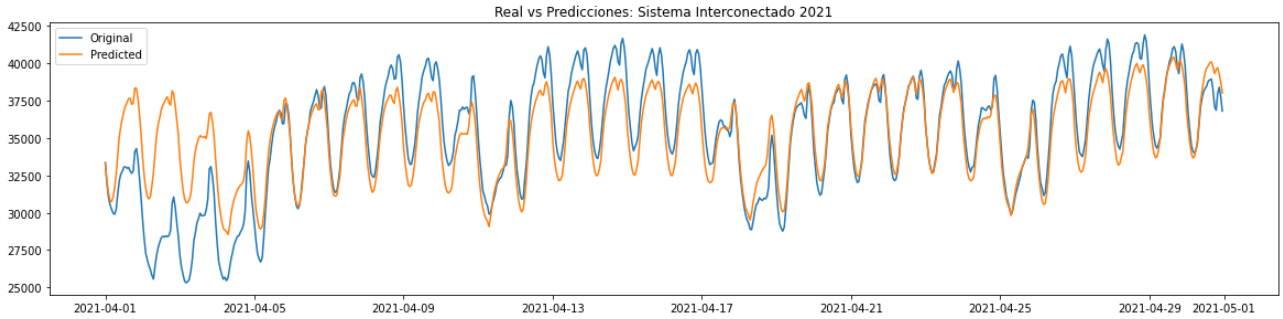


Figura 4.8: Errores de predicción en el Sistema Interconectado Nacional

En la figura 4.8, al comparar el comportamiento de la demanda real en los años 2019 y 2020, como era de esperarse, son diferentes. Sin embargo, al comparar el 2019 contra el 2021, existen varias similitudes en sus patrones de comportamiento, lo cual, es un indicativo que el consumo general de la demanda energética está volviendo a la "normalidad", es decir, al que se tenía antes de la pandemia.

Comentarios finales

Con el objetivo de mejorar las predicciones individuales de los modelos implementados en el capítulo 3, en este capítulo se realizó un ensamble de ellos utilizando la técnica de bagging. Sin embargo, al comparar la eficiencia de cada una de estas técnicas, los resultados mostraron que, tanto XGBoost como el ensamble resultaron ser los mejores, ya que sus pronósticos tuvieron los menores errores de predicción en una cantidad similar de casos, siendo estos la mayoría.

Finalmente se analizaron las gráficas del comportamiento real y pronosticado de los mejores modelos de predicción, lo cual permitió identificar características como: la persistencia de los efectos de la pandemia en el año 2021, y una mayor estabilidad en los patrones de comportamiento de la demanda energética durante los años de pandemia.

Capítulo 5

Conclusiones y trabajo futuro

A continuación, se describen las conclusiones de los análisis realizados en el capítulo 4. Además, se presentan algunos de los posibles trabajos futuros que pueden continuar desarrollándose como resultado de la investigación.

5.1. Conclusiones

1. El brote de coronavirus (codiv-19) junto con las regulaciones gubernamentales de imponer el cierre de diversas actividades tanto laborales como recreativas condujeron a un cambio en la conducta de la sociedad, lo cual, modifico los patrones de comportamiento de diversos aspectos sociales, de entre los cuales, está el consumo energético, tratado en esta tesis.

2. A través del análisis exploratorio realizado sobre cada una de las regiones que conforman el Sistema Interconectado Nacional, se mostró que los efectos de la pandemia causaron un cambio en diferentes aspectos de la demanda energética, siendo los principales, la estacionalidad y volatilidad. Naturalmente, los efectos de estos cambios no fueron los mismos para todas las regiones, ya que sus poblaciones se dedican a diferentes actividades, por lo que sus necesidades energéticas son diferentes.

3. En esta tesis, para evidenciar el efecto de este nuevo comportamiento sobre la eficiencia de los predictores, se implementaron modelos de predicción utilizando diferentes técnicas de machine learning, de entre las cuales, tanto XGBoost como el ensamble mostraron hacer las mejores predicciones en comparación de las otras técnicas implementadas, sin embargo, un aspecto a considerar del ensamble es que este es más costoso en tiempo, ya que su implementación requiere del previo entrenamiento de varios modelos de predicción. Por esta razón, una recomendación para lidiar con las limitaciones del tiempo seria implementar las técnicas por orden de prioridad, así, en caso de llegar al límite de tiempo, se podría disponer de los resultados individuales de los modelos más eficientes, o alternativamente, realizar el ensamble con ellos.

4. De acuerdo con los pronósticos realizados durante los años 2019, 2020 y 2021, los modelos de predicción mostraron una mejor recepción al nuevo comportamiento de la demanda energética causado por la pandemia, ya que, en seis de las siete regiones analizadas, los errores de predicción obtenidos durante los años de pandemia fueron menores.

5. A pesar de haber contado con más datos e información de la demanda energética pre – pandemia estos resultados indica una mejora en la eficiencia de los predictores durante estos dos últimos años. Y es que, un aspecto positivo que surgió con la llegada de la pandemia fue una mayor estabilidad en los patrones de comportamiento de la demanda, provocando que los modelos mejoraran en el aprendizaje del nuevo comportamiento de los datos, y por consecuencia, aumentara la precisión de los pronósticos realizados.

5.2. Trabajo futuro

A continuación, se presentan algunos trabajos que podrían tomarse como futuras investigaciones para darle seguimiento a esta investigación o que pueden desarrollarse como resultado de esta tesis:

- Considerar si la eficiencia de otras técnicas de pronóstico, tales como los modelos autorregresivos (AR) o redes neuronales, se vieron afectada de la misma forma.
- Realizar investigaciones tomando en cuenta el comportamiento de la sociedad para averiguar cuál será el futuro del consumo energético en México, si se restablecerá el comportamiento presentado antes de la pandemia, o si se mantendrá este nuevo comportamiento.

Bibliografía

[1] www.freecodecamp.org, «www.freecodecamp.org,» [En línea]. Available: https://www.freecodecamp.org/news/machine-learning-for-managers-what-you-need-to-know/. [Último acceso: 15 09 2021].

[2] Morales, L. G. G., & Perucci, F. (2021). COVID-19: How the data and statistical community stepped up to the new challenges.

[3] Eftimov, T., Popovski, G., Petković, M., Seljak, B. K., & Kocev, D. (2020). COVID-19 pandemic changes the food consumption patterns. Trends in Food Science & Technology, 104, 268.

[4] Carvalho, M., Bandeira de Mello Delgado, D., de Lima, K. M., de Camargo Cancela, M., dos Siqueira, C. A., & de Souza, D. L. B. (2021). Effects of the COVID-19 pandemic on the Brazilian electricity consumption patterns. International Journal of Energy Research, 45(2), 3358-3364.

[5] Gulati, P., Kumar, A., & Bhardwaj, R. (2021). Impact of Covid19 on electricity load in Haryana (India). International Journal of Energy Research, 45(2), 3397-3409.

[6] Tiwari, D., Bhati, B. S., Al-Turjman, F., & Nagpal, B. (2021). Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques. Expert Systems.

[7] Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., & Alsulmi, E. S. (2021). Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients. Scientific Programming, 2021.

[8] G.d.Mexico, «www.gob.mx,» [En línea]. Available: https://www.gob.mx/sener/articulos/el-gobierno-de-mexico-fortalece-el-sistema-electrico-nacional. [Último acceso: 13 09 2021].

[9] C. R. d. Energia, «www.gob.mx,» [En línea]. Available: https://cutt.ly/OW6SPEQ. [Último acceso: 15 09 2021].

[10] C. N. d. C. d. Energia, «www.cenace.gob.mx,» [En línea]. Available: https://www.cenace.gob.mx/Paginas/SIM/Reportes/EstimacionDemandaReal.aspx. [Último acceso: 15 09 2021].

[11] S. d. G. (SEGOB), «dof.gob.mx,» [En línea]. Available: https://www.dof.gob.mx/nota_detalle.php?codigo=5590914&fecha=31/03/2020. [Último acceso: 22 09 2021].

[12] datavizcatalogue.com, «datavizcatalogue.com,» [En línea]. Available: https://datavizcatalogue.com/ES/metodos/diagrama_cajas_y_bigotes.html. [Último acceso: 25 09 2021].

[13] N. K. M. Altman, «The curse(s) of dimensionality,» Nat Methods, vol. 15, nº 6, p. 399–400, 2018.

[14] J. Fischer, «statworx.com,» [En línea]. Available: https://www.statworx.com/at/blog/what-the-mape-is-falsely-blamed-for-its-true-weaknesses-and-better-alternatives/#h-particularly-small. [Último acceso: 13 11 2021].

[15] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[16] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

[17] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[18] IArtificial.net, «IArtificial.net,» [En línea]. Available: https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/. [Último acceso: 19 10 2021].

[19] Fix, E., & Hodges Jr, J. L. (1952). Discriminatory analysis-nonparametric discrimination: Small sample performance. California Univ Berkeley.

A) Análisis de normalidad mediante histogramas para el SIN y sus regiones

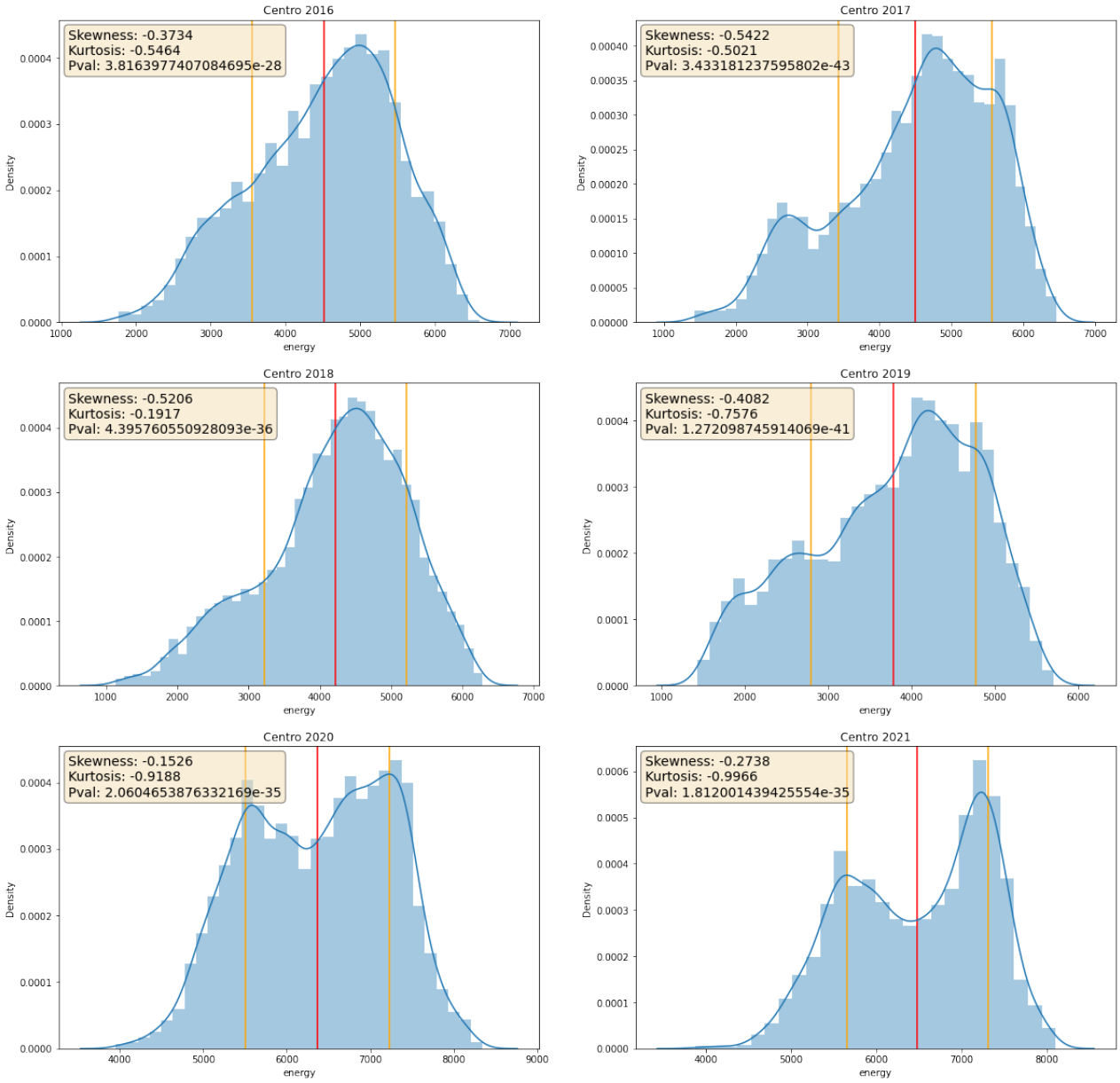


Figura 1: Análisis de Normalidad en la región centro

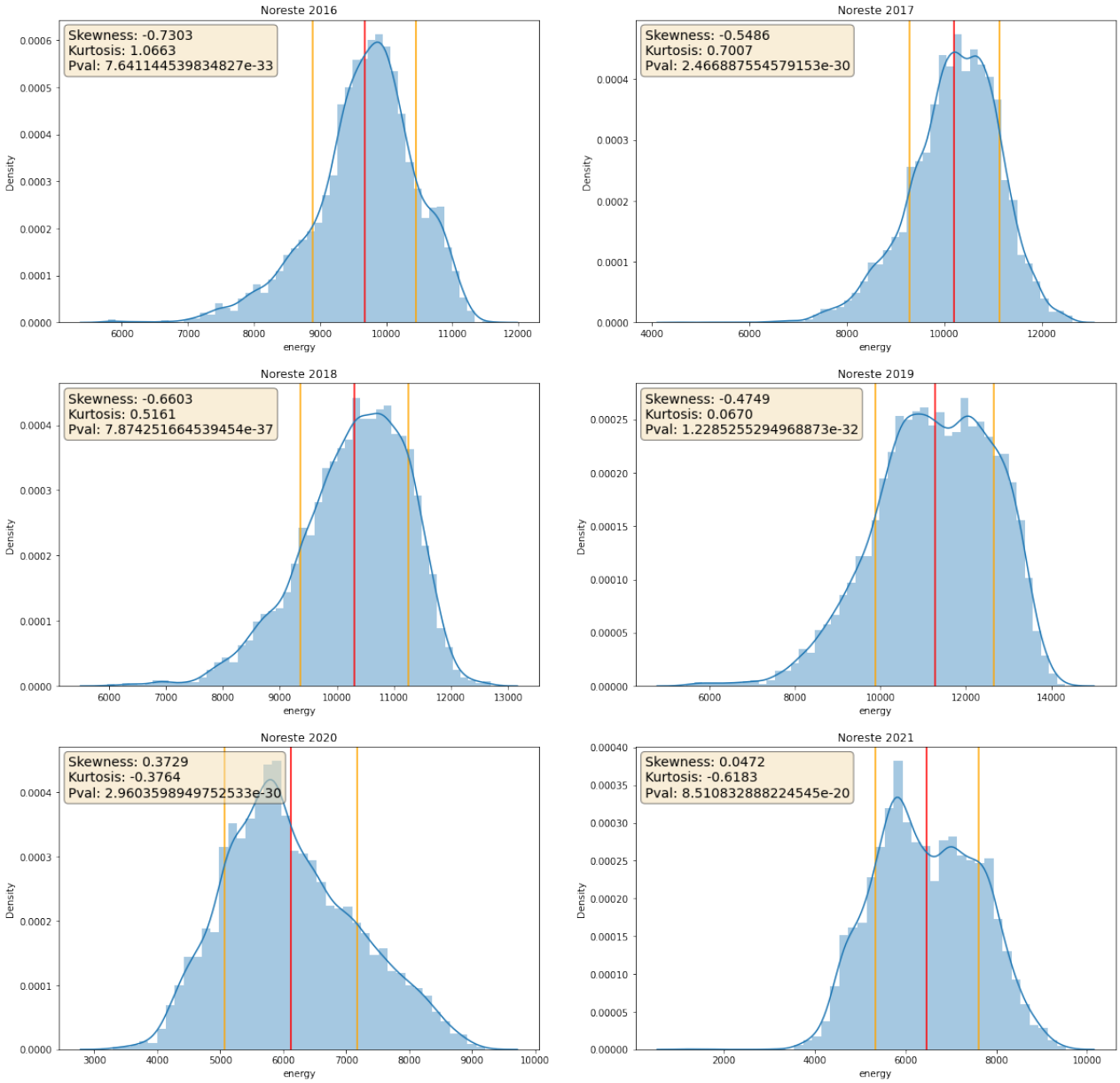


Figura 2: Análisis de Normalidad en la región Noreste

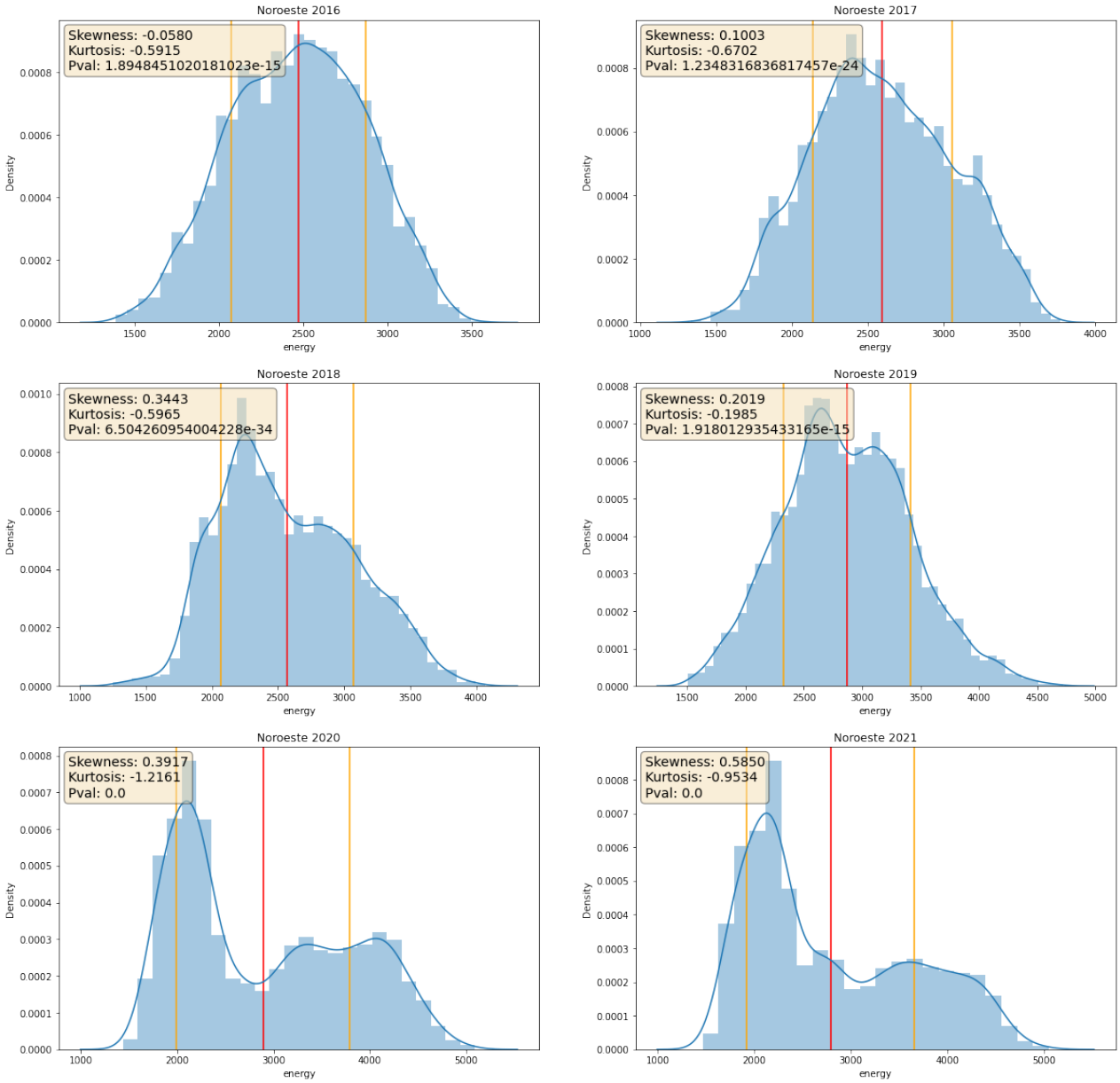


Figura 3: Análisis de Normalidad en la región Noroeste

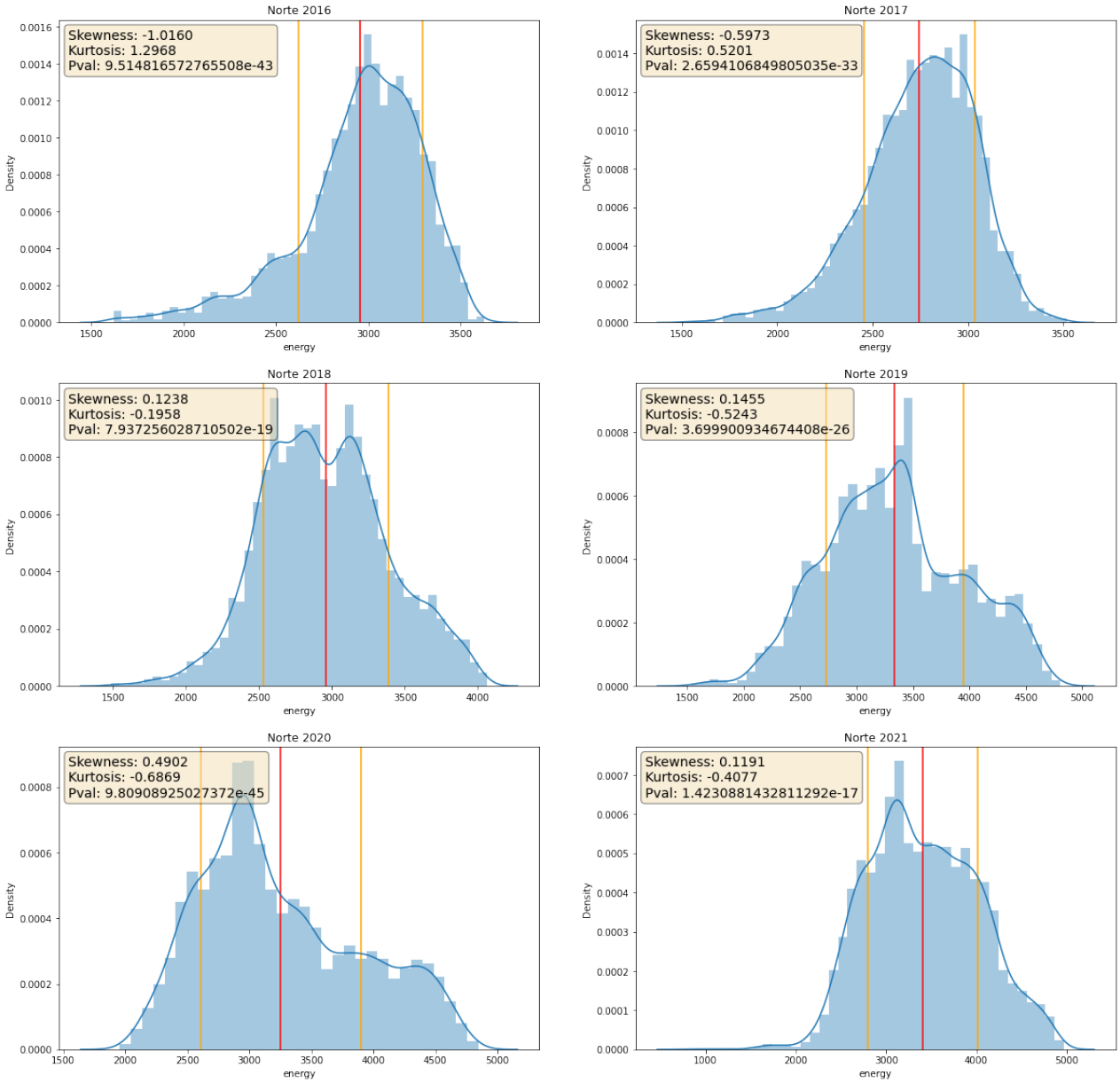


Figura 4: Análisis de Normalidad en la región Norte

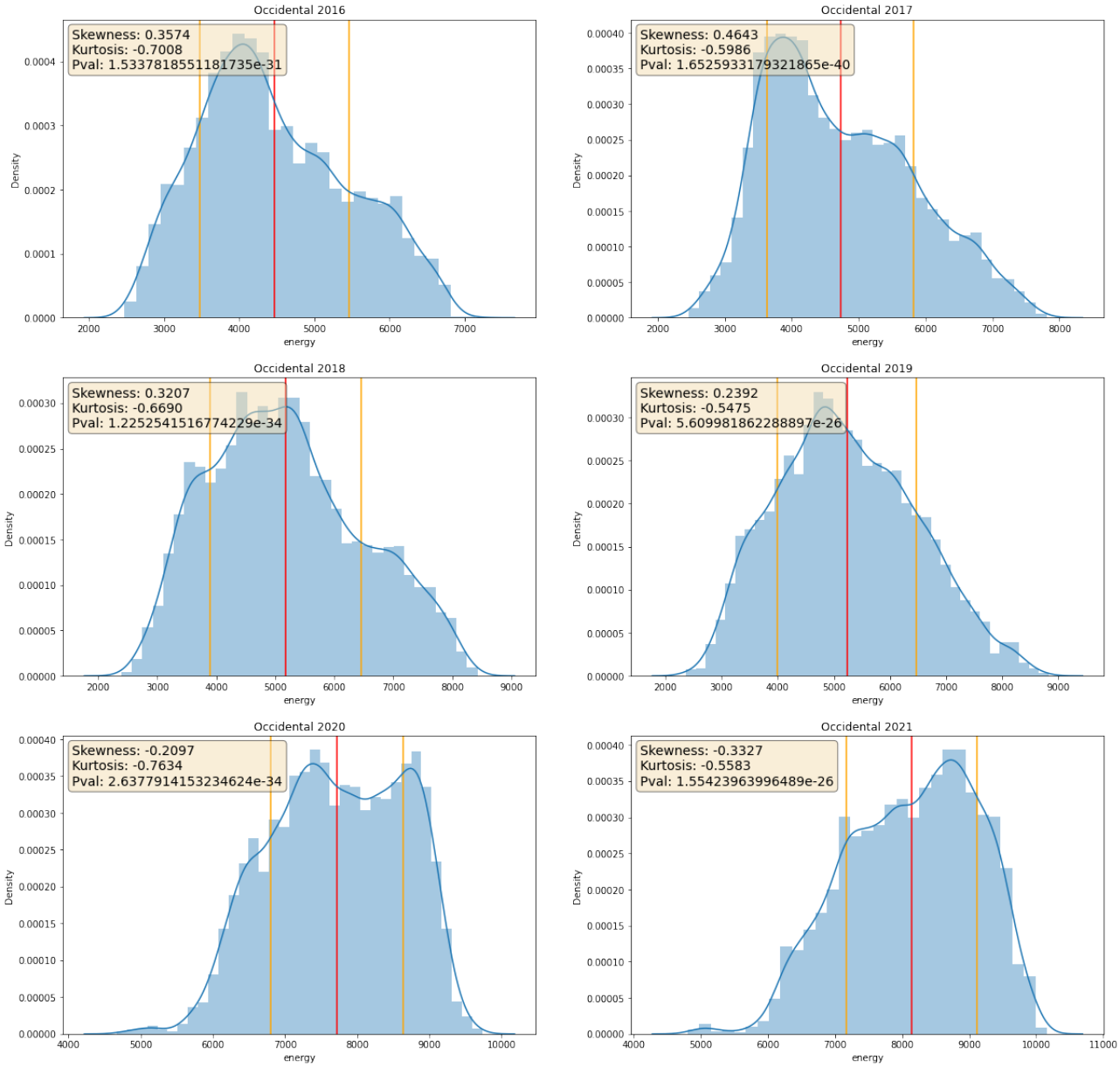


Figura 5: Análisis de Normalidad en la región Occidente

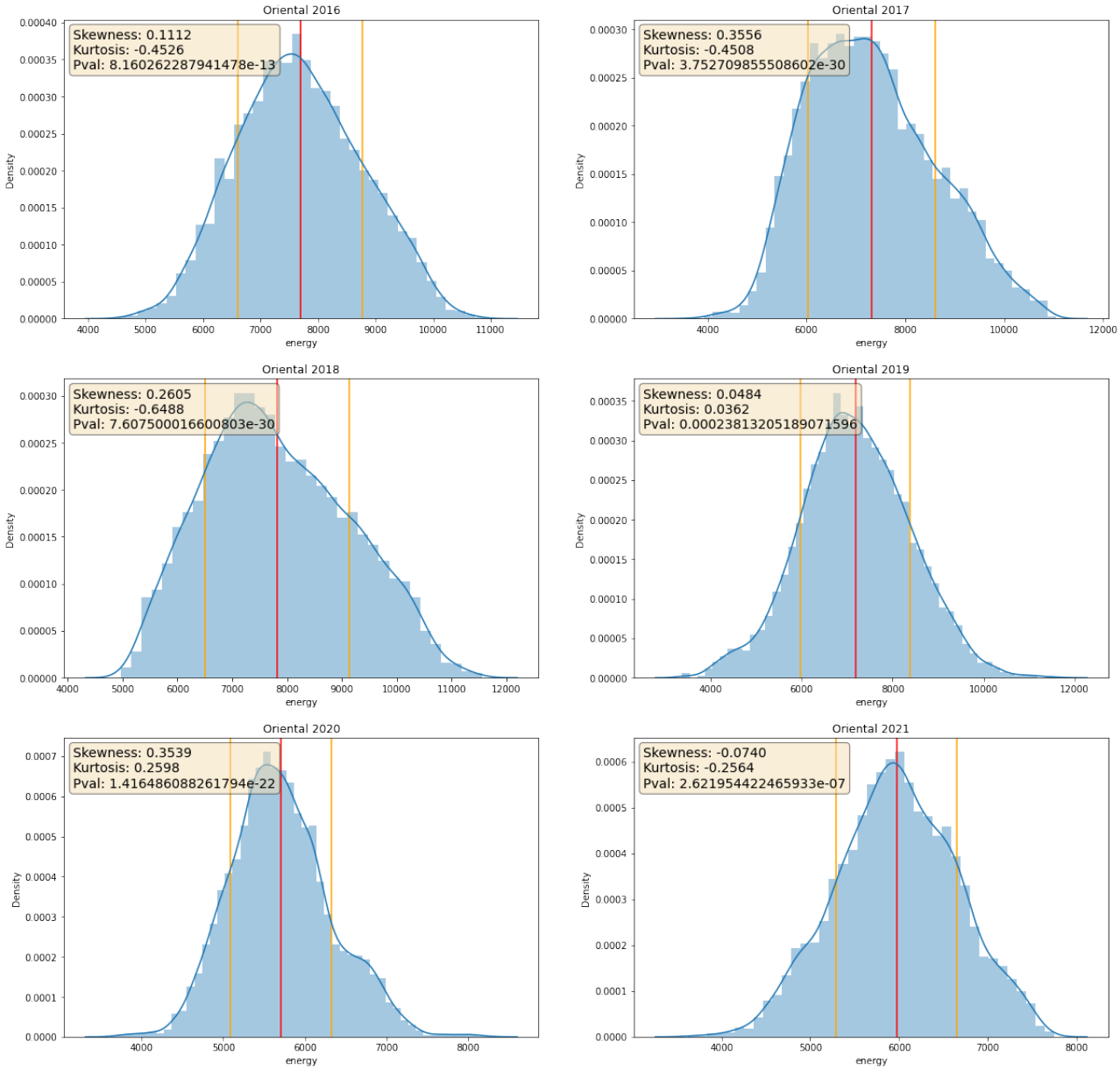


Figura 6: Análisis de Normalidad en la región Oriental

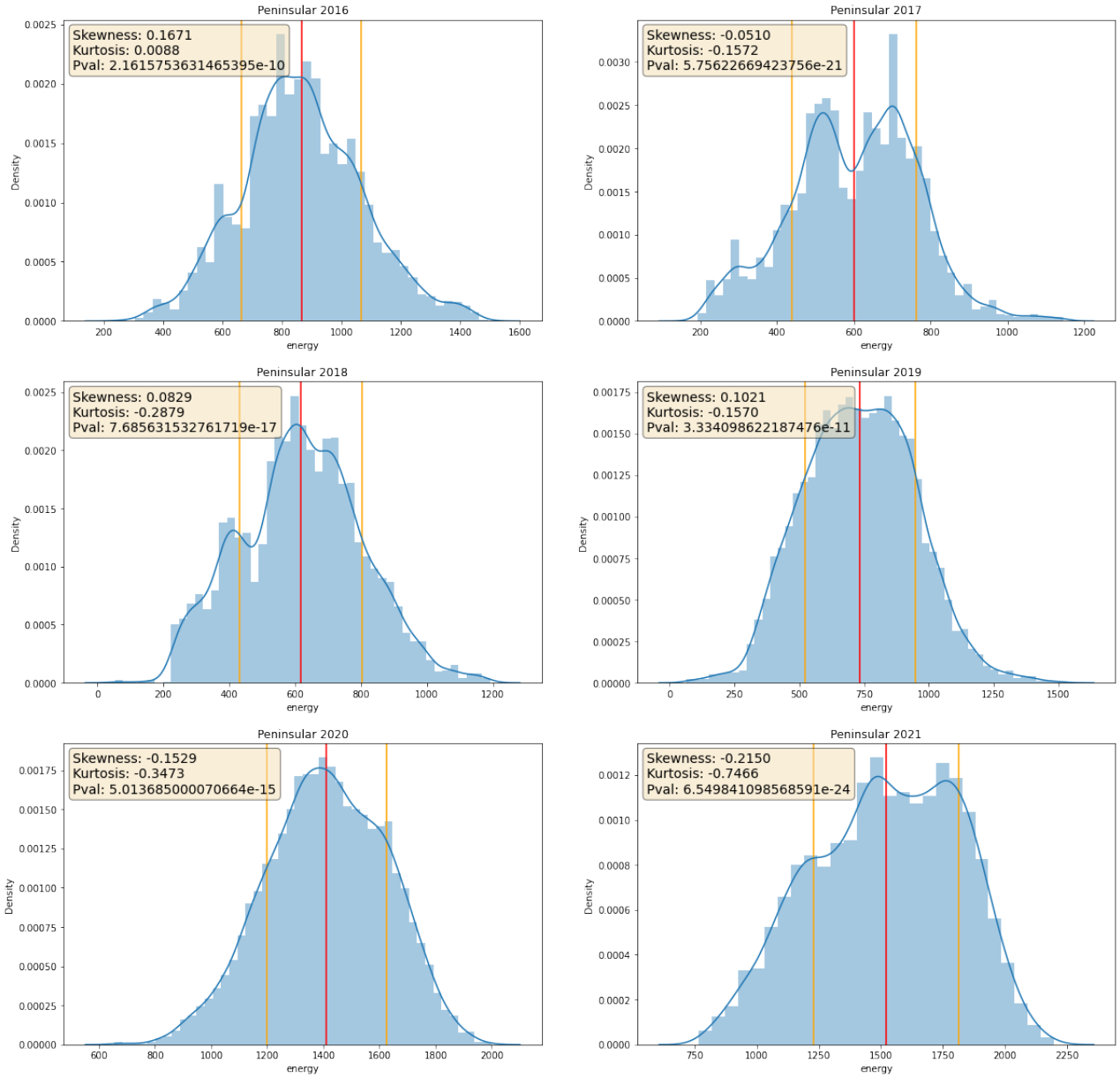


Figura 7: Análisis de Normalidad en la región Peninsular

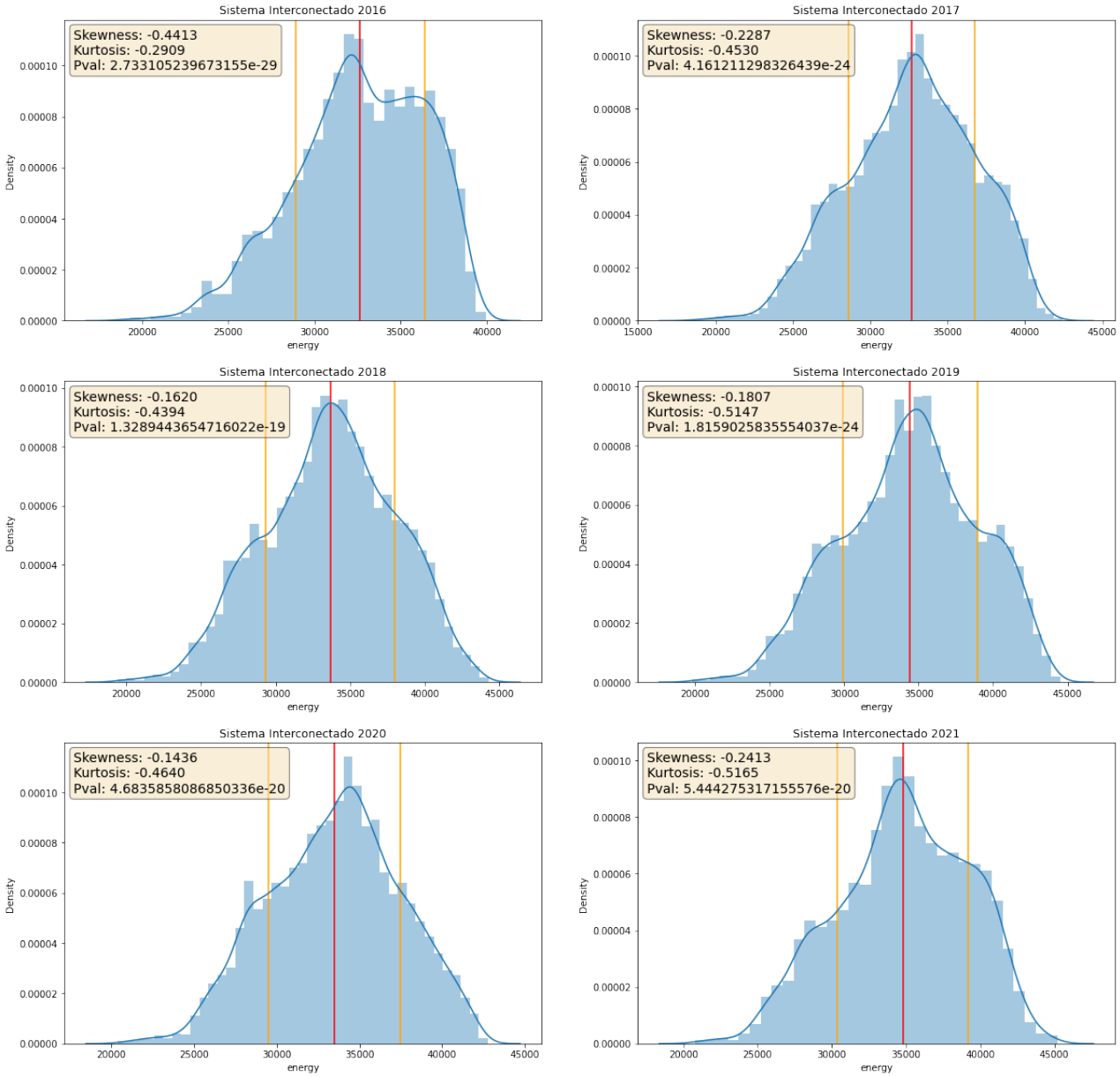


Figura 8: Análisis de Normalidad en el Sistema Interconectado

B) Analisis de normalidad mediante graficos QQ para el SIN y sus regiones

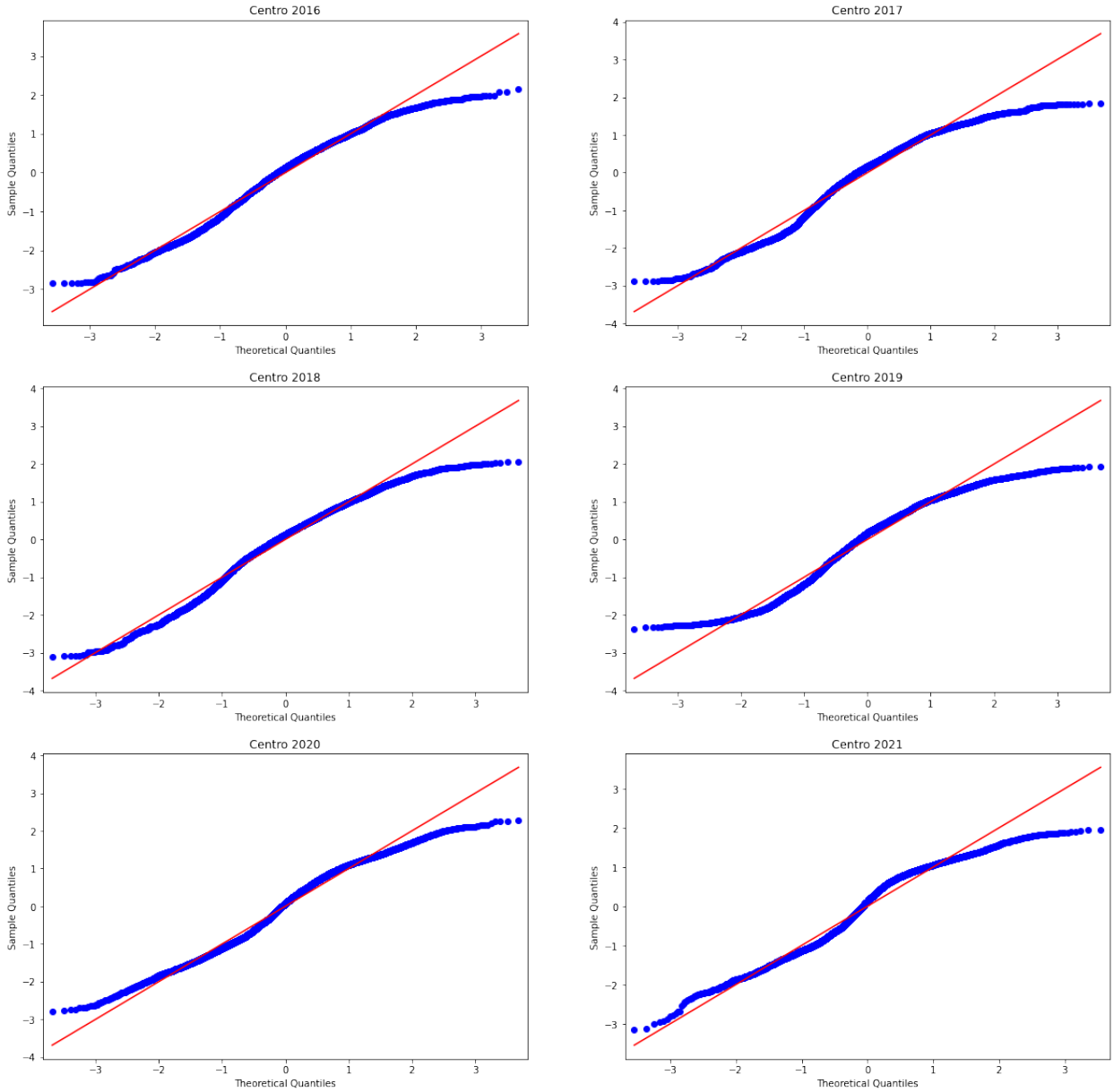


Figura 9: Análisis de Normalidad en la región centro

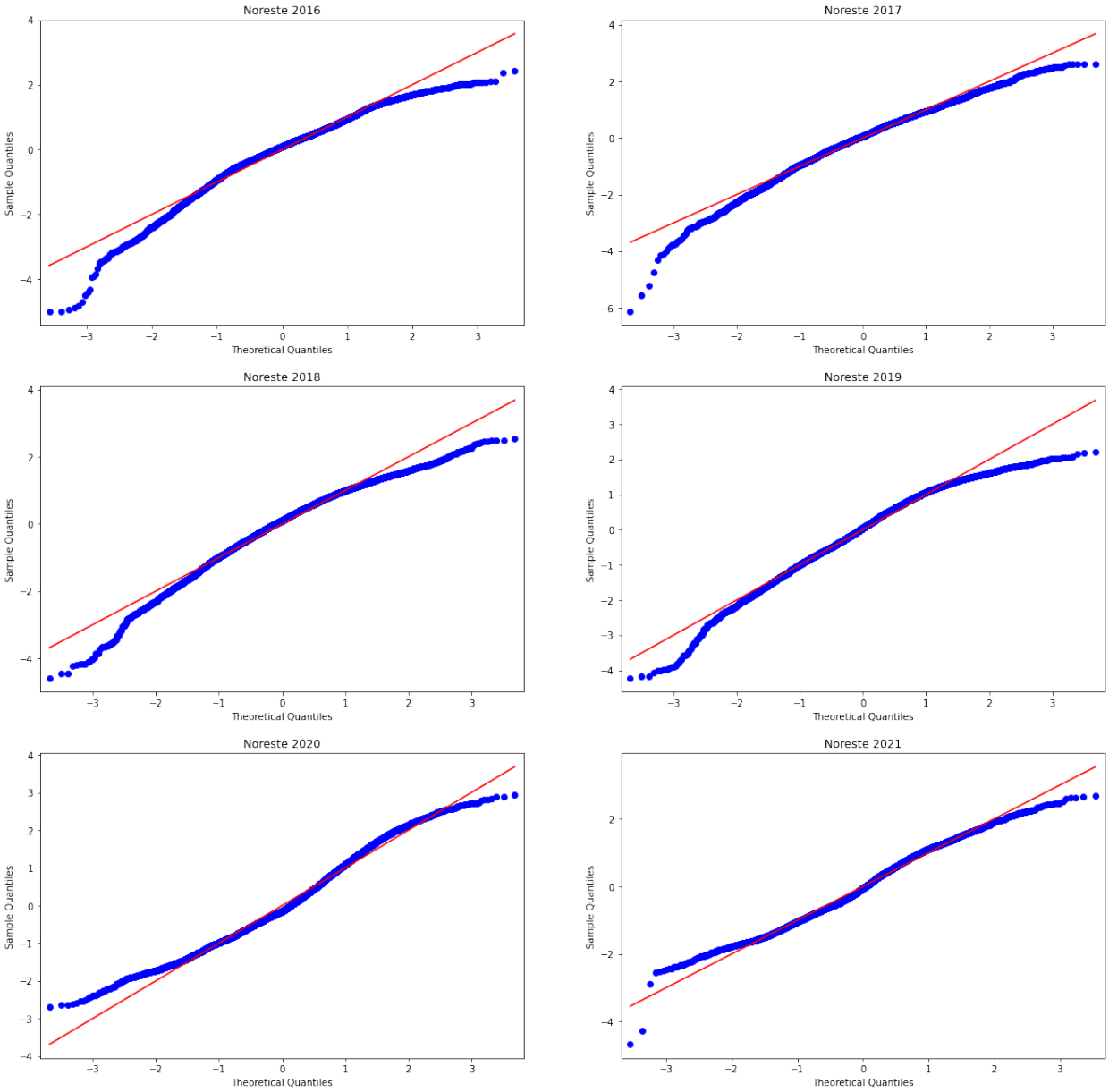


Figura 10: Gráficos QQ para el análisis de normalidad de la región Noreste

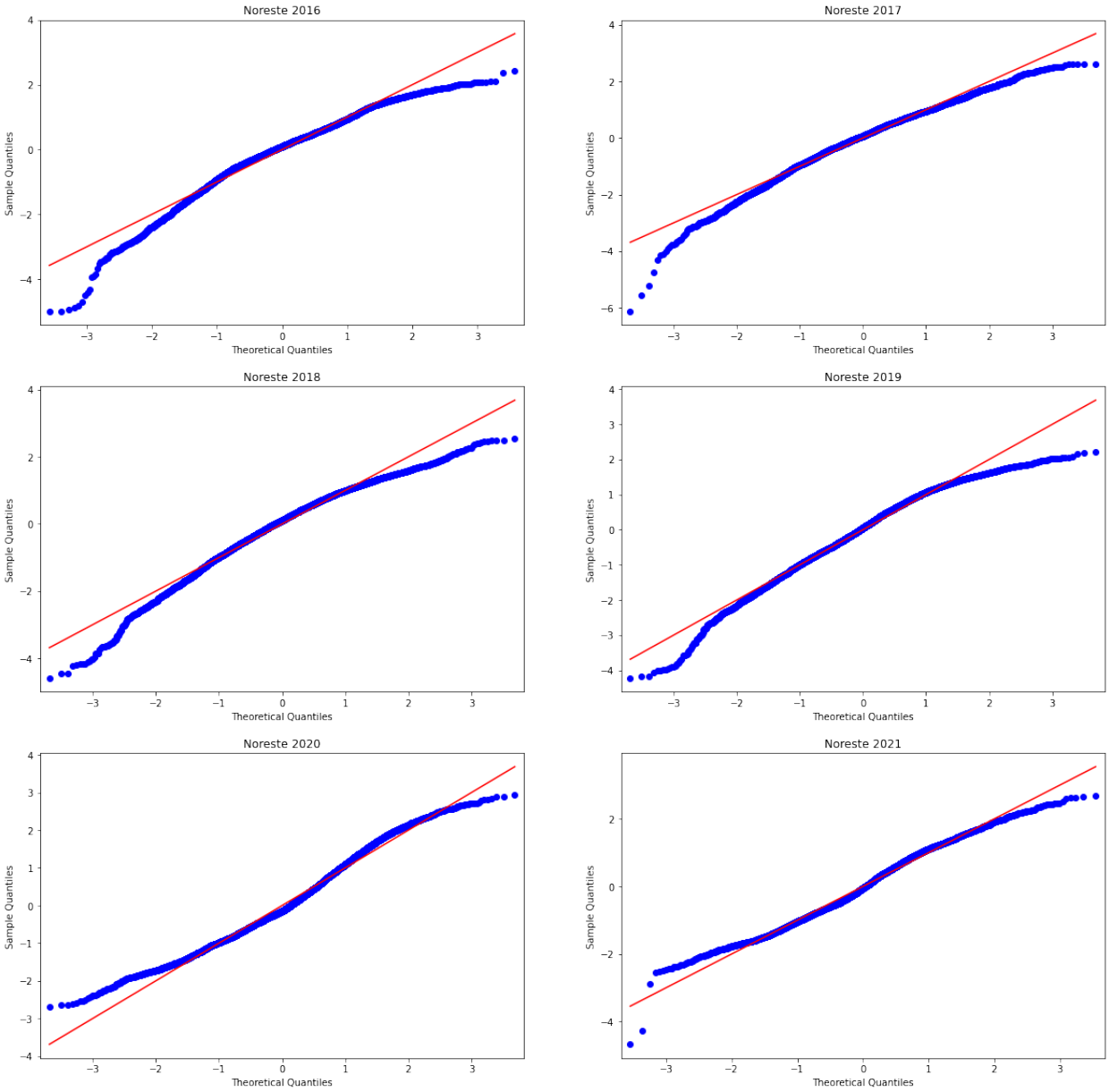


Figura 11: Gráficos QQ para el análisis de normalidad de la región Noroeste

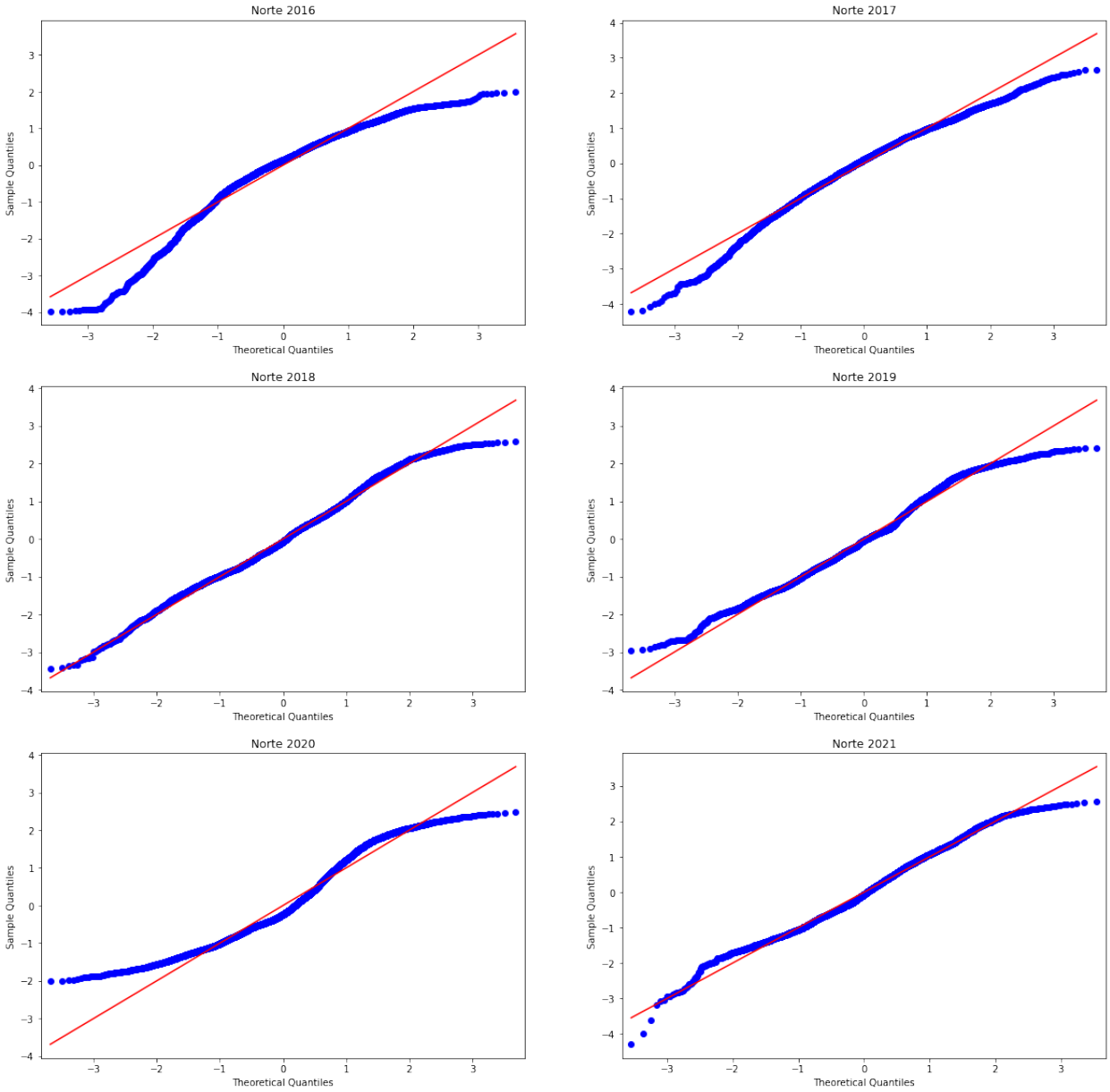


Figura 12: Gráficos QQ para el análisis de normalidad de la región Norte

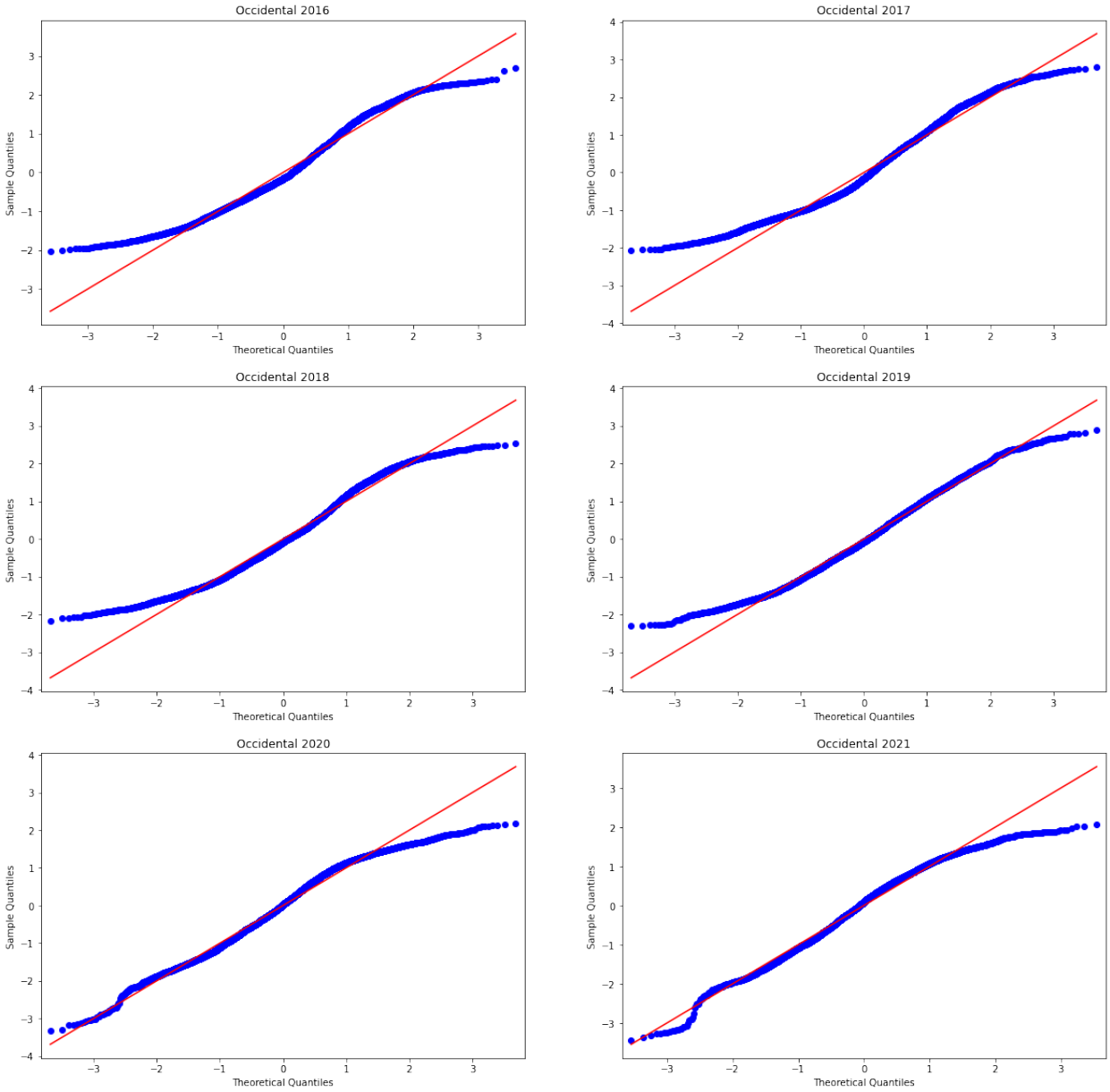


Figura 13: Gráficos QQ para el análisis de normalidad de la región Occidente

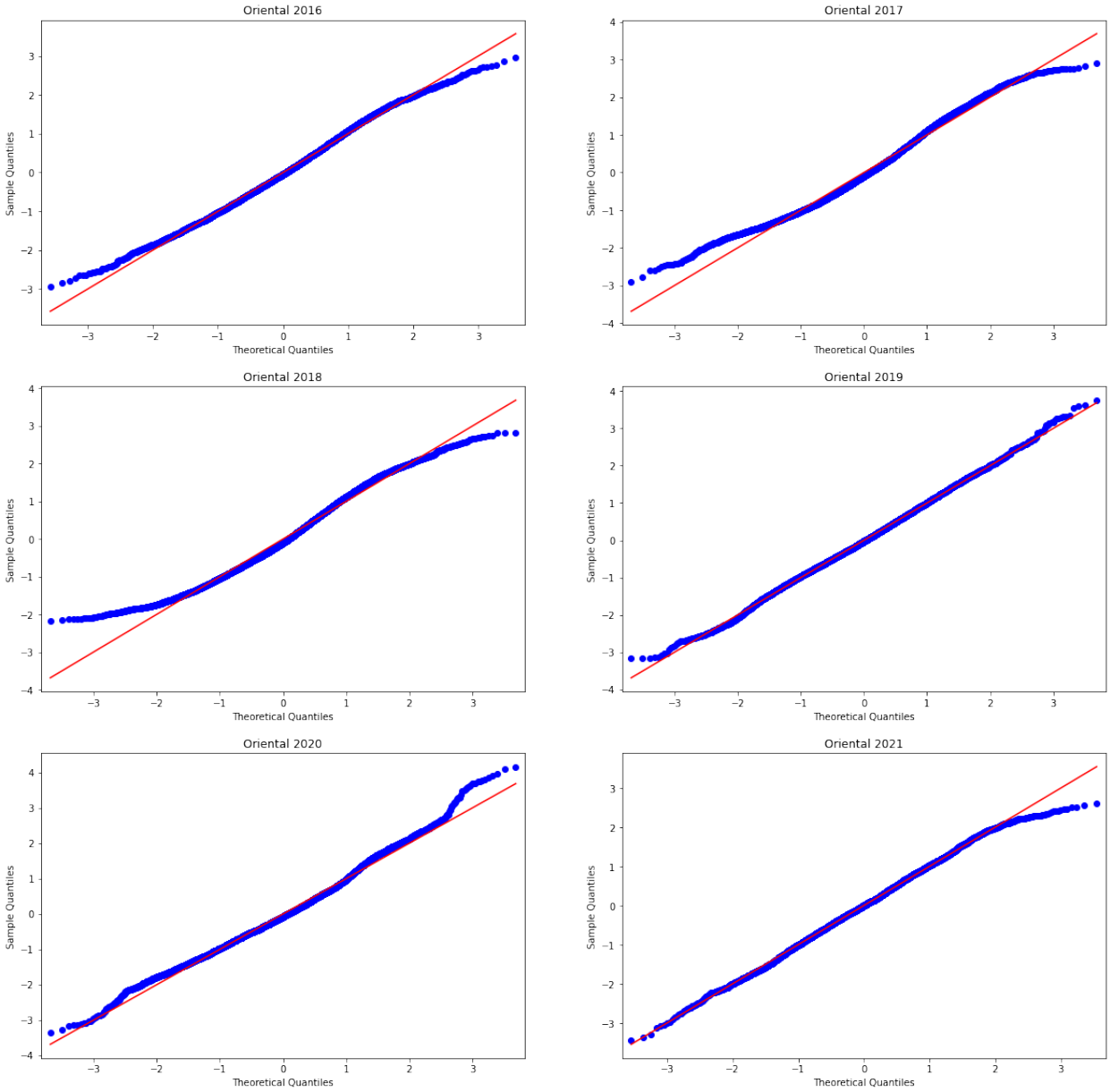


Figura 14: Gráficos QQ para el análisis de normalidad de la región Oriental

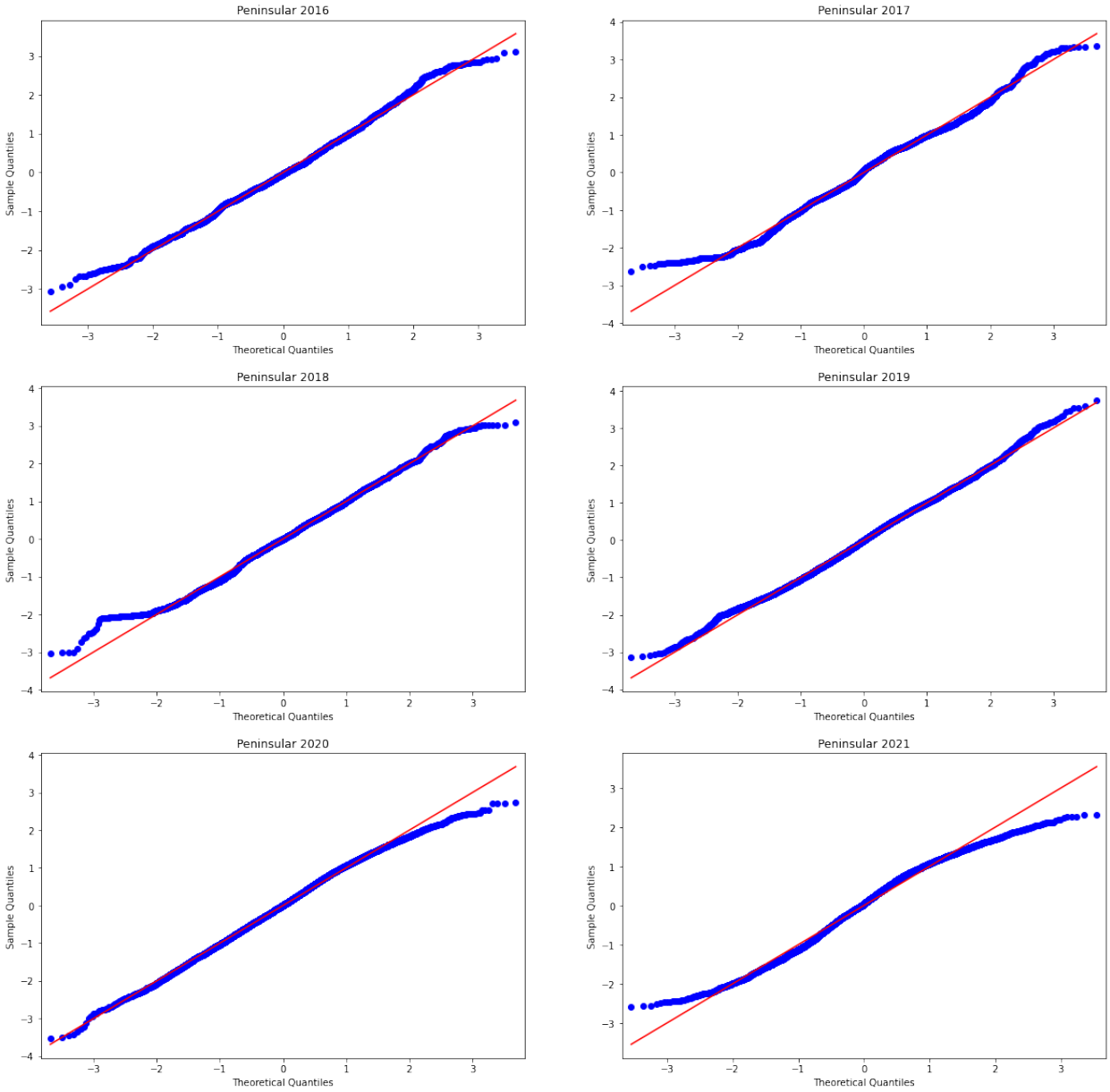


Figura 15: Gráficos QQ para el análisis de normalidad de la región Peninsular

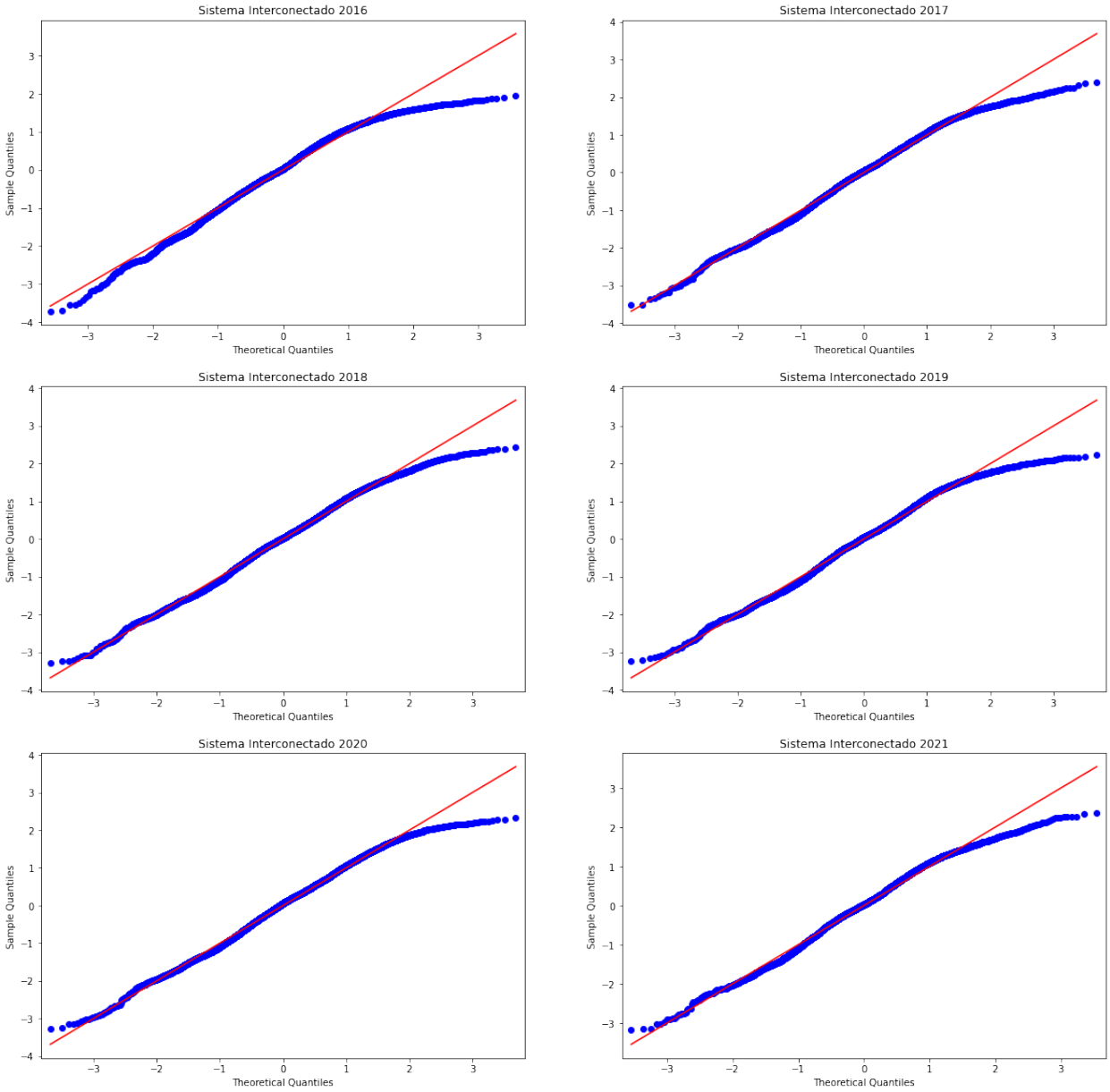


Figura 16: Gráficos QQ para el análisis de normalidad del Sistema Interconectado

C) Coeficiente de variación actual y pre-pandemia del SIN y sus regiones

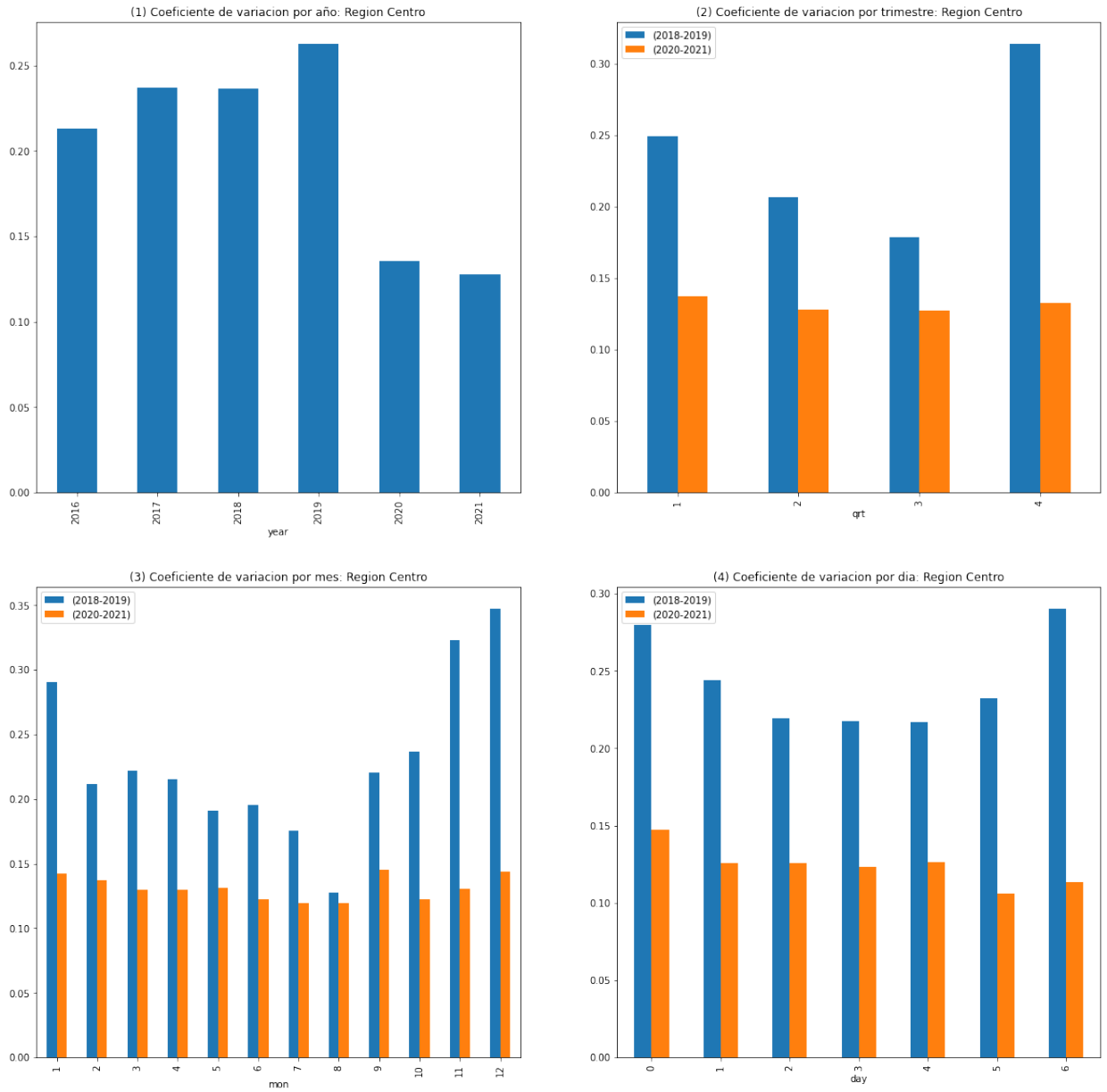


Figura 17: Coeficientes de variación de la región Central

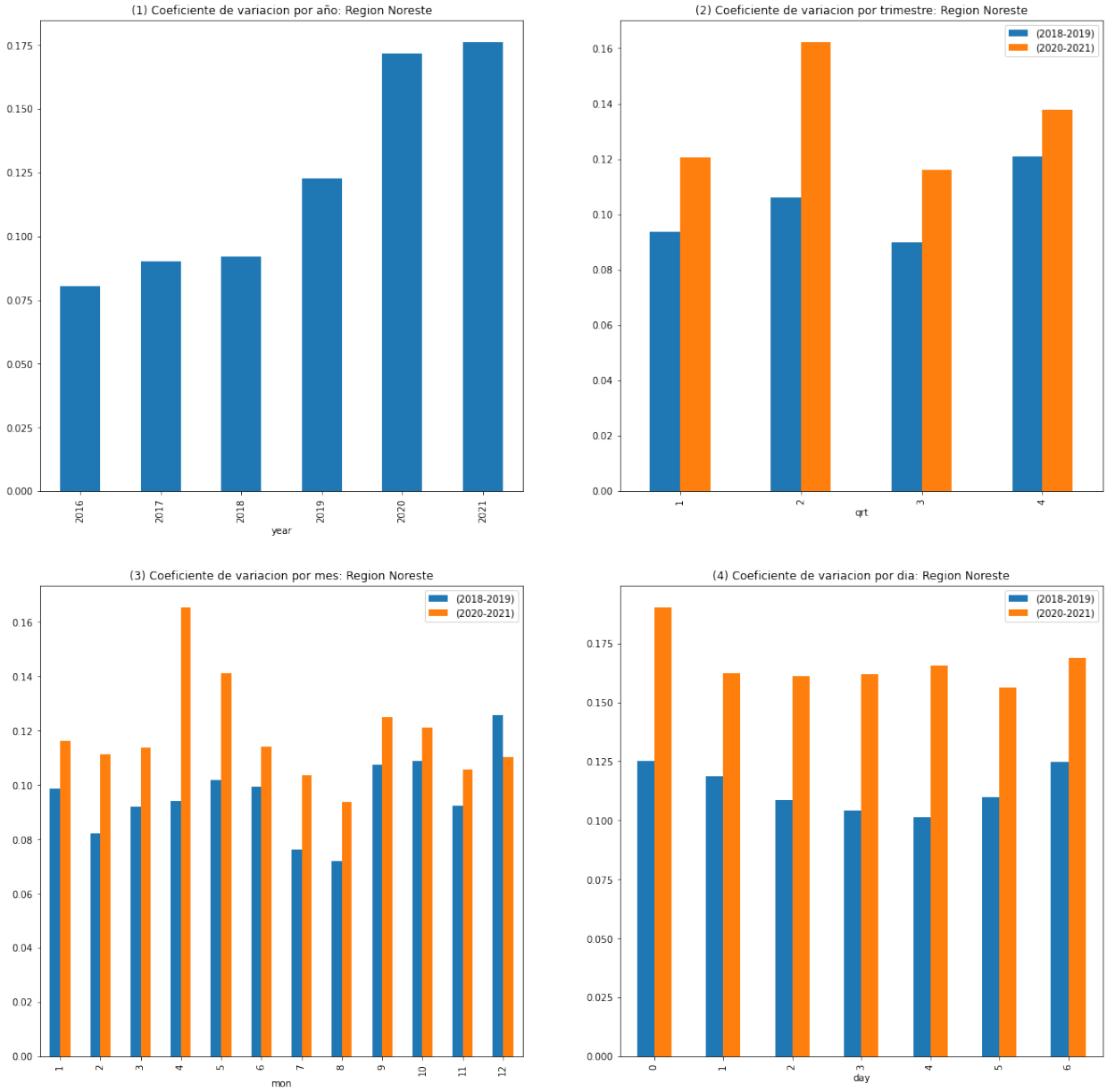


Figura 18: Coeficientes de variación de la región Noreste

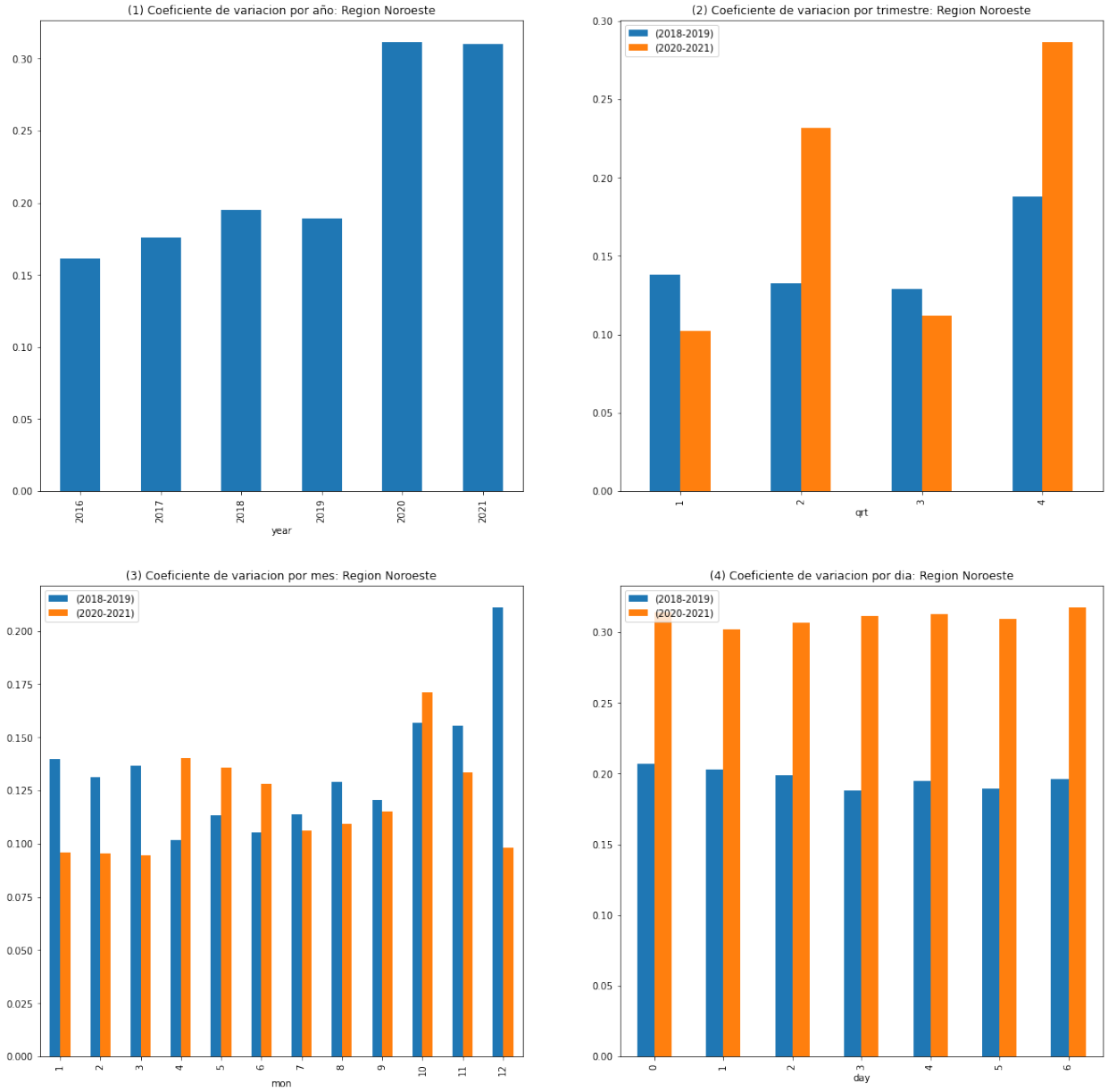


Figura 19: Coeficientes de variación de la región Noroeste

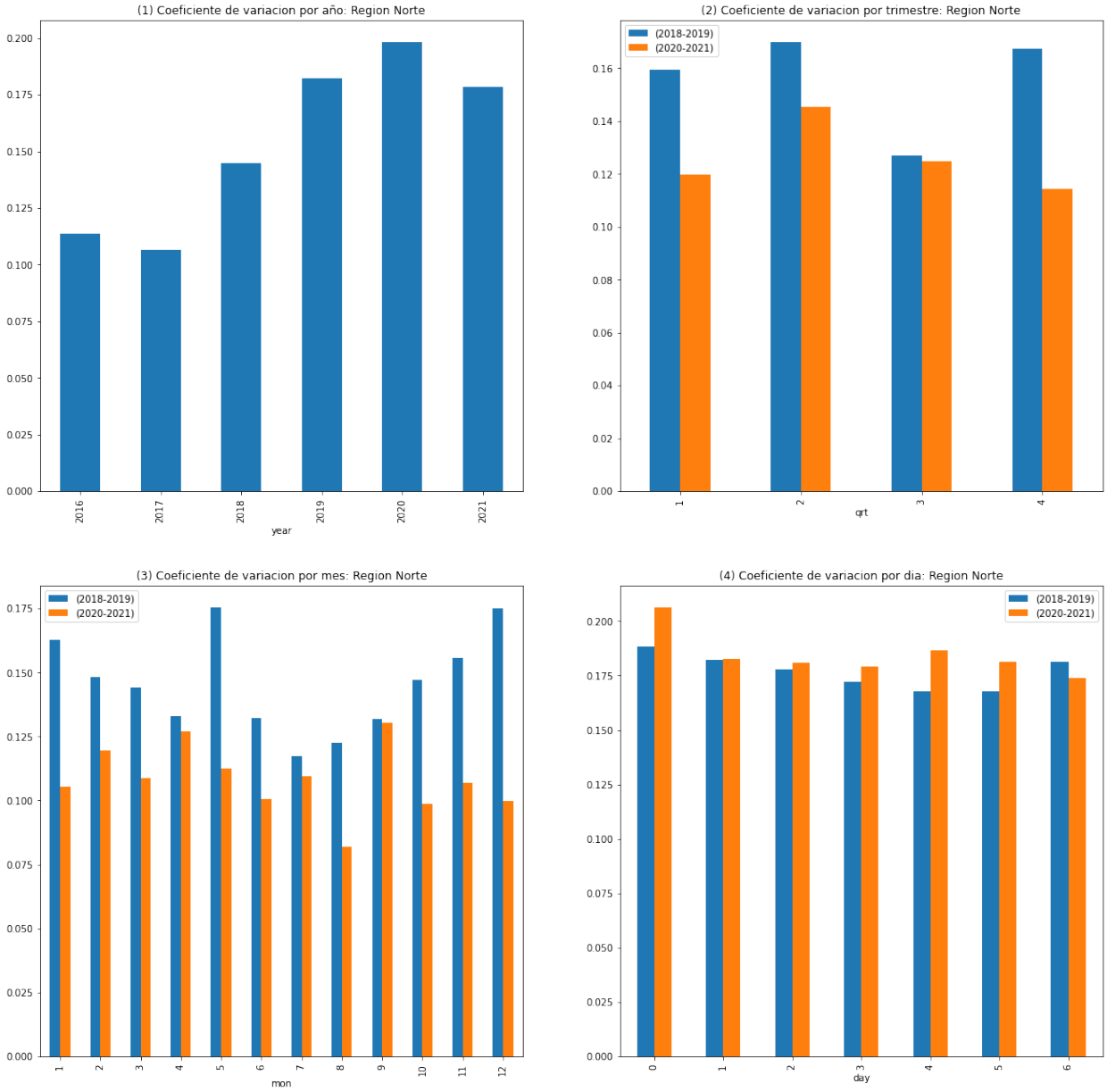


Figura 20: Coeficientes de variación de la región Norte

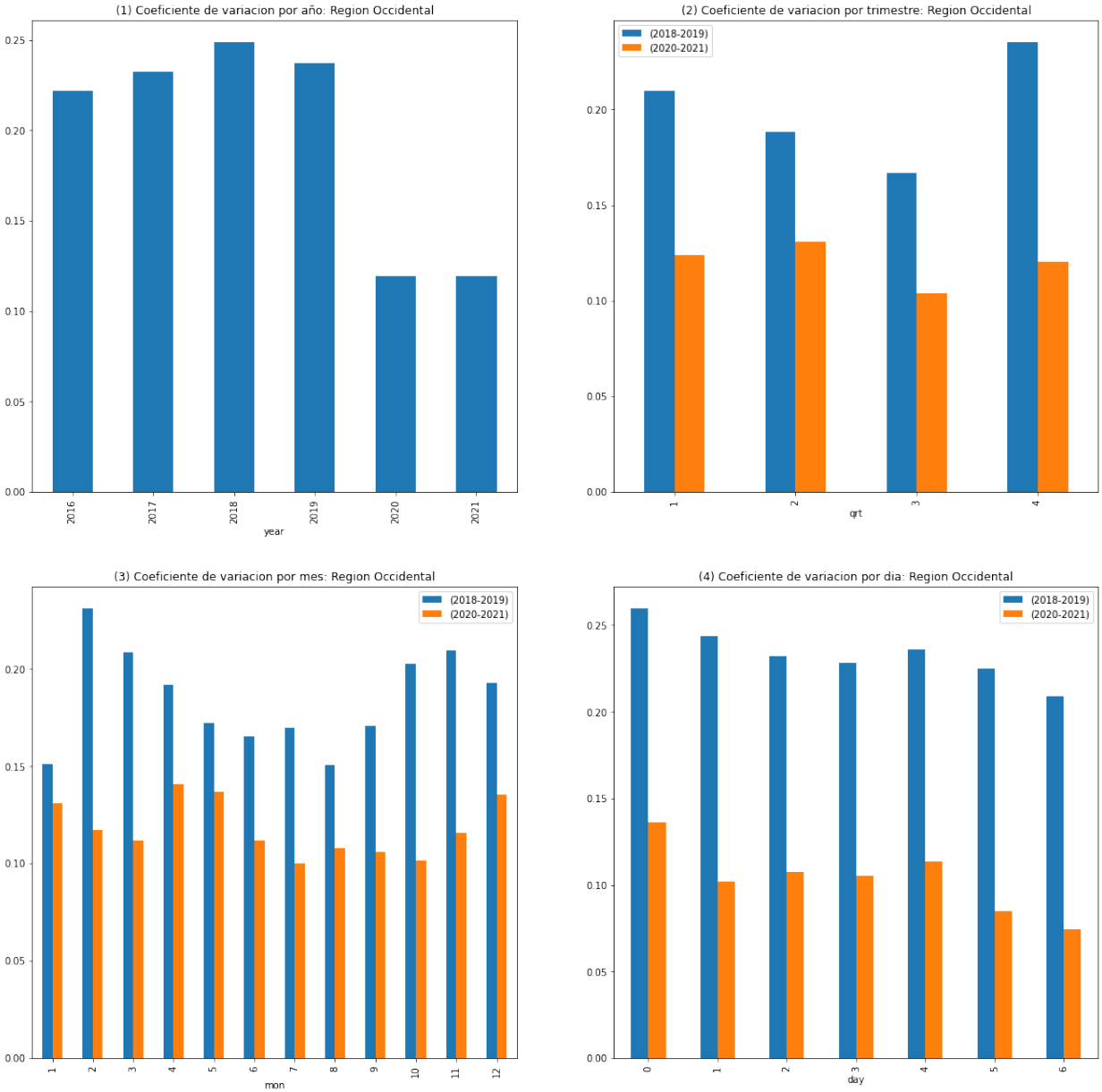


Figura 21: Coeficientes de variación de la región Occidente

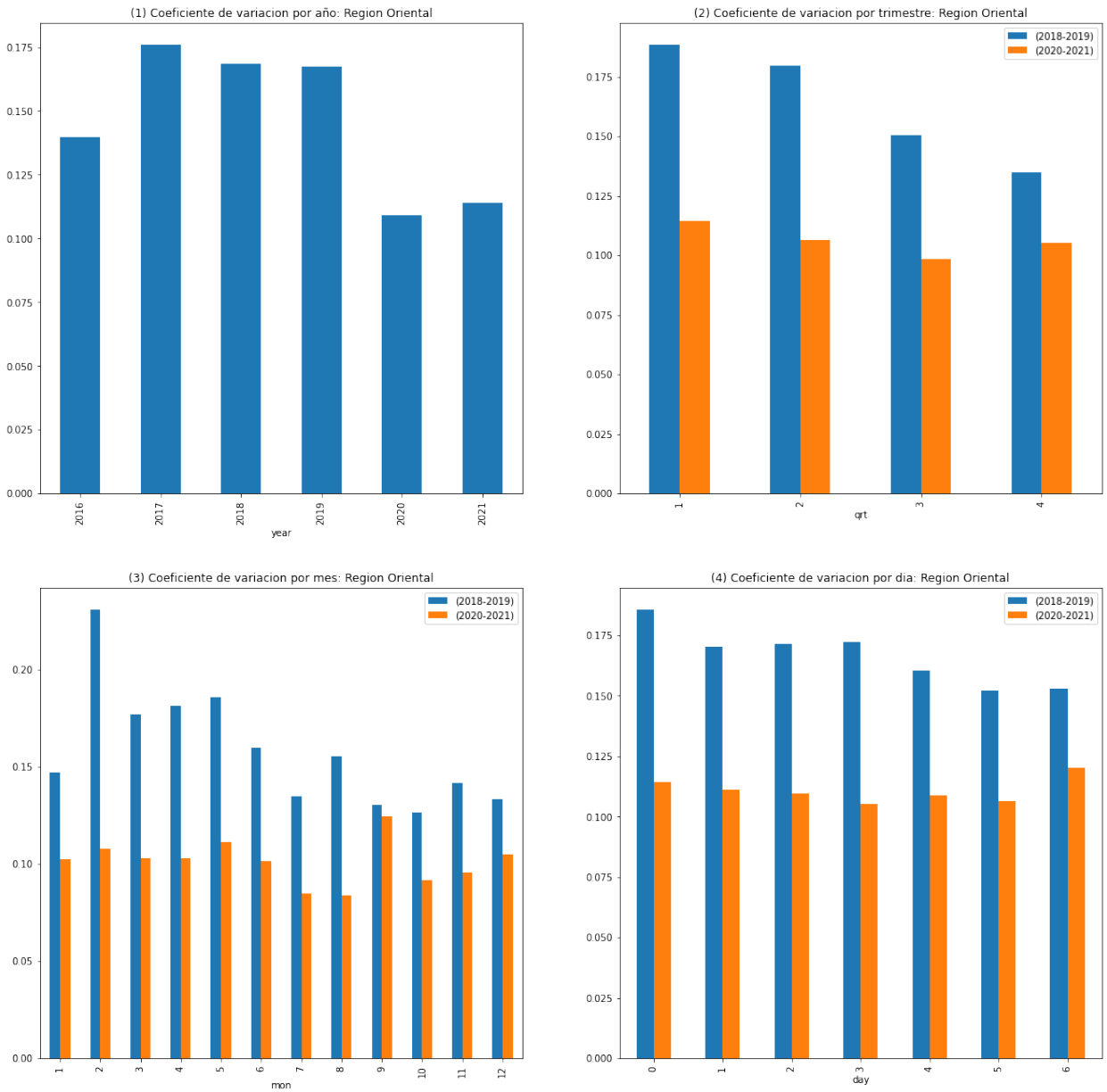


Figura 22: Coeficientes de variación de la región Oriente

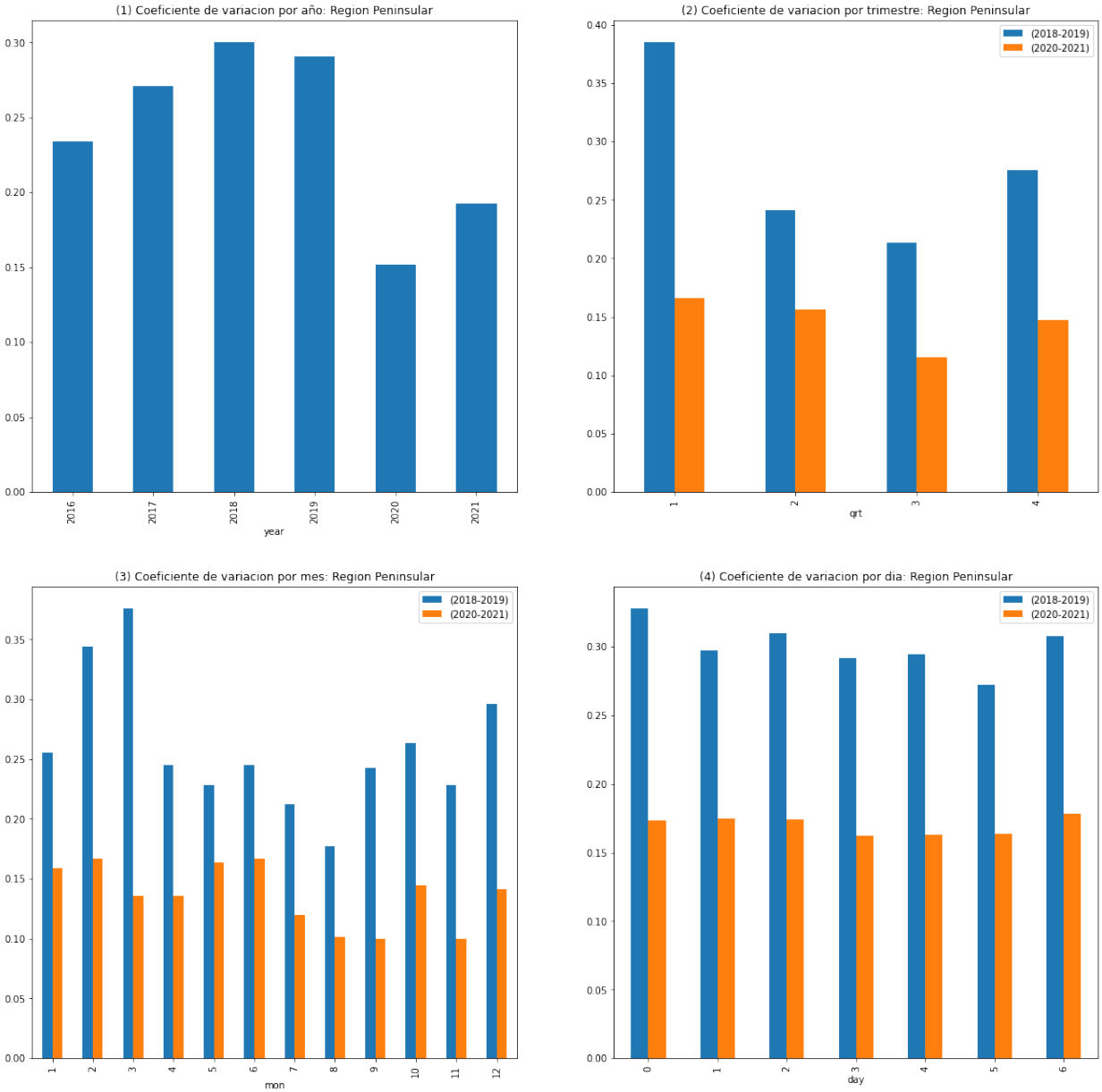


Figura 23: Coeficientes de variación de la región Peninsular

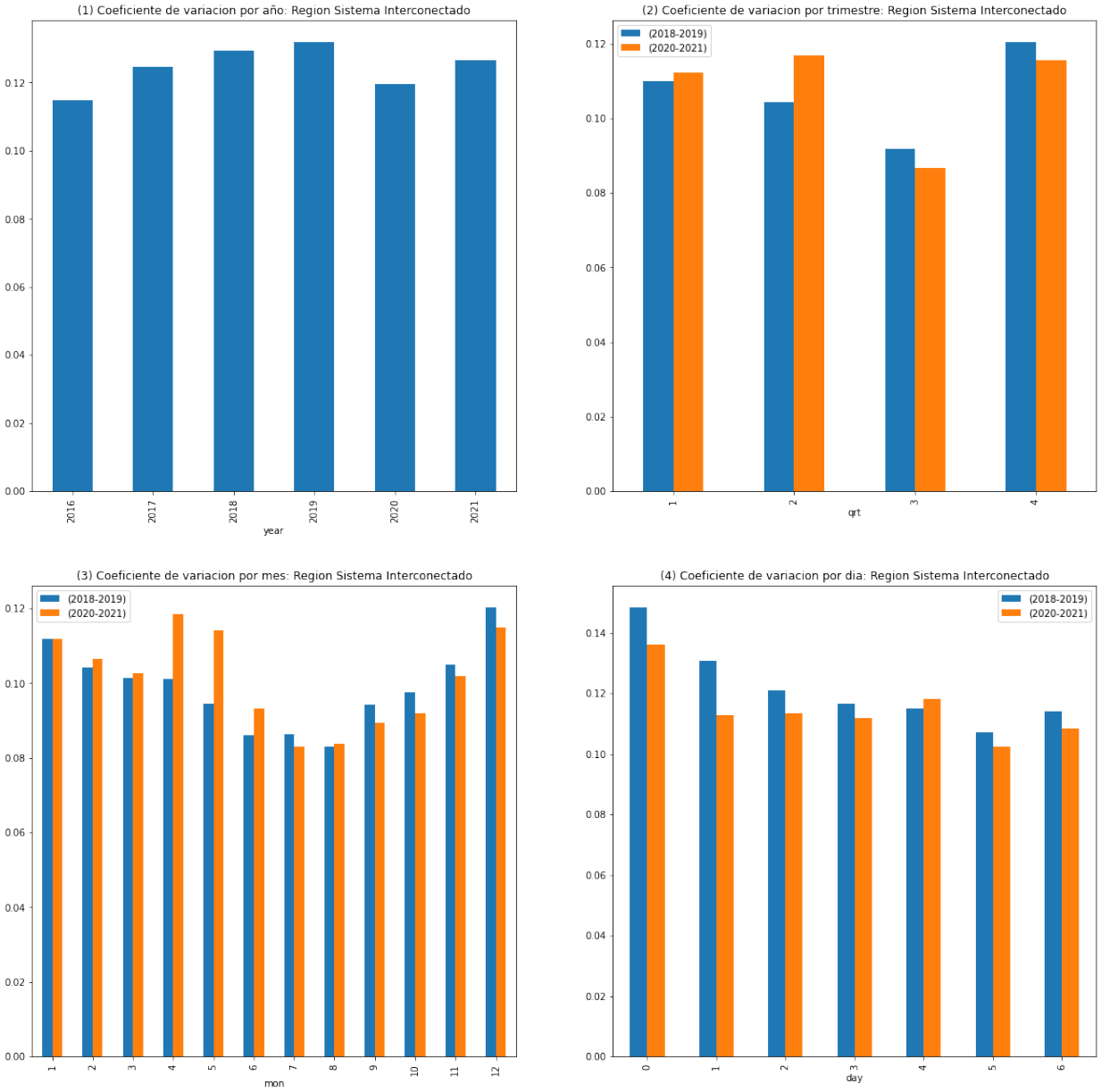


Figura 24: Coeficientes de variación del Sistema Interconectado

D) Analisis de estacionalidad con diagramas de cajas para el SIN y sus regiones

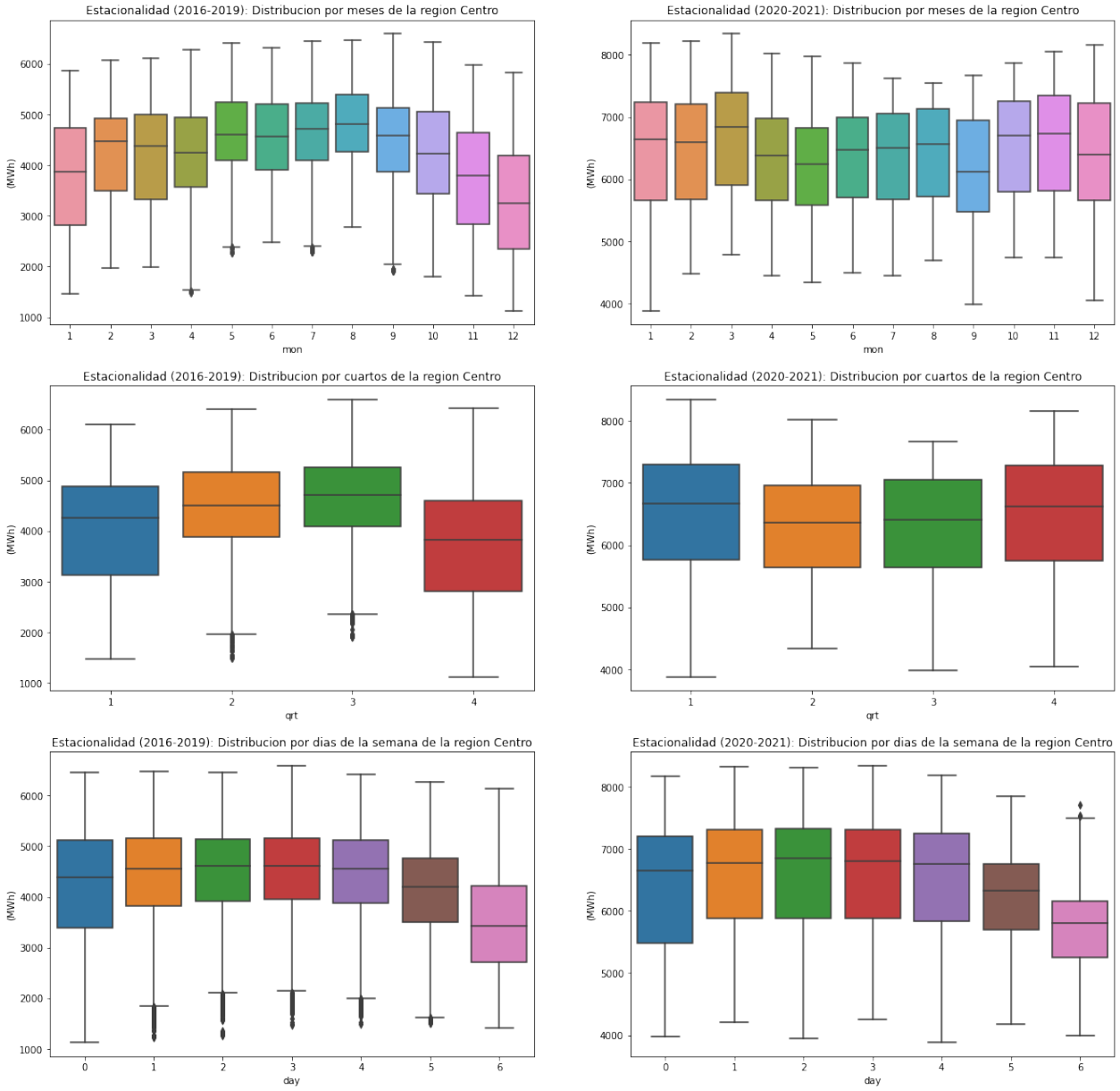


Figura 25: Diagramas de caja para el análisis de estacionalidad en la región Centro

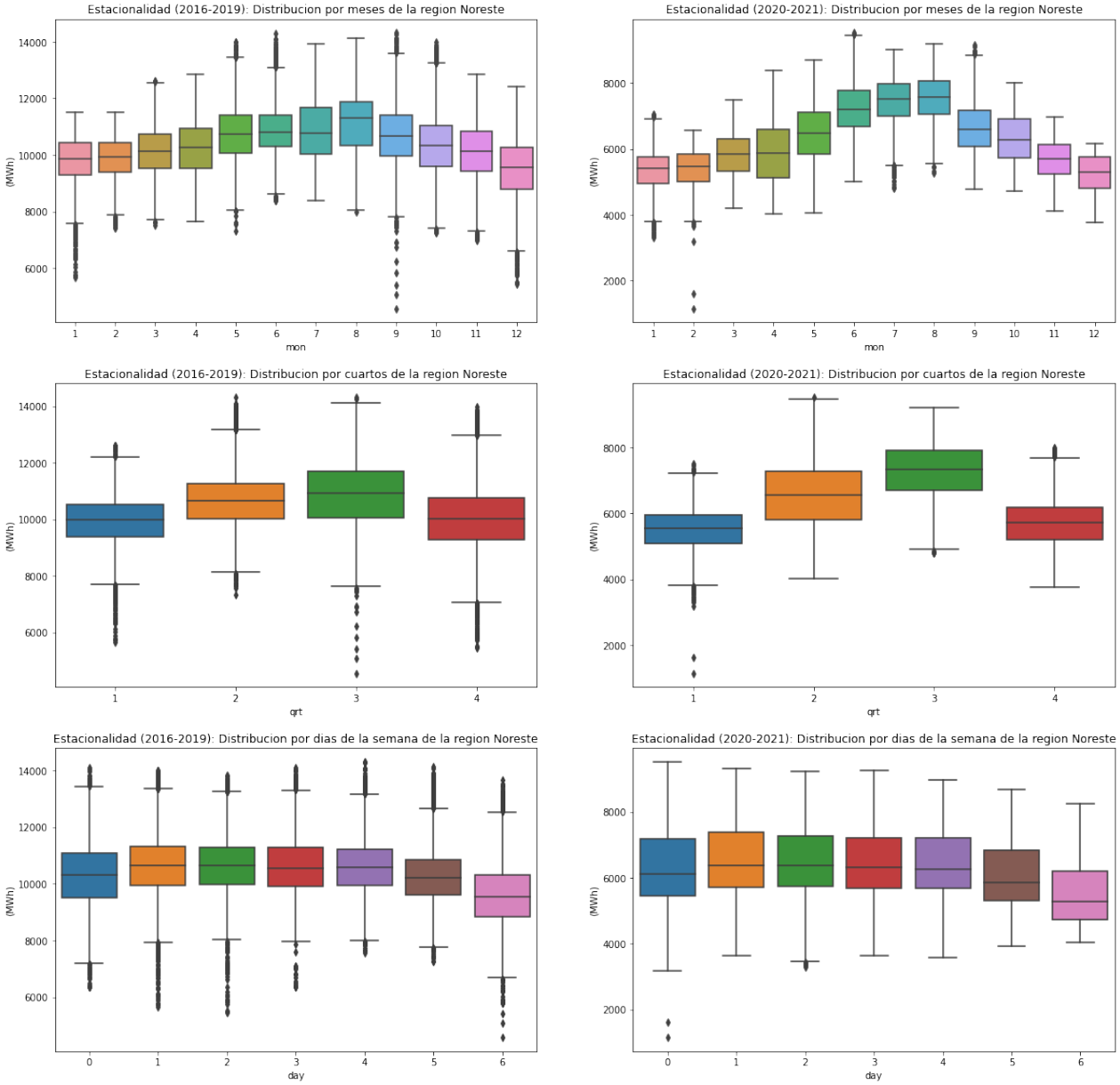


Figura 26: Diagramas de caja para el análisis de estacionalidad en la región Noreste

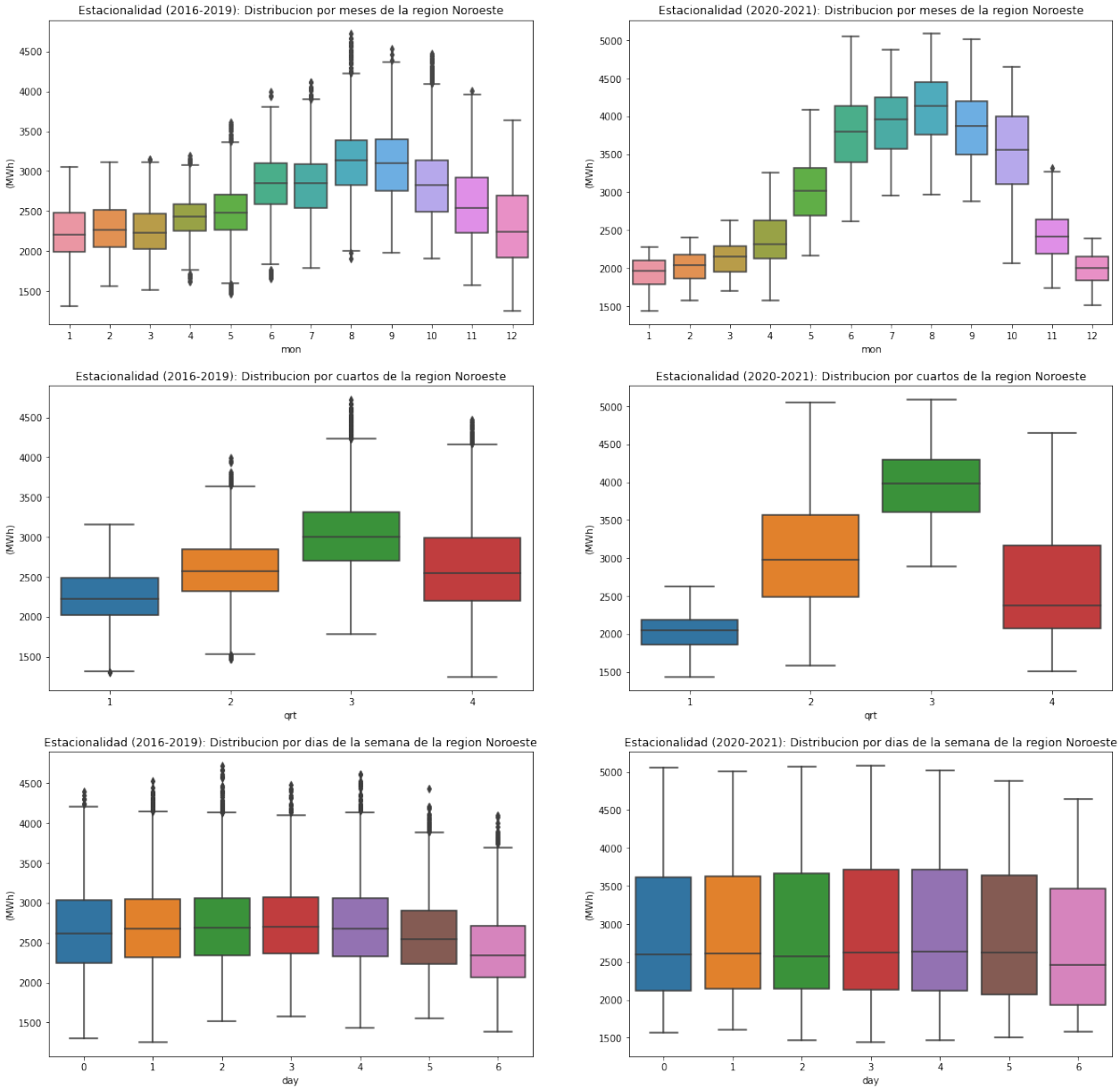


Figura 27: Diagramas de caja para el análisis de estacionalidad en la región Noroeste

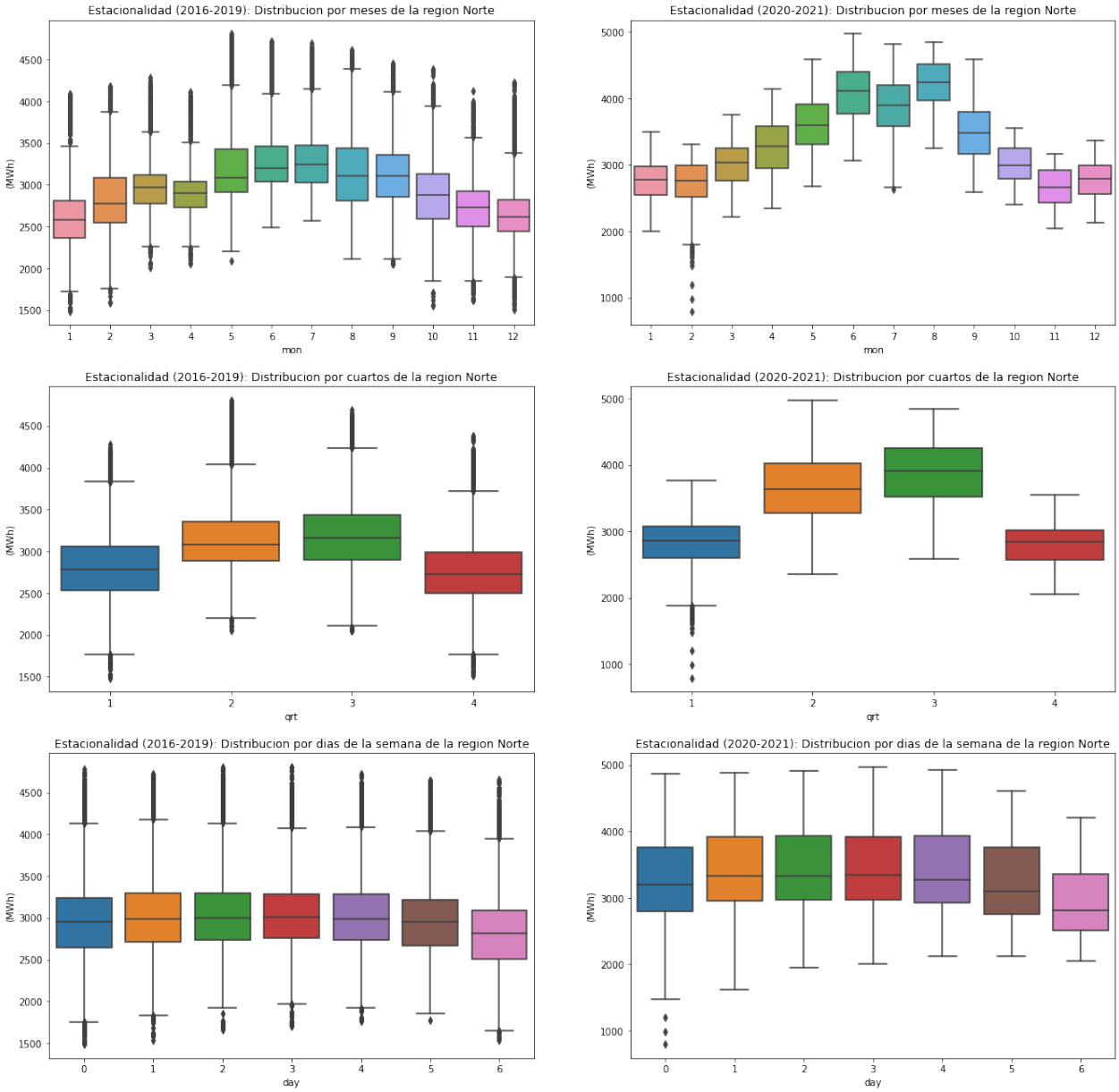


Figura 28: Diagramas de caja para el análisis de estacionalidad en la región Norte

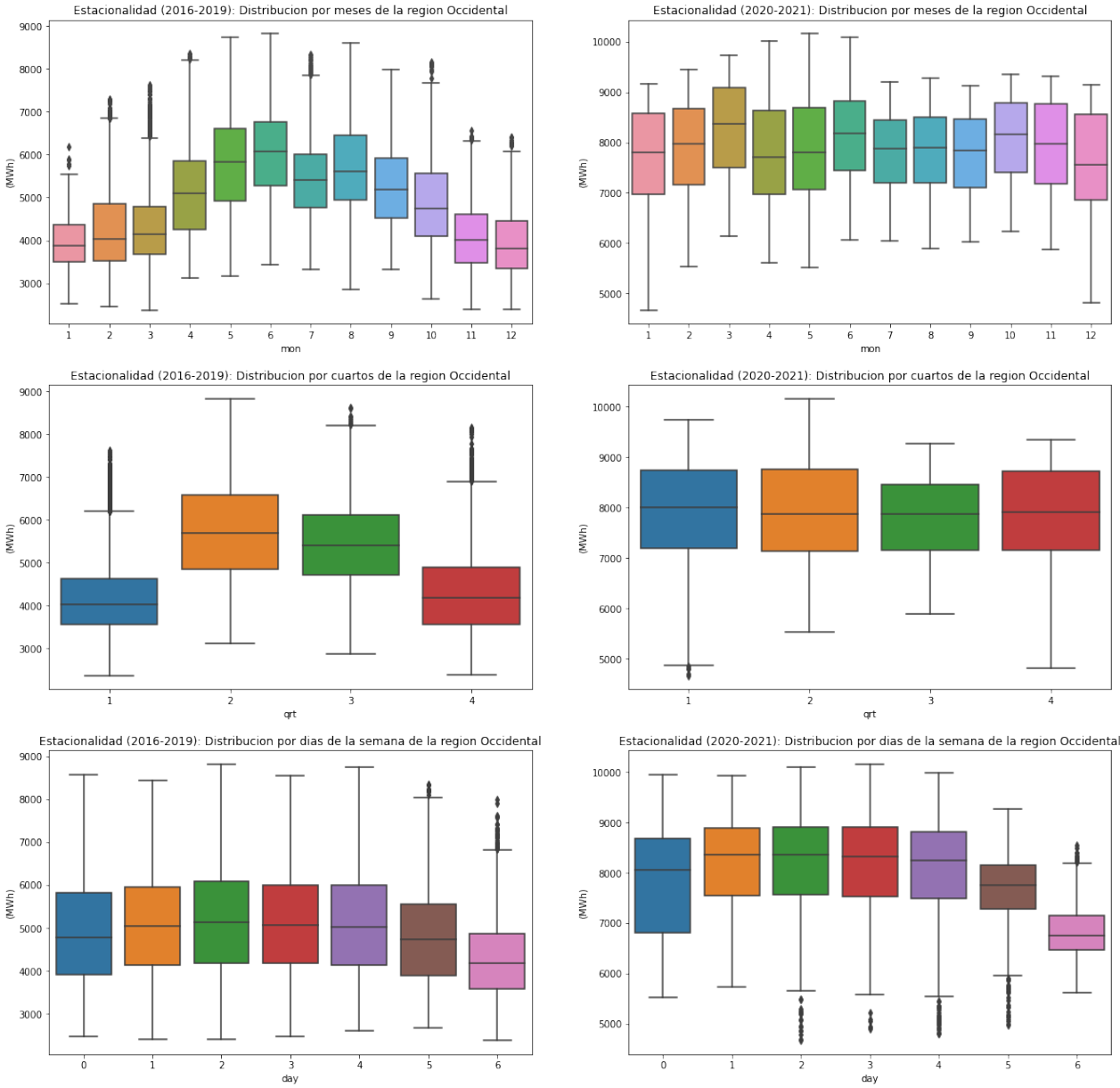


Figura 29: Diagramas de caja para el análisis de estacionalidad en la región Occidente

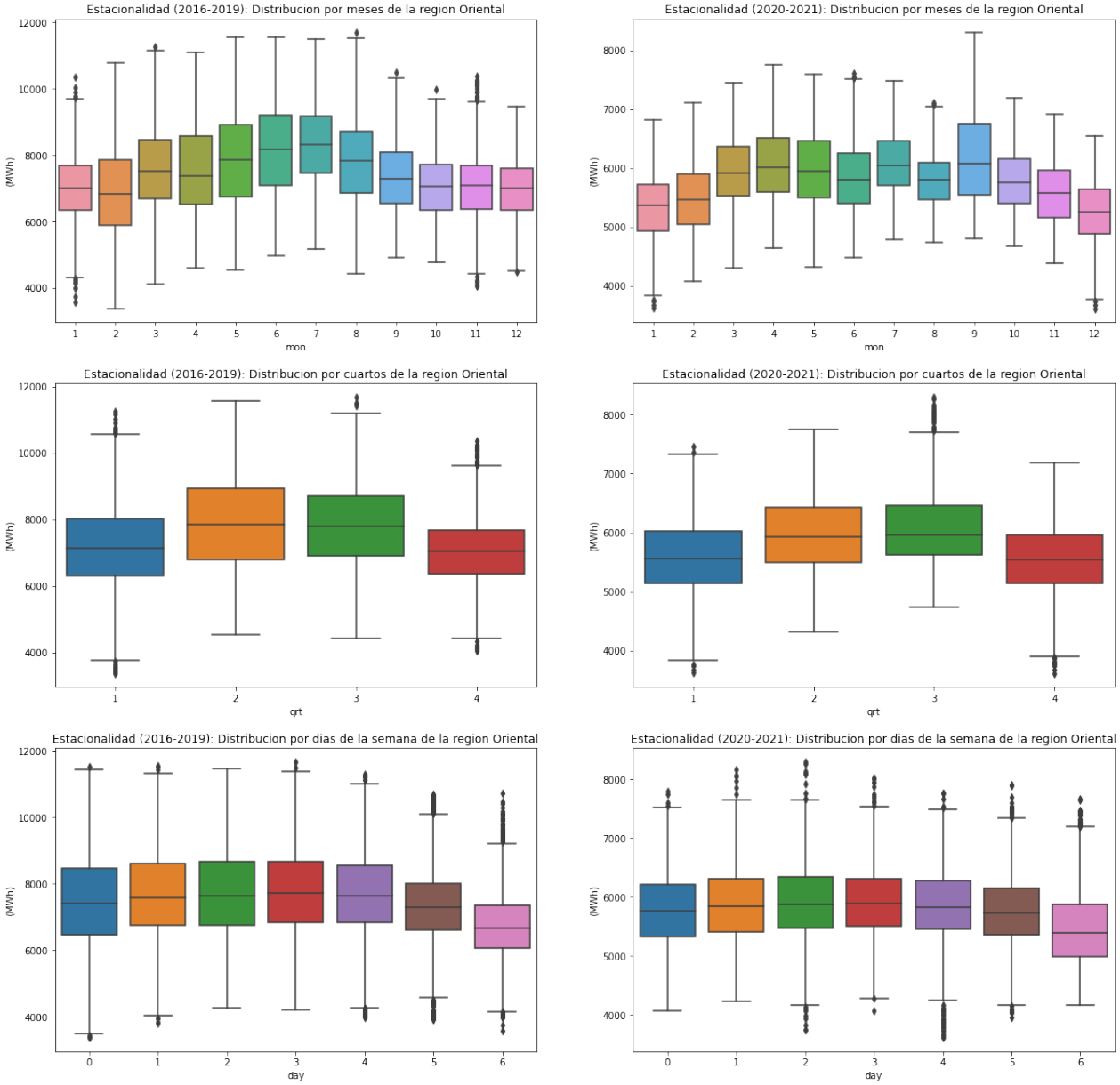


Figura 30: Diagramas de caja para el análisis de estacionalidad en la región Oriente

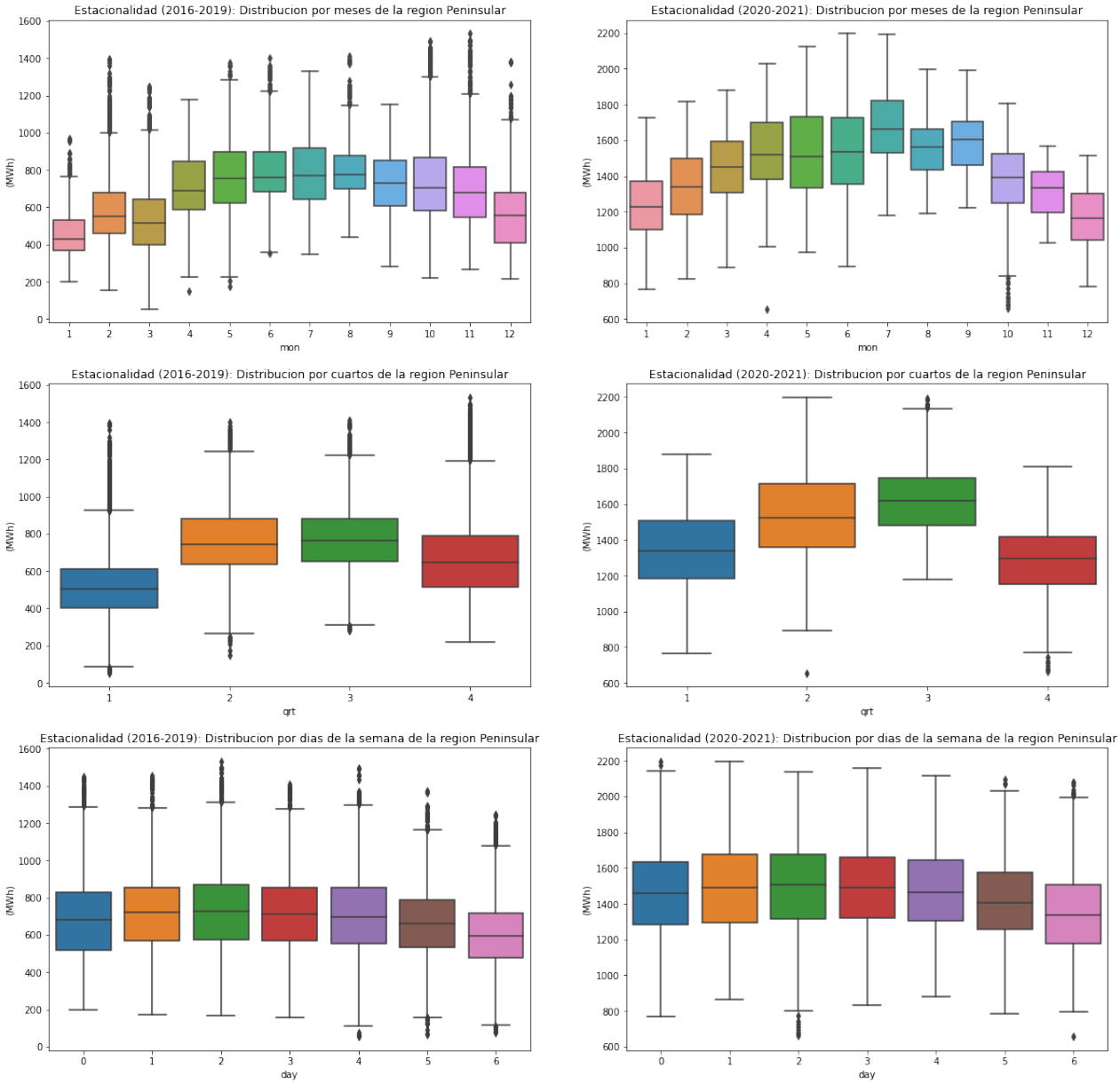


Figura 31: Diagramas de caja para el análisis de estacionalidad en la región Peninsular

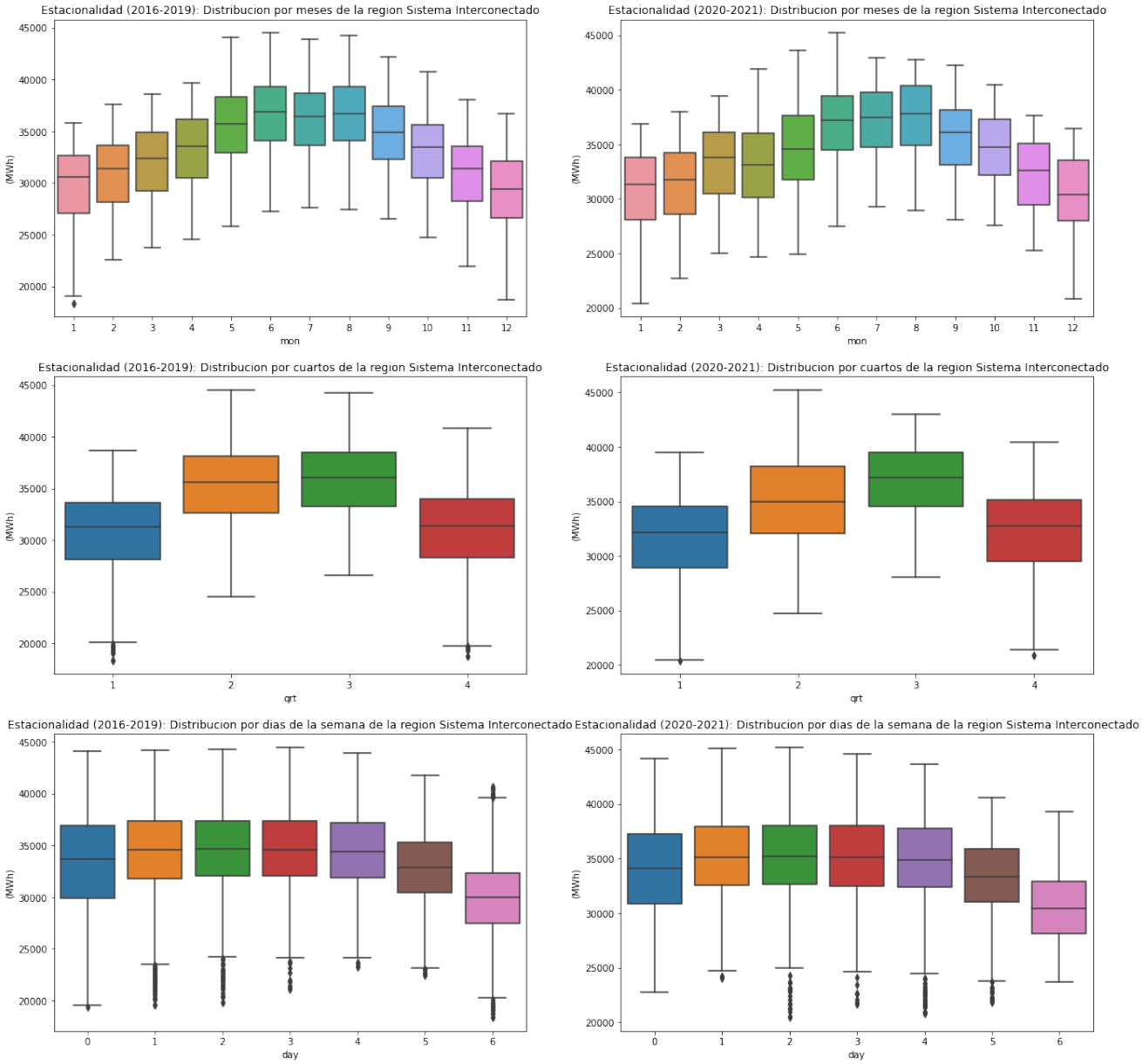


Figura 32: Diagramas de caja para el análisis de estacionalidad en el Sistema Interconectado

E) Analisis de tendencia para el SIN y sus regiones

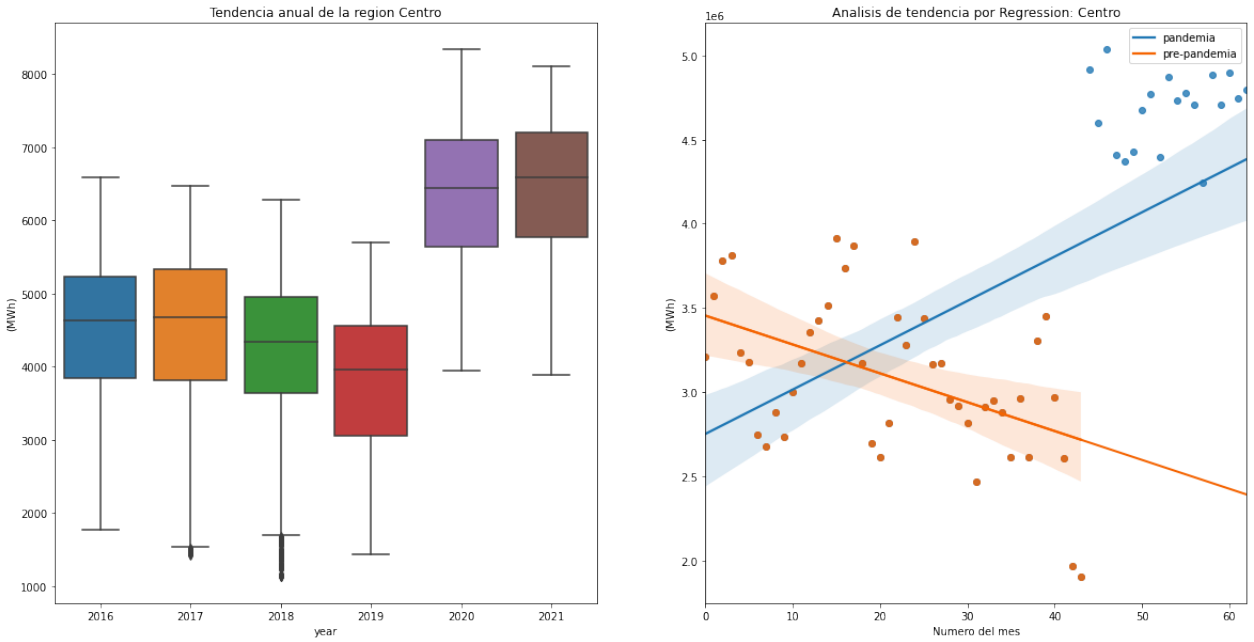


Figura 33: Análisis de tendencia para la región Centro

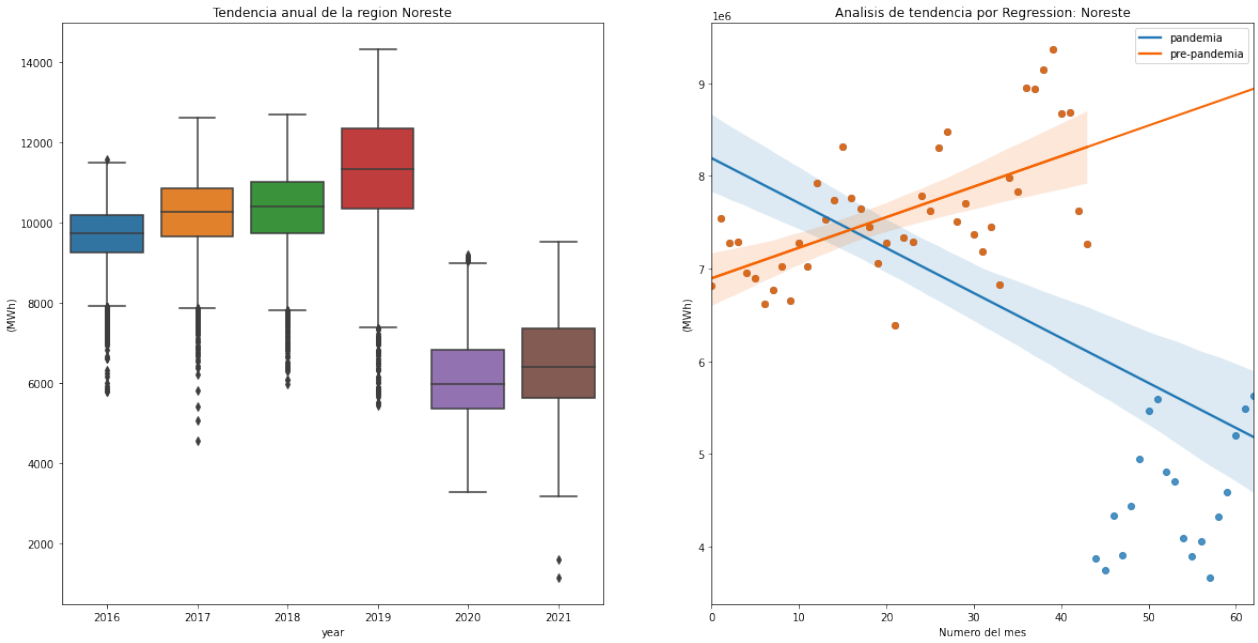


Figura 34: Análisis de tendencia para la región Noreste

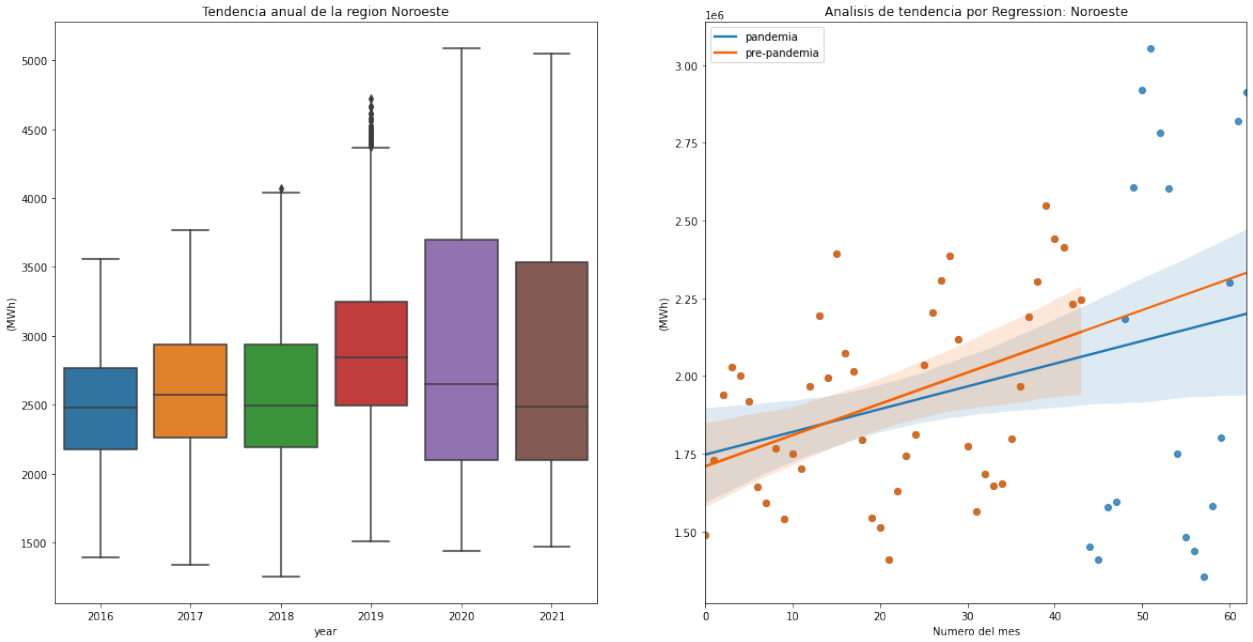


Figura 35: Análisis de tendencia para la región Noroeste

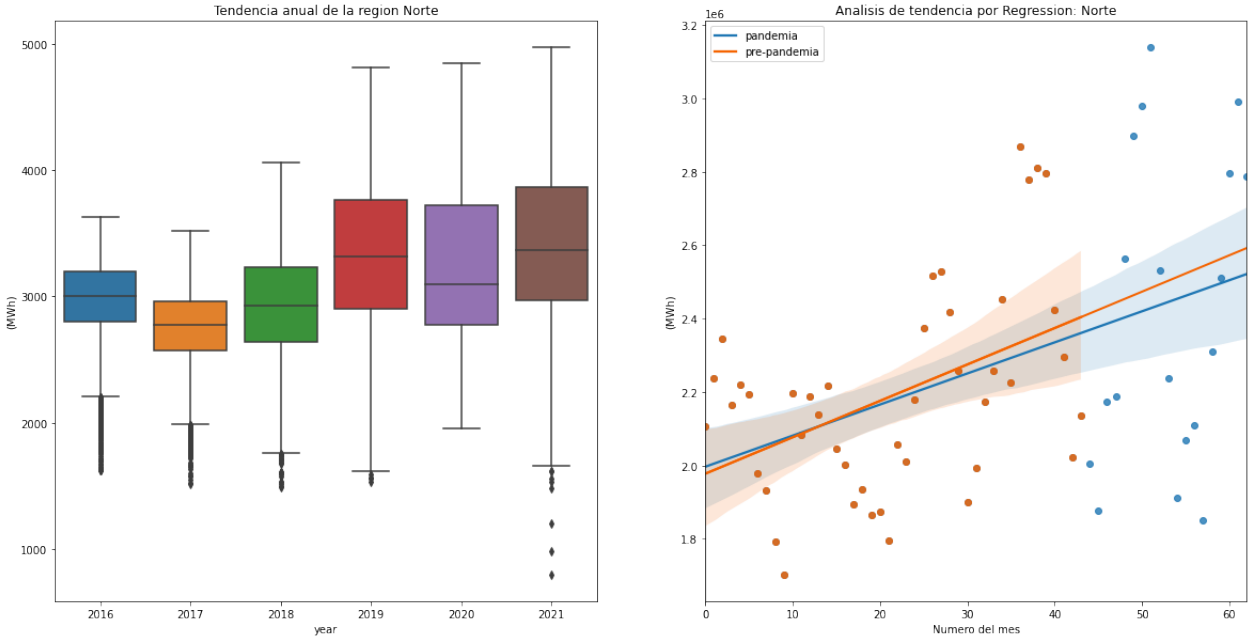
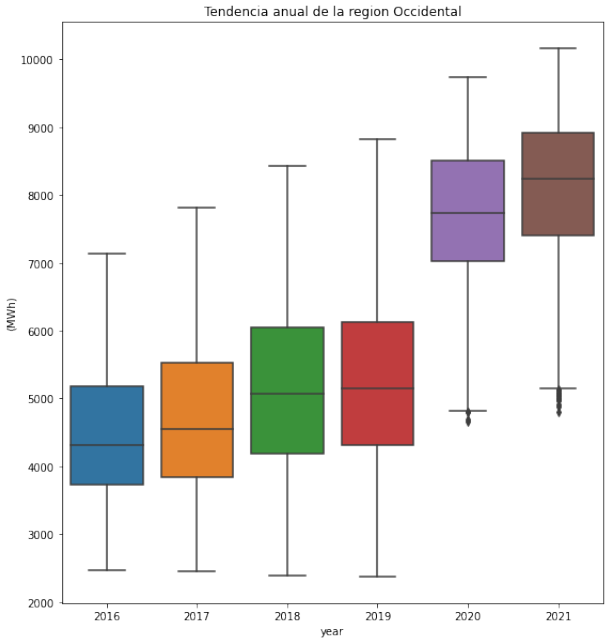


Figura 36: Análisis de tendencia para la región Norte



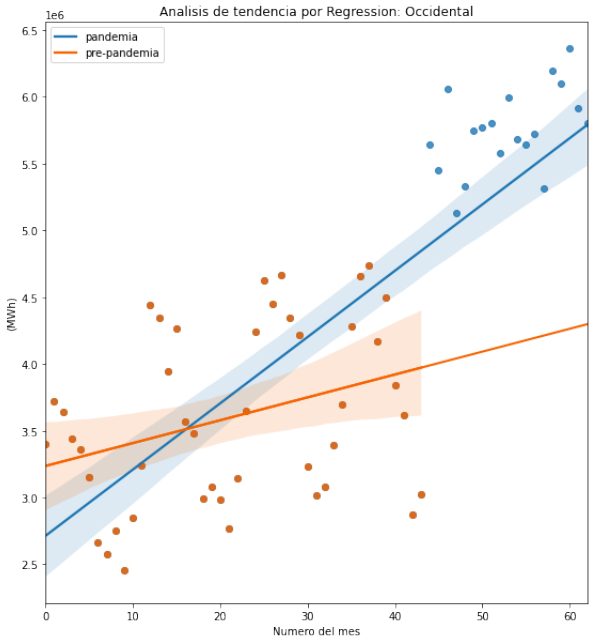
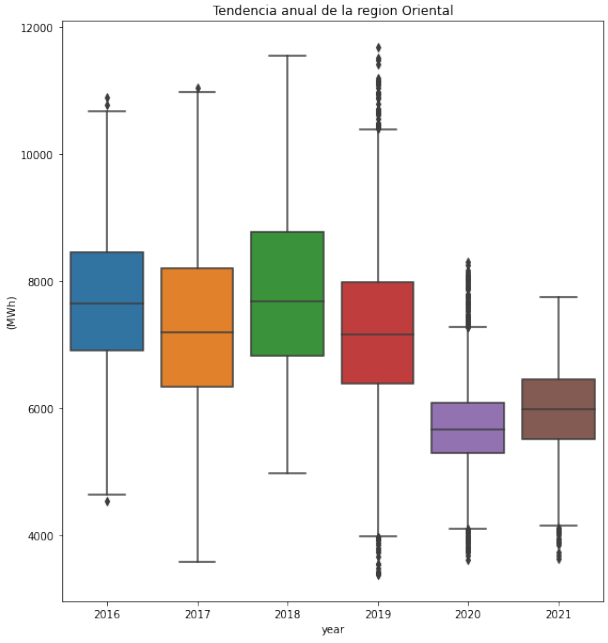


Figura 37: Análisis de tendencia para la región Occidental



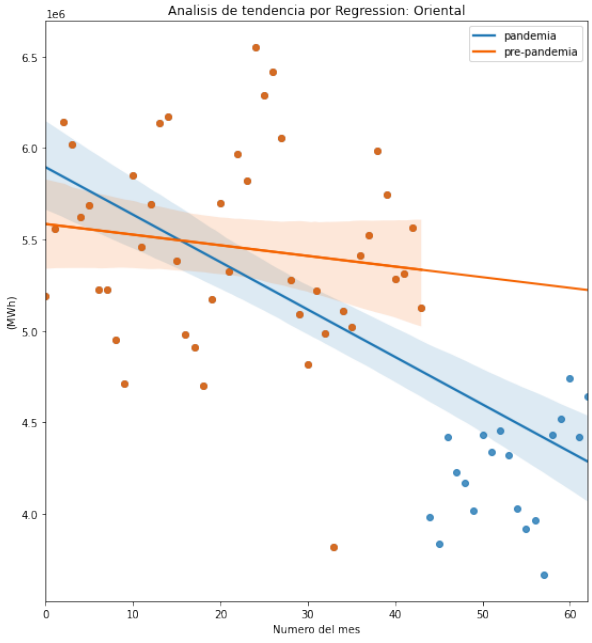


Figura 38: Análisis de tendencia para la región Oriental

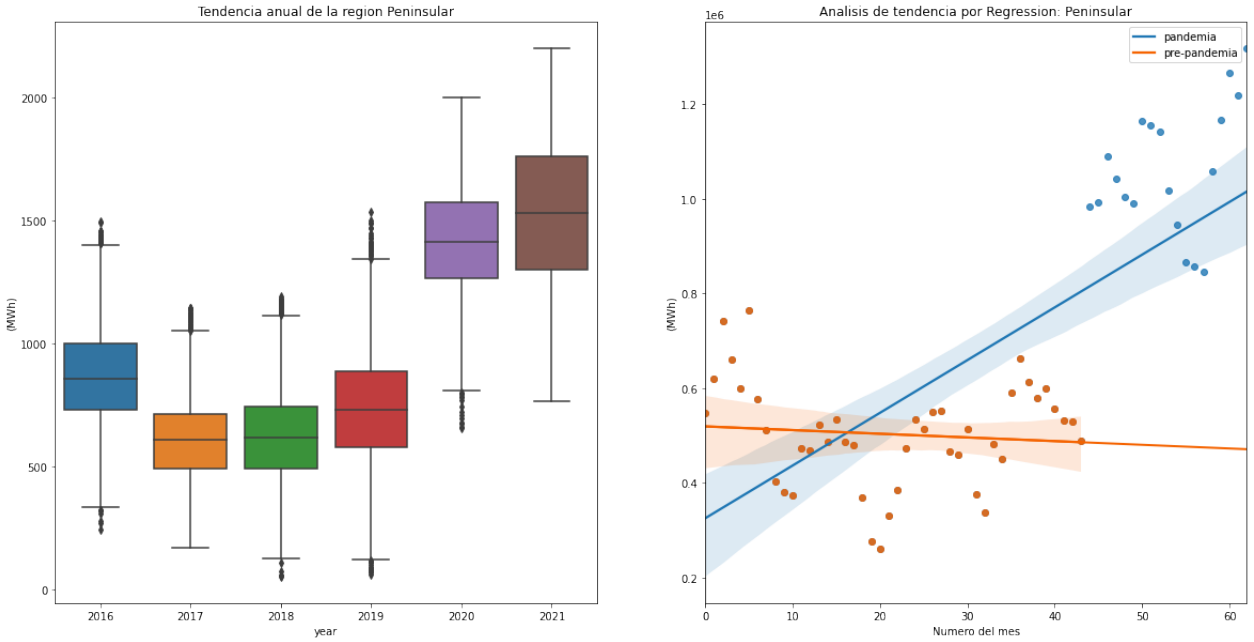


Figura 39: Análisis de tendencia para la región Peninsular

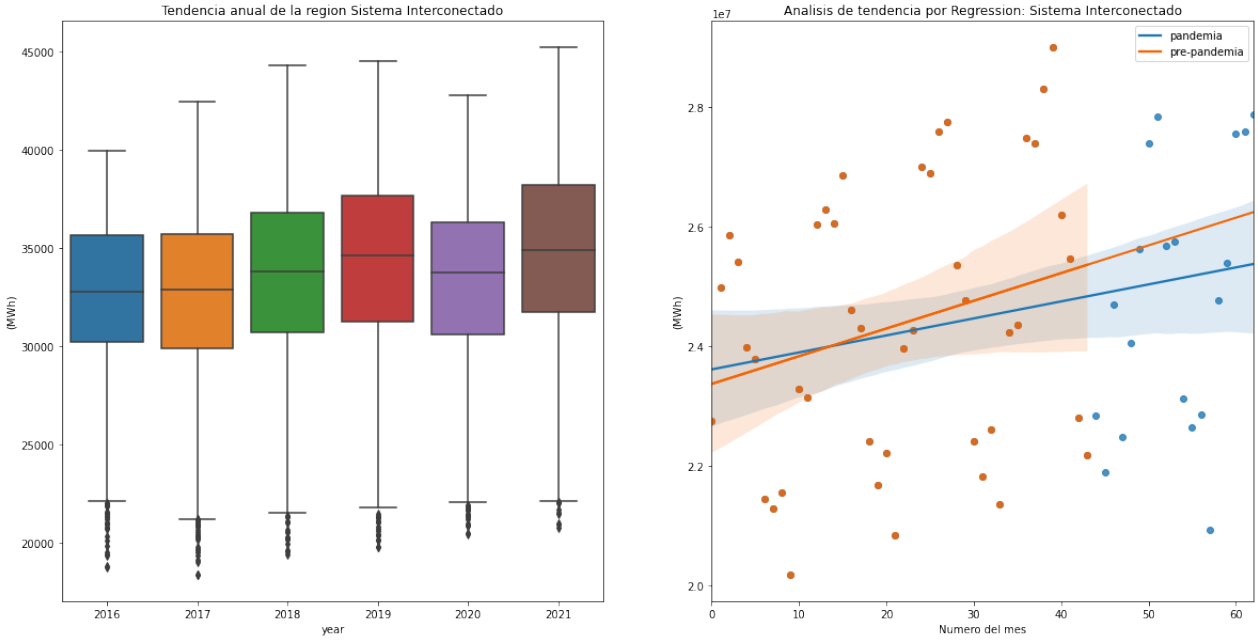


Figura 40: Análisis de tendencia del Sistema Interconectado

F) Graficas de autocorrelacion de las ultimas 24hrs del SIN y sus regiones

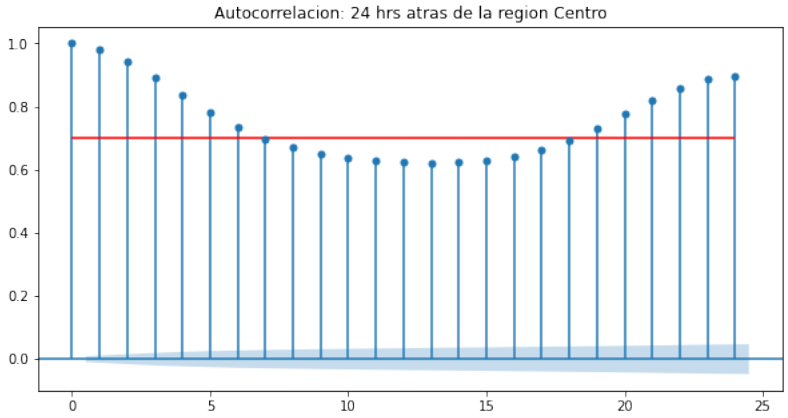


Figura 41: Grafica de autocorrelación de la region centro con 24 retrasos

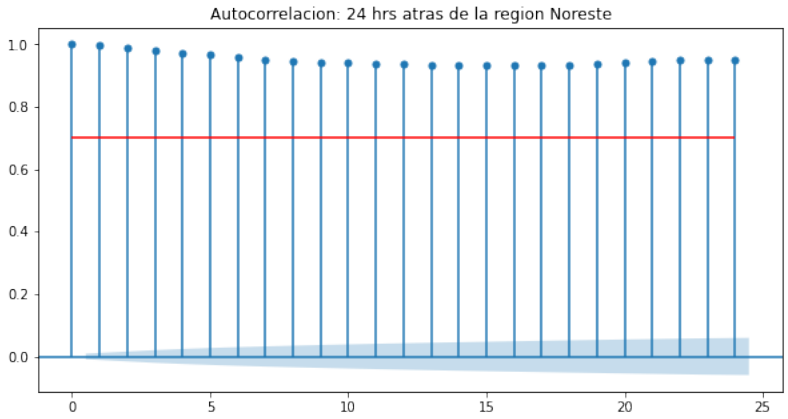


Figura 42: Grafica de autocorrelación de la region noreste con 24 retrasos

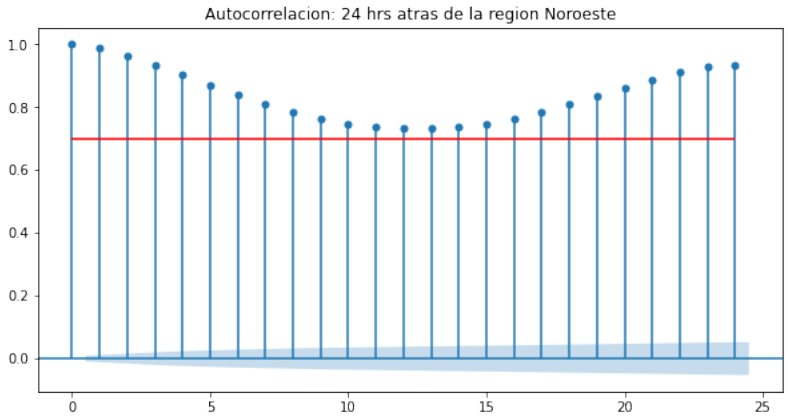


Figura 43: Grafica de autocorrelación de la region noroeste con 24 retrasos

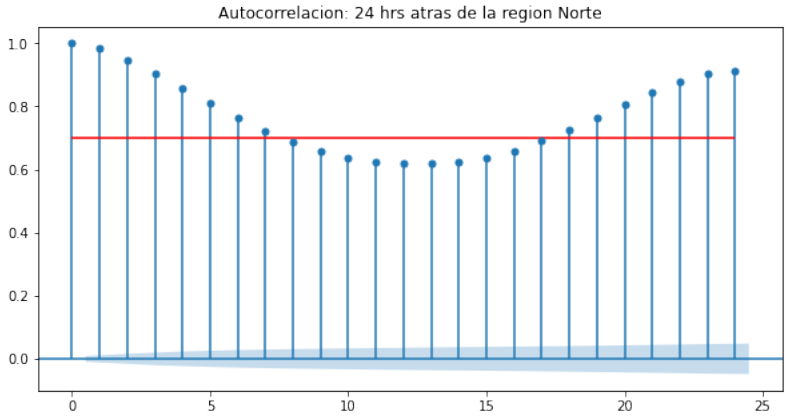


Figura 44: Grafica de autocorrelación de la region Norte con 24 retrasos

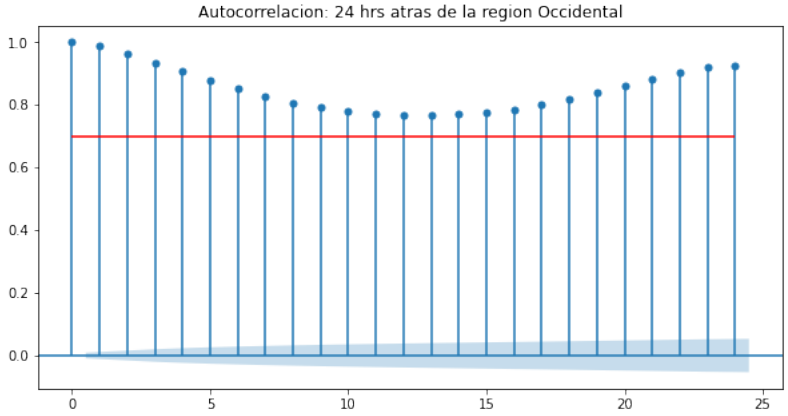


Figura 45: Grafica de autocorrelación de la region occidente con 24 retrasos

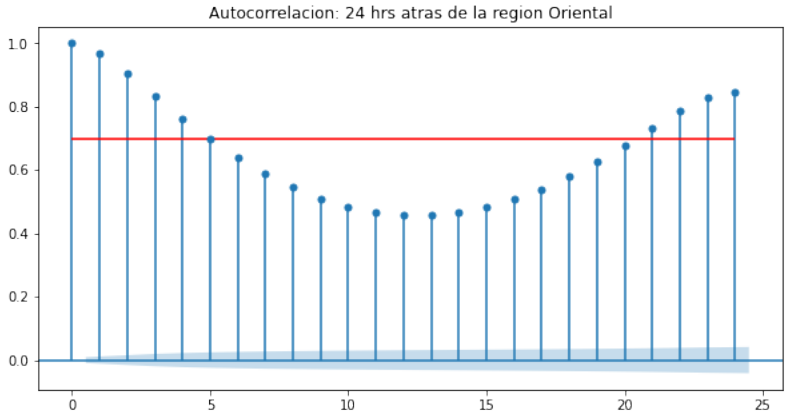


Figura 46: Grafica de autocorrelación de la region occidente con 24 retrasos

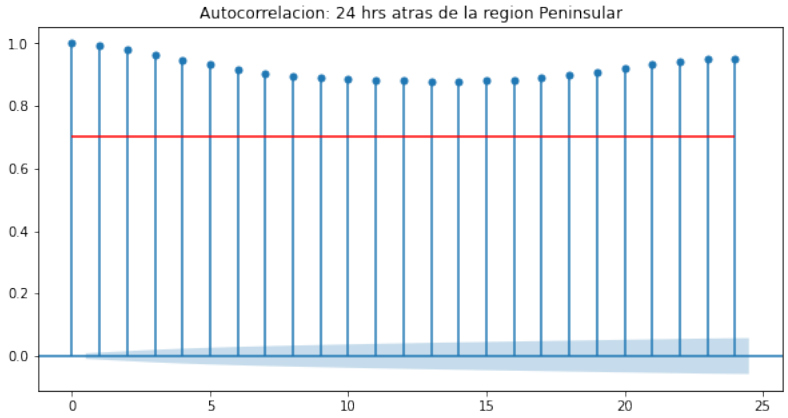


Figura 47: Grafica de autocorrelación de la region peninsular con 24 retrasos

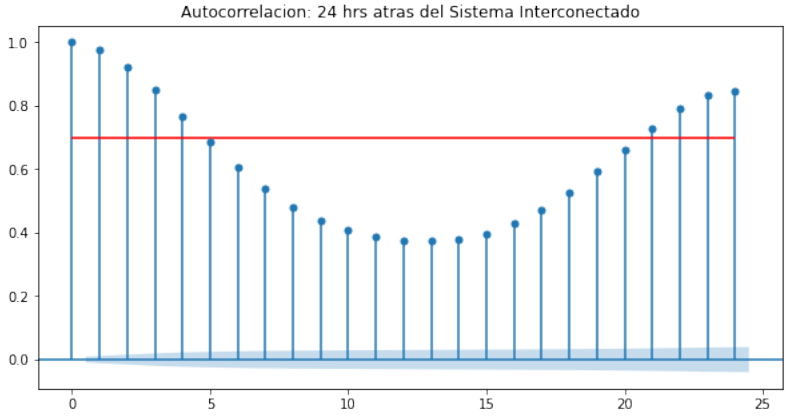


Figura 48: Grafica de autocorrelación del SIN con 24 retrasos