UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE COMERCIO Y TURISMO

MASTER EN BIG DATA & DATA SCIENCE

APLICACIONES AL COMERCIO, EMPRESA Y FINANZAS



TRABAJO FINAL DE MÁSTER

2022 – 2023

**MODELO DE ESTIMACIÓN DE LA DEMANDA DE ENERGÍA O ELECTRICIDAD SECTORIAL DE UN PAÍS**

Autores: (GRUPO 1)

* Ruth Fernández Padilla
* Ángel Martínez Barrial
* Alejandro Minguez Bonache
* Eduardo Urrutia Rivas
* Nam Nguyen Thi
* Antonio Villamayor Delgado

Supervisor:

* Tutor

Septiembre, 2023

**Declaration of Authorship**

MRC Consultants and Transaction Advisers

Signed:

Date: September, 2023

*“Los datos no son solo información, son la clave para el conocimiento y la toma de decisiones informadas en la era digital”*

Peter Drucker

**Abstract**

The purpose of this project was to predict energy consumption in Brazil up to year 2040 using data from matrices and balances published on the official website of OLADE. The objective of this project was to test whether a machine learning model could produce sufficiently accurate results for a complex forecasting problem. We explored various machine learning techniques and developed a data-driven model for energy consumption forecasting. The dataset included a 52-year record (1970-2021), and we used an ARIMA model to train the data. The model's performance was evaluated using the mean absolute percentage error (MAPE) to ensure direct comparability with energy readings in the dataset. The results indicate that machine learning algorithms can effectively predict energy consumption. These findings can be applied to other countries in South America for energy consumption forecasting.

**ÍNDICE**

[**1. Introducción 4**](#_heading=h.gjdgxs)

[Motivación y objetivos 4](#_heading=h.30j0zll)

[Estructura del documento 4](#_heading=h.1fob9te)

[**2.  Análisis descriptivo de la demanda de energía eléctrica 4**](#_heading=h.3znysh7)

[Análisis descriptivo del conjunto de datos. 4](#_heading=h.2et92p0)

[Factores económicos / no económicos que influyen en la demanda. 4](#_heading=h.tyjcwt)

[Exploración y limpieza del conjunto de datos (tratamiento de los valores nulos).](#_heading=h.3dy6vkm) 4

Selección de variables relevantes para el estudio. [4](#_heading=h.3dy6vkm)

[Preprocesamiento de datos. 4](#_heading=h.1t3h5sf)

[**3. Desarrollo del modelo de Machine Learning 4**](#_heading=h.4d34og8)

[Introducción a los modelos desarrollados 4](#_heading=h.2s8eyo1)

[Selección el modelo óptimo 4](#_heading=h.17dp8vu)

[**4. Análisis y visualización de los resultados 4**](#_heading=h.3rdcrjn)

[Estimaciones y predicciones de la demanda de energía sectorial](#_heading=h.26in1rg) 4

Evaluación del rendimiento de los modelos estimados y análisis de los errores. [4](#_heading=h.26in1rg)

[Visualización e interactiva aplicación 4](#_heading=h.lnxbz9)

[**5. Conclusiones 4**](#_heading=h.35nkun2)

[Conclusiones obtenidas 4](#_heading=h.1ksv4uv)

[Limitaciones y futuras líneas investigaciones 4](#_heading=h.44sinio)

[**6. Recomendaciones de negocio. 4**](#_heading=h.2jxsxqh)

[**Referencias bibliográficas/ Anexos**](#_heading=h.z337ya)

1. **Introduction**

Global energy demand for energy consumption is increasing day by day, and it seems complicated for most countries to meet energy demand with total energy production.

In an increasingly tense world, BRICS countries－Brazil, Russia, India, China and South Africa－are very complementary with the member countries being the world's major energy producers as well as major fuel consumers.

The Brazilian energy sectors have witnessed numerous technological changes and has evolved to become a global leader in clean technology sales, both to the domestic and foreign market. A lot of factors contributed to the innovative activities in its electricity sector which includes both government and FDI contribution.

[Graphic of energy consumption by energy source 1975 – 2021]

In recent years, artificial intelligence in general and machine learning in specific present solutions for a huge number of problems. One of the promising applications is predicting future energy consumption using historic time series by acquiring its imitation model.

This project aims to test whether it is possible to apply the model ARIMA to the time series data. The aim is to yield good enough results in forecasting energy consumption.

**3. Fundamental**

3.1 Time series

Time series data tracks values over a specific period, typically with time as the independent variable, often used for future forecasting. It can be categorized as Continuous Time Series and Discrete Time Series based on the time intervals. Additionally, time series data can be divided into univariate (tracking a single variable over time) and multivariate (tracking multiple variables over time) based on the number of dependent variables.

3.2 Time series characteristics

Time series has three key characteristics crucial for accurate forecasting

* **Autocorrelation**: Measures the relationship between a variable's current value and its past values.
* **Seasonality**: Refers to regular, short-term variations that occur, such as weekly, monthly, or quarterly patterns.
* **Stationarity**: Implies that statistical properties of a time series remain constant over time, with a consistent mean and variance, and covariance independent of time.

3.3 Time Series Forecast Methods

Over the years, numerous studies have assessed the performance of classical and machine learning methods. Some well-known classical methods include ARMA (Autoregressive Moving Average), ARIMA (Autoregressive Integrated Moving Average), SARIMA (Seasonal Autoregressive Integrated Moving-Average), while modern algorithms like MLP (Multi-Layer Perceptron), BNN (Bayesian Neural Networks), GRNN (Generalized Regression Neural Networks)

Two other modern algorithms are Recurrent Neural Networks RNN (Recurrent Neural Networks), and LSTM (Long Short- Term Memory) have also gained prominence (Brownlee, 2016).

3.4 ARIMA (p,d,q)

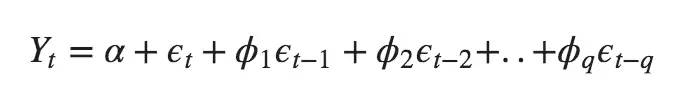
ARIMA stands for auto-regressive integrated moving average. It’s a way of modelling time series data for forecasting (i.e., for predicting future points in the series), in such a way that:

* **AR** : *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations.
* **I** : *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
* **MA** : *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

The parameters of the ARIMA model are defined as follows:

* **p**: The number of lag observations included in the model, also called the lag order.
* **d**: The number of times that the raw observations are differenced, also called the degree of differencing.
* **q**: The size of the moving average window, also called the order of moving average.

An ARIMA model is one where the time series was differenced at least once to make it stationary and you combine the AR and the MA terms. So the equation becomes:



To apply ARIMA model, follow these steps:

* **Data Prep**: Get the time series data in order.
* **Stationarize Data**: Remove trends and seasonality.
* **Identify Parameters**: Find the right model order (d, p, q).
* **Fit ARIMA**: Apply the ARIMA model to your data.
* **Check Residuals**: Ensure residuals resemble white noise.
* **Forecast**: Use the model for future predictions.
* **Evaluate**: Assess forecast accuracy and adjust if needed.
* **Validate**: Test on new data to confirm reliability.
* **Visualize**: Communicate results visually.
* **Deploy**: Put the model into action if it meets your requirements.

**2. Data processing**

**2.1. Data**

The dataset utilized for this study encompasses Brazil's energy matrix balance spanning from 1970 to 2021, sourced from the OLADE website ([https://www.olade.org/](https://www.olade.org/" \t "_new)). This comprehensive dataset is structured within an Excel file featuring multiple sheets. We imported this Excel file directly from a repository on [GitHub](https://github.com/NamNguyen2015/TFM/blob/main/datas/Option_B/Brazil_Energy%20balance%20matrix.xlsx) as a dictionary, with each key corresponding to a specific year within the 1970 to 2021 timeframe, after some data cleansing.

Each year, the data is organized as a DataFrame, with the columns representing the energy sources, and the rows depicting the sectors where the energies are distributed.

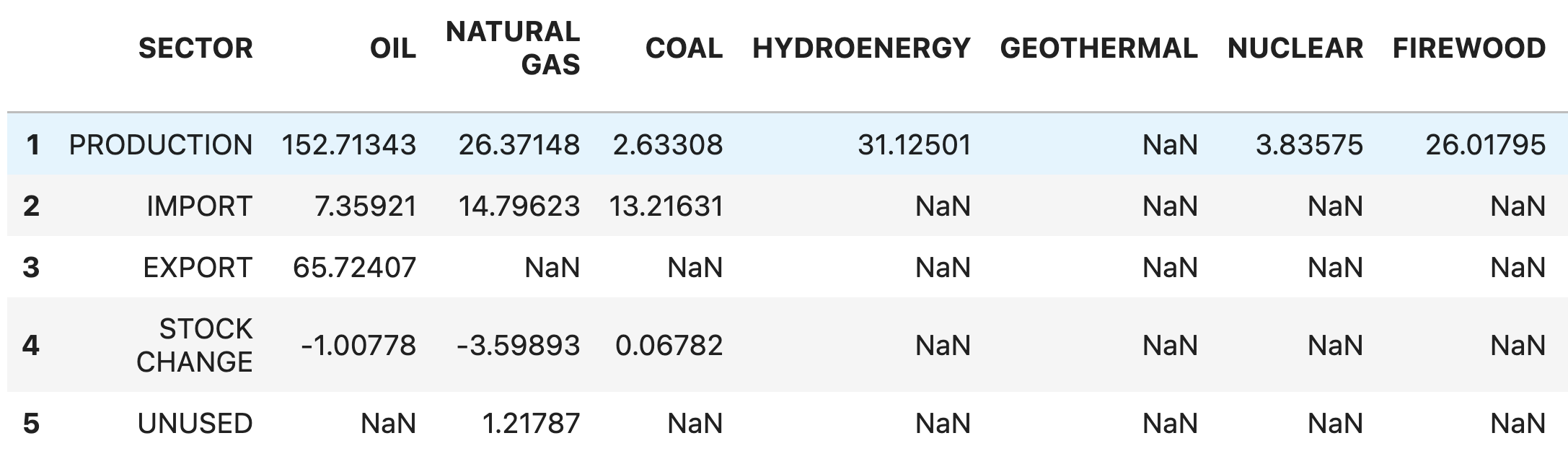


Figure 3: Some first rows and columns in each data frame.

All columns in the dataset are important for our study due to their role in providing a comprehensive overview of Brazil's energy matrix balance from 1970 to 2021. Each column represents a specific aspect of energy production, consumption, and related factors, including various energy sources and sectors.

While outliers exist and can influence model outcomes, they have been retained for analysis to capture potential valuable insights or anomalies in the energy data. The maximum and minimum values generally align with the dataset's central tendencies, suggesting that they are within reasonable bounds. Understanding and considering these aspects are crucial for robust data analysis and model development.

The presence of NaN (Not-a-Number) values in the database is primarily due to missing or unavailable information. In this dataset, the NaN values are particularly abundant because they represent cases where the actual values ware not recorded or not available. Therefore, these missing values have been represented as NaN to indicate the absence of data.

**2.2 Data Pre-processing**

We performed actions such as rounding values, renaming columns, removing spaces, and correcting names. Additionally, we added the year as a new column. We converted units of measurement from Ktoe (kilotonnes of oil equivalent) to Mtoe (megatonnes of oil equivalent) in all numeric columns. Numeric values were rounded to two decimal places. Row headers containing units of measurement were removed, and the columns 'Unnamed: 0' were renamed to 'SECTOR,' and 'OTHER PRIMARY\_x000d\_' was renamed to 'OTHER PRIMARY.' Furthermore, the sector name 'COKE PLANTS AND BLAST FURNACES\_x000d\_' was corrected to 'COKE PLANTS AND BLAST FURNACES.' These transformations ensure that the data is in a consistent format and ready for further analysis. The processed data is stored in a new dictionary called 'modified\_dict,' which replaces the original one. Finally, the cleaned dictionary is saved in a new Excel file named "Brazil\_Energy balance matrix\_cleaned.xlsx," preserving the original sheets for subsequent analysis and visualization. This process ensures data quality and consistency before its utilization in more in-depth analyses.

**2.3. First visualization of original data**

This project tries to answer the question if it’s possible to create an ARIMA model to predict future energy consumption.

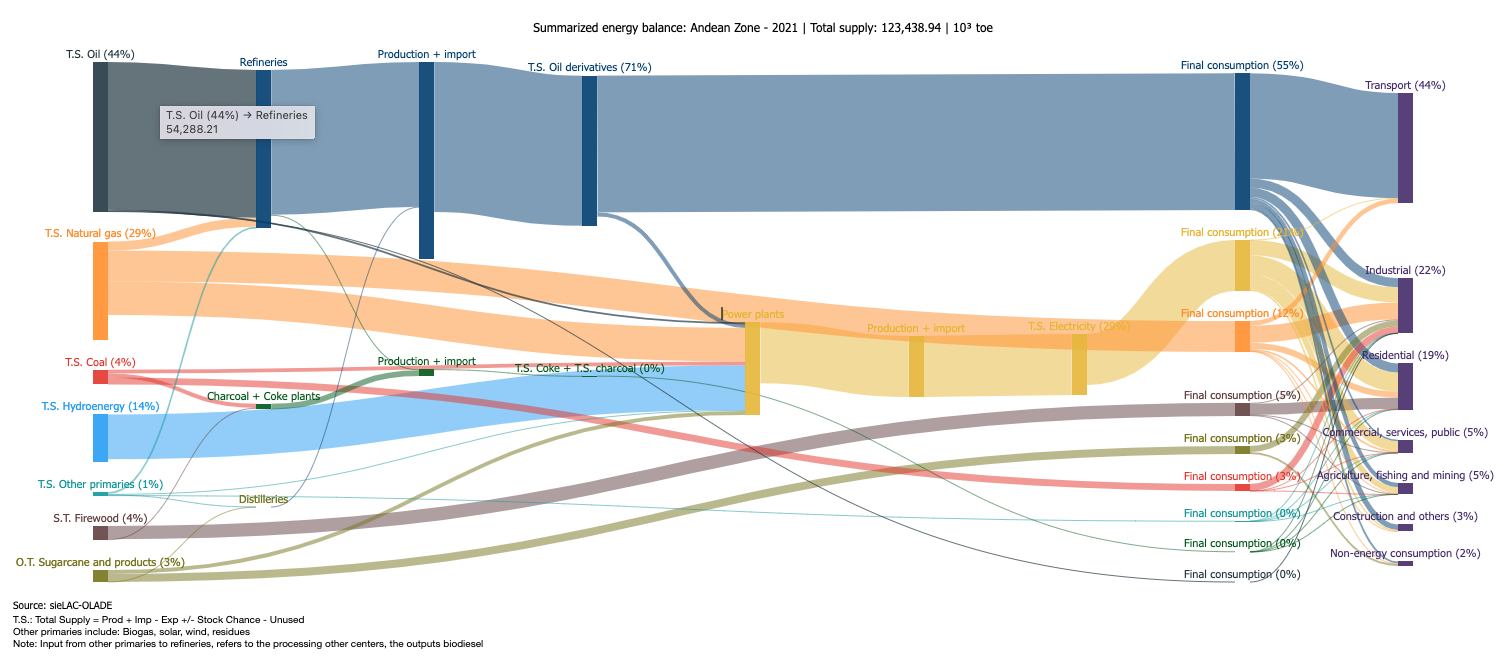
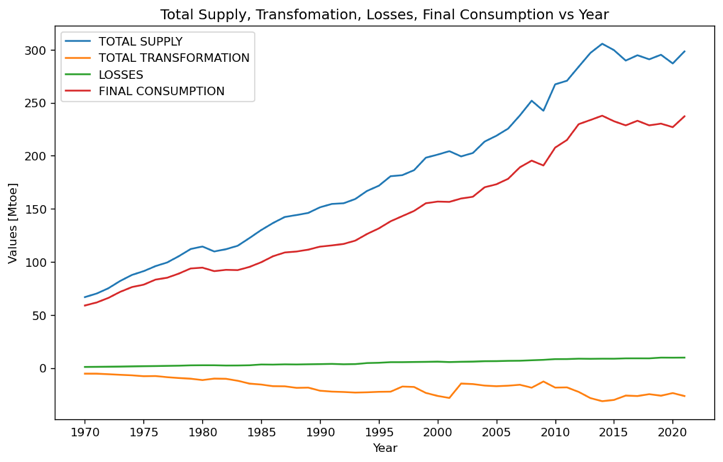


Figure 1: Sankey diagram of energy balance in Brazil, 2021



Energy supplies depend on both primary and secondary energy sources.

Gráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

Final energy consumption is allocated across various sectors within the country, including transportation, industry, residential, commercial- public services, agriculture, fishing-mining, construction, and non-energy consumption.

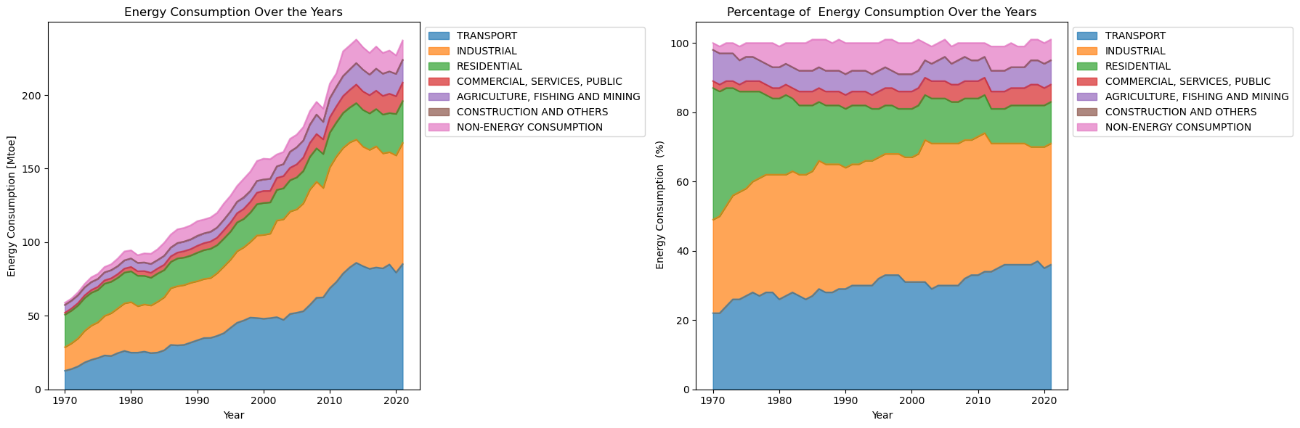


Figure 2: Energy consumption trends over time (Volume - a, Percentage - b)

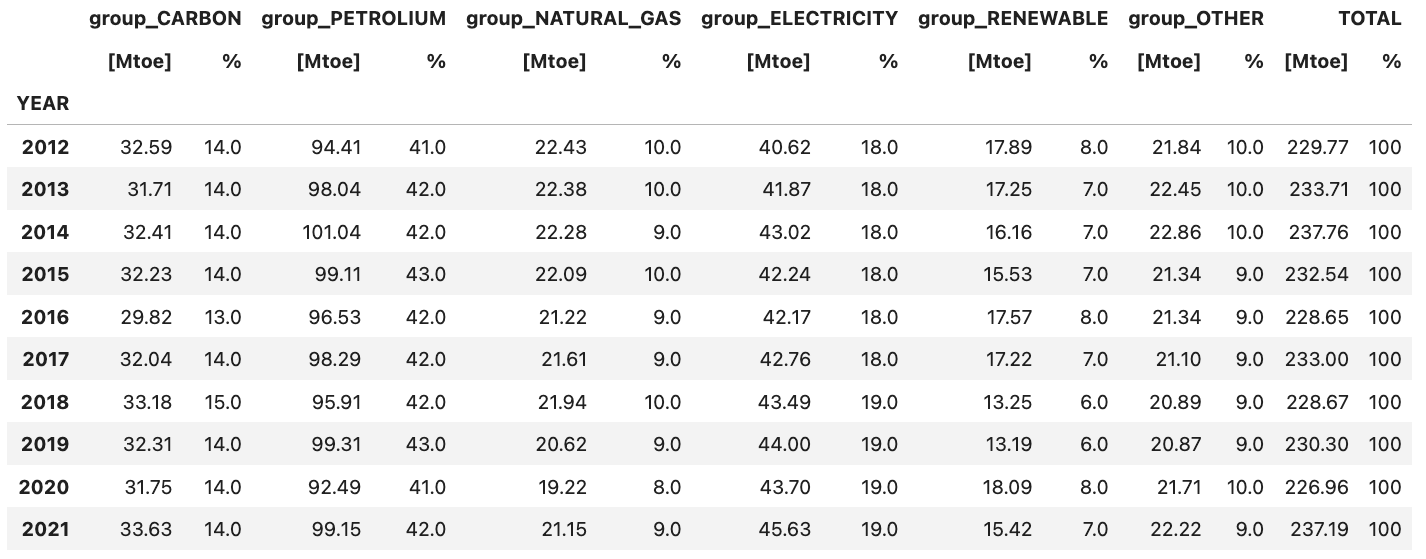


Table 1: Energy sources in final consumption over the recent 10 years

2.2 ARIMA Model – Univariate problem.

We apply the ARIMA model to all sectors, but first, we need to perform some checks to select the best parameters (p, d, q).

In time series analysis, whether we need to scale the data (i.e., standardize or normalize it) depends on the specific characteristics of the data and the modeling techniques we plan to use. Scaling may or may not be required.

In this section, we are focusing on modeling the univariate problem.

2.2.1 ARIMA for all sectors

Step 1: Manually check to chose d:

In this step, we first check the stationarity of a time series by performing a statistical test, such as the Augmented Dickey-Fuller (ADF) test, and examining the p-value associated with the test. If the p-value is smaller than 0.05, we can consider the series as stationary; otherwise, we make the series stationary by applying differencing. The minimum order of differencing is the value of d.

Additionally, we plot the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) at each differencing order. This step allows us to select the appropriate values for 'p' and 'q' based on the characteristics of the ACF and PACF plots.

For all sectors, the second differencing (d = 2) renders the time series stationary, indicating that a second-order differencing will be applied in the ARIMA modeling of each sector to achieve the required stationarity. (CHECK ANEXOS).

Step 2: Find the best (p,q) manually

This step is optional, as p and q can also be selected by finding the minimum values of AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) if d is already determined.

The optimal values of p and q vary across sectors, and the choice of criteria (AIC or BIC) can influence the selection. These variations indicate the complexity of modeling energy consumption in different sectors, and the selection of the best ARIMA model may depend on the specific context and objectives of the analysis.

2.2.2 Model Training, Validation and Prediction

In our modeling process, we allocated a significant portion of the dataset for training and validation purposes. Specifically, we used 90% of the data for training and 10% for validation. This split allowed us to effectively train our ARIMA models and assess their performance.

For the validation phase, we employed the validation dataset to evaluate the accuracy of our models' predictions. This involved comparing the predicted values generated by our ARIMA models to the actual data in the validation set. We then calculated the Mean Absolute Percentage Error (MAPE) to quantify the level of accuracy achieved. MAPE provided valuable insights into the models' performance, helping us understand how well they could generalize to unseen data. The MAPE values for each sector were as follows:

TRANSPORT: 9.37%

INDUSTRIAL: 9.76%

RESIDENTIAL: 14.27%

COMMERCIAL, SERVICES, PUBLIC: 9.69%

AGRICULTURE, FISHING AND MINING: 9.03%

CONSTRUCTION AND OTHERS: 9165359077.49% (an outlier)

NON-ENERGY CONSUMPTION: 22.68%

FINAL CONSUMPTION: 8.57%

These MAPE values provided valuable insights into the models' performance, helping us understand how well they could generalize to unseen data. While most sectors achieved reasonable MAPE values, it's essential to address the outlier in the "CONSTRUCTION AND OTHERS" sector, where the MAPE appears exceptionally high, indicating potential issues with the model's performance for this sector.

In the prediction phase, we projected future energy consumption for each sector. Specifically, we forecasted energy consumption trends for the next 20 years, allowing us to make informed predictions about energy consumption in the years ahead.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 4: Prediction FINAL ENERGY CONSUMPTION 20 more years

Gráfico

Descripción generada automáticamente

Figure 5: Prediction ENERGY CONSUMPTION BY SECTORS 20 more years

In the prediction phase, we projected future energy consumption for each sector. Specifically, we forecasted energy consumption trends for the next 20 years. To do this, we used the trained ARIMA models and specified the number of periods we wanted to forecast. This allowed us to make informed predictions about energy consumption in the years ahead.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 6: Prediction SUM OVER TOTAL SECTORIALS 20 more years

By following this approach of training, validating, and predicting, we ensured that our models were robust and capable of making reliable forecasts for various energy consumption sectors. These steps were essential in preparing our data for further analysis and decision-making processes.

2.2.3 Scenarios

In this section, we have performed a series of data preparation and analysis steps to create scenarios for future energy consumption across different energy source groups. The main objectives of this analysis were to understand historical energy consumption patterns, classify energy sources into meaningful groups, and estimate how these groups might contribute to future energy consumption.

We began by selecting the "FINAL CONSUMPTION" sector from our dataset and cleaning the data to remove unnecessary columns. Next, we categorized energy sources into distinct groups, such as CARBON, PETROLEUM, NATURAL GAS, RENEWABLE, ELECTRICITY, and OTHER. This grouping allowed us to aggregate energy consumption within each category.

To understand the historical distribution of energy sources, we calculated the percentage distribution of each group relative to the total energy consumption for each year. These percentages were rounded to whole numbers for simplicity.

The resulting data was then combined into a comprehensive DataFrame, final\_df, which includes historical energy consumption values (in Mtoe) and the percentage distribution of energy sources for each year. This DataFrame provides a clear overview of energy consumption trends and source distribution up to the present year.

Additionally, we created scenarios for future energy consumption by estimating how energy source groups might contribute to total energy consumption in the coming years. These scenarios are based on the assumption that the distribution of energy sources will remain consistent with recent historical trends.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figure 7: Prediction SUM OVER TOTAL SECTORIALS 20 more years

For future energy consumption scenarios (from 2022 to 2042), we observe several key trends:

Carbon-Based Sources: The consumption of carbon-based energy sources (CARBON) remains relatively stable over the years, with a slight upward trend.

Petroleum-Based Sources: Energy consumption from petroleum-based sources (PETROLEUM) also shows a gradual increase, reflecting continued demand.

Natural Gas: The consumption of natural gas (NATURAL GAS) maintains a consistent growth pattern, contributing significantly to overall energy consumption.

Electricity: Electricity consumption (ELECTRICITY) steadily rises throughout the forecast period, indicating its importance in various sectors.

Renewable Energy: The consumption of renewable energy sources (RENEWABLE) sees a continuous, albeit moderate, increase, highlighting the growing emphasis on sustainable energy.

Other Sources: Energy from other sources (OTHER) exhibits a steady upward trajectory, likely influenced by diverse energy alternatives.

These scenarios can serve as valuable tools for stakeholders involved in energy planning, policymaking, and decision-making. They provide insights into potential shifts in energy consumption patterns, enabling informed strategies and actions to address future energy challenges effectively.

### Análisis descriptivo del conjunto de datos.

**Suministro Total, Transformación Total y Consumo Total en el período de estudio.**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

* **Total Supply (Suministro Total):** La línea azul representa la tendencia del Suministro Total de energía. Inicialmente, en la década de 1970, el suministro total estaba en torno a los 65,000-70,000 unidades. A lo largo de los años, hay una tendencia creciente, con algunos altibajos, llegando a un pico en los años más recientes (aproximadamente en 2021) en alrededor de 280,000 unidades.
* **Total Transformation (Transformación Total):** La línea naranja representa la tendencia de la Transformación Total de energía. En 1970, la transformación total estaba cerca de -10,000 unidades. A lo largo de los años, hay una tendencia creciente en la transformación total, llegando a un pico positivo alrededor de 2021.
* **Total Consumption (Consumo Total):** La línea verde muestra la tendencia del Consumo Total de energía. Comienza en alrededor de 60,000 unidades en 1970 y muestra un aumento gradual en los primeros años. Sin embargo, a partir de alrededor de 1980, hay un aumento más significativo en el consumo total, alcanzando un pico en alrededor de 2021 en alrededor de 240,000 unidades.

**Suministro de energía primaria en el período de estudio.**

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

A lo largo del tiempo, Brasil ha mostrado una diversificación en sus fuentes de energía primaria. En la década de 1970, el país dependía en gran medida de la quema de leña, así como del petróleo y sus derivados. Con el paso de los años, se ha producido un cambio significativo en la matriz energética, incorporando fuentes como el gas natural, la hidroenergía y el carbón.

En el **primer gráfico**, que representa el suministro de energía primaria a lo largo de los años, se destacan varias tendencias clave. En las décadas iniciales, el petróleo (OIL) y el gas natural (NATURAL GAS) fueron las principales fuentes de energía primaria, con un suministro que continuó aumentando a lo largo del tiempo. A partir de la década de 1980, se observa un marcado aumento en la contribución del gas natural, lo que sugiere su creciente importancia en la matriz energética del país.

El **segundo gráfico**, que muestra la contribución porcentual de cada fuente de energía primaria con respecto al suministro total, revela aún más detalles sobre la transformación de la matriz energética de Brasil. A medida que avanzan los años, el gas natural emerge como un componente vital, mostrando un aumento constante en su porcentaje de contribución. Esta tendencia sugiere que Brasil ha dependido cada vez más del gas natural para satisfacer sus necesidades energéticas.

Aunque se ha observado una diversificación en las fuentes de energía, el petróleo y sus derivados, así como el gas natural, siguen siendo componentes fundamentales en la matriz energética de Brasil. Ambos siguen representando una parte significativa del suministro total.

La energía nuclear y otras fuentes no especificadas de energía primaria también tienen una presencia notable en la tabla. Sin embargo, su participación en el suministro total sigue siendo relativamente baja.

En resumen, los gráficos resaltan cómo Brasil ha transitado hacia una matriz energética más sostenible y diversificada a lo largo del tiempo, con un aumento en la participación de la hidroenergía y un crecimiento moderado en el uso del gas natural, así como un decrecimiento moderado en el uso del petróleo. Estos cambios indican una estrategia de diversificación energética en Brasil en busca de una mayor seguridad energética y una menor dependencia de los combustibles fósiles, aunque es fundamental seguir avanzando hacia fuentes de energía verdaderamente limpias y renovables.

**Evolución de Importación y Exportación de Fuentes Primaria en el período de estudio.**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

De 1970 a 2000, Brasil experimentó una transformación en su posición energética, pasando de ser un país importador de energía a convertirse en un importante exportador neto de energía en algunos aspectos. Aquí hay un resumen de cómo evolucionó la situación energética de Brasil durante ese período:

* **Hasta la década de 1970:**

Brasil dependía en gran medida de la importación de petróleo y derivados para satisfacer sus necesidades energéticas. La economía brasileña estaba en crecimiento, lo que aumentaba la demanda de energía y, por lo tanto, la dependencia del petróleo importado.

* **Década de 1970:**

En la década de 1970, Brasil comenzó a enfrentar problemas con la inflación y la deuda externa, que se agravaron por el alto costo de las importaciones de petróleo. Como respuesta a estos desafíos, Brasil inició programas para reducir su dependencia del petróleo importado y diversificar su matriz energética. Década de 1980:

Durante esta década, Brasil implementó políticas para promover la producción interna de petróleo y gas. Se descubrieron y desarrollaron importantes yacimientos de petróleo en la cuenca de Campos, lo que contribuyó a aumentar la producción de petróleo en el país. Década de 1990:

Brasil continuó avanzando en la explotación de sus recursos energéticos, incluidos el petróleo y la energía hidroeléctrica. El país también comenzó a desarrollar biocombustibles, especialmente etanol derivado de la caña de azúcar, como una alternativa más sostenible a los combustibles fósiles.

* **Para la década de 2000**:

Brasil había avanzado en la producción interna de petróleo y gas, lo que le permitió reducir significativamente sus importaciones de petróleo. El país también dependía en gran medida de la energía hidroeléctrica, con la construcción de grandes represas que le proporcionaron una fuente significativa de energía renovable. Además, Brasil se convirtió en un exportador neto de petróleo y productos energéticos, gracias a su creciente producción interna y la explotación de sus recursos naturales. En resumen, a lo largo de las décadas de 1970 a 2000, Brasil pasó de ser un país importador de energía a ser un exportador neto en ciertos aspectos, particularmente en lo que respecta al petróleo y sus derivados. La diversificación de su matriz energética y el desarrollo de sus recursos naturales jugaron un papel crucial en esta transformación.

**Suministro de energía secundaria en el período de estudio.**

Gráfico, Histograma

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

* Primer Gráfico - Energía Secundaria (Cantidades Absolutas):

El **primer gráfico** de área apilada muestra la evolución de la energía secundaria suministrada en Brasil a lo largo del tiempo en cantidades absolutas. Cada área coloreada representa un tipo de fuente de energía secundaria, y la altura total de la pila en cada año representa el suministro total de energía secundaria.

En las décadas iniciales (1970-1980), la mayoría de las fuentes de energía secundaria experimentaron un crecimiento constante. La electricidad (ELECTRICITY) y los derivados del petróleo, como el diésel (DIESEL OIL) y la gasolina (GASOLINE/ALCOHOL), fueron las principales contribuyentes al suministro de energía secundaria.

A medida que avanzan los años, hay un aumento significativo en el suministro de electricidad y productos derivados del petróleo, como el gasóleo y la gasolina. Sin embargo, a partir de mediados de la década de 1990, se observa una tendencia decreciente en el suministro de derivados del petróleo y un aumento en la contribución de otras fuentes de energía secundaria, como el carbón (COKE) y la biomasa (CHARCOAL).

* Segundo Gráfico - Porcentaje de Energía Secundaria:

El **segundo gráfico** de área apilada muestra la composición de la matriz de energía secundaria de Brasil en términos de porcentaje de cada fuente con respecto al suministro total de energía secundaria en cada año.

Durante las primeras décadas, los derivados del petróleo y la electricidad dominaron la matriz de energía secundaria, con un enfoque particular en el diésel y la gasolina. Sin embargo, a medida que avanzan los años, se pueden observar varios cambios en la composición de la matriz energética.

A finales de la década de 1990 y principios de la década de 2000, hay una disminución en la participación de los derivados del petróleo, mientras que otras fuentes de energía, como la electricidad y el carbón, mantienen una presencia significativa. Además, a medida que avanzan los años, se observa un aumento en la participación de fuentes de energía más limpias y renovables, como la biomasa y la electricidad.

En resumen, los gráficos muestran una transformación en la matriz de energía secundaria de Brasil a lo largo del tiempo. Se ha producido una disminución gradual en la dependencia de los derivados del petróleo y un aumento en la participación de fuentes de energía más diversificadas, incluidas las renovables. Esto sugiere una tendencia hacia una matriz energética más sostenible y diversificada en Brasil, con un enfoque en la reducción de la dependencia de los combustibles fósiles.

**Consumo de energía por sectores en el período de estudio***.*

Gráfico

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

* Gráfico 1: Energy Consumption by Sector Over the Years

Este gráfico representa la distribución del consumo de energía en diferentes sectores a lo largo de los años. Cada área coloreada en el gráfico corresponde a un sector específico, como Agricultura, Industria, Transporte, Residencial, Comercial y otros. La altura de cada área en un año determinado muestra la cantidad total de energía consumida por ese sector en ese año.

* Agricultura: Durante los primeros años, la agricultura tenía una participación significativa en el consumo de energía, pero a medida que avanzó el tiempo, su contribución disminuyó.
* Industria: La industria es uno de los principales consumidores de energía y su contribución ha sido constante, con algunos aumentos notables en ciertos años.
* Transporte: El consumo de energía en el sector del transporte ha aumentado de manera constante a lo largo de los años, lo que indica un crecimiento en la movilidad y la demanda de combustibles.
* Residencial: La energía utilizada en los hogares ha aumentado, posiblemente debido al aumento de la población y las necesidades de calefacción, refrigeración y electrónica.
* Comercial y Otros: Este sector incluye edificios comerciales y otros usos diversos de energía. Su contribución generalmente se mantiene constante.
* Gráfico 2: Percentage of Energy Consumption by Sector Over the Years

Este gráfico muestra la proporción de consumo de energía que cada sector representa en relación con el total. En lugar de valores absolutos, se centra en la distribución relativa en forma de porcentaje.

* A lo largo del tiempo, la industria sigue siendo una parte significativa del consumo total de energía.
* El transporte también ha aumentado en términos de porcentaje, lo que indica un aumento en la importancia y demanda del transporte en el consumo de energía.
* El sector residencial muestra una disminución relativa en comparación con otros sectores, lo que sugiere un cambio en la composición del consumo de energía.
* El sector agrícola ha disminuido constantemente en términos de porcentaje.
* Comercial y Otros se mantiene relativamente constante en términos de porcentaje.

Exploración y limpieza del conjunto de datos (tratamiento de los valores nulos).

El dataset que hemos utilizado para este trabajo se trata del balance de la matriz energética desde 1970 a 2021 de Brasil descargado de la página web de OLADE.

Los datos de la matriz energética los tenemos contenidos en un archivo Excel con múltiples hojas llamado "Brazil\_Energy balance matrix.xlsx".

Realizamos acciones como redondear valores, renombrar columnas, eliminar espacios y corregir nombres. Además, agregamos el año como una columna. Convertimos las unidades de medida de Ktoe (kilotoneladas equivalentes de petróleo) a Mtoe (megatoneladas equivalentes de petróleo) en todas las columnas numéricas. Los valores numéricos se redondean a dos decimales. Se eliminan los encabezados de fila que contienen unidades de medida, y se renombran las columnas 'Unnamed: 0' a 'SECTOR' y 'OTHER PRIMARY\_x000d\_' a 'OTHER PRIMARY'. Además, se corrige el nombre del sector 'COKE PLANTS AND BLAST FURNACES\_x000d\_' a 'COKE PLANTS AND BLAST FURNACES'. Estas transformaciones aseguran que los datos estén en un formato coherente y listos para su análisis posterior. Los datos procesados se almacenan en un nuevo diccionario llamado modified\_dict, que reemplaza el original. Finalmente, se guarda el diccionario limpio en un nuevo archivo Excel llamado "Brazil\_Energy balance matrix\_cleaned.xlsx", preservando las hojas originales para su análisis y visualización posteriores. Este proceso asegura la calidad y coherencia de los datos antes de su utilización en análisis más profundos.

Después de esta conversión empezamos a analizar nuestro dataset.

**3. Desarrollo del modelo de predicción**

En la sección de "Desarrollo del modelo de Machine Learning," se abordará la parte central de este proyecto de análisis de energía en Brasil. El objetivo principal en esta fase es construir un modelo capaz de realizar pronósticos precisos sobre el consumo energético en el país. En este contexto, se ha elegido el modelo ARIMA (Autoregressive Integrated Moving Average) como enfoque para abordar esta tarea. La elección de ARIMA se basa en diversas razones fundamentales que respaldan su idoneidad para modelar y predecir series temporales, como la demanda de energía.

Esas razones consisten en:

1. **Adecuado para Series Temporales**: ARIMA es ampliamente reconocido y utilizado en la modelización de series temporales, lo que lo hace altamente relevante para el pronóstico de datos que varían con el tiempo, como el consumo de energía.
2. **Capacidad para Capturar Patrones Temporales**: El modelo ARIMA es eficaz para capturar patrones, tendencias y estacionalidades en los datos temporales, lo que lo hace apropiado para predecir la demanda de energía que puede mostrar variaciones a lo largo del año.
3. **Estacionalidad y Tendencias**: Permite la identificación y tratamiento de componentes estacionales y tendencias en los datos, lo que es crucial para el análisis de la demanda de energía, que a menudo exhibe patrones estacionales y cambios a lo largo del tiempo debido a factores económicos y estacionales.
4. **Comportamiento de Series no Estacionaria**s: ARIMA puede manejar series temporales no estacionarias mediante la diferenciación, lo que es común en datos energéticos que pueden mostrar cambios a lo largo del tiempo.
5. **Base en Principios Estadísticos**: Este modelo se basa en principios estadísticos sólidos, lo que lo convierte en una elección confiable y respaldada por la teoría.

Introducción a los modelos desarrollados

**Selección del modelo óptimo**

Los modelos ARIMA se aplicarán a todos los sectores de consumo energético en este análisis. Sin embargo, antes de implementar el modelo, es esencial realizar una selección cuidadosa de los parámetros óptimos (p, d, q) para garantizar que el modelo se ajuste adecuadamente a las características de las series temporales no estacionarias de consumo de energía en Brasil.

Hemos decidió no escalar las variables de nuestro dataframe debido a las siguientes casuísticas:

1. **Naturaleza de las Series Temporales:** Las series temporales ya representan datos en función del tiempo, lo que significa que las unidades y las magnitudes de las variables pueden tener sentido en su forma original. En el caso del consumo de energía, las unidades generalmente son medidas físicas como en nuestro caso, megatoneladas equivalentes de petróleo (Mtoe), y estas unidades son inherentemente relevantes para la interpretación de los datos.

1. **Conservación de la Interpretabilidad:** Al no escalar las variables, los valores en las predicciones del modelo ARIMA se mantendrán en la misma escala y unidades que los datos originales. Esto facilita la interpretación de los resultados, ya que los pronósticos seguirán siendo comprensibles y significativos en términos del consumo real de energía, lo que es fundamental para los responsables de la toma de decisiones y los expertos en energía.
2. **Sensibilidad al Contexto:** Escalar las variables puede afectar la magnitud de los coeficientes en el modelo y, en consecuencia, las estimaciones de las predicciones. En un contexto como el análisis de energía, donde las cifras reales son críticas para la planificación y la política, realizar un escalado podría introducir cierta opacidad en la interpretación de los resultados.
3. **Enfoque Tradicional de ARIMA:** El modelo ARIMA se desarrolló originalmente para trabajar con series temporales en su forma original, sin la necesidad de escalamientos adicionales. Esta metodología se basa en la idea de que los patrones y las relaciones inherentes en los datos se pueden modelar sin requerir escalas uniformes.

Para asegurar que nuestros modelos ARIMA sean robustos y precisos en la predicción de la demanda de energía, se realizarán análisis exhaustivos para determinar la mejor combinación de parámetros. Esto permitirá que el modelo se ajuste de manera óptima a las variaciones y patrones específicos presentes en los datos de consumo de energía de Brasil, lo que, a su vez, respaldará decisiones fundamentadas en políticas y planificación energética.

En el proceso de selección del modelo óptimo en el contexto de modelos ARIMA aplicados al análisis de series temporales de consumo de energía para cada sector en Brasil, se utilizan varias funciones fundamentales. Estas funciones desempeñan un papel crítico en la evaluación y ajuste de los modelos para garantizar que sean apropiados para las características específicas de los datos, las cuales son:

1. **Función MAPE (Error Porcentual Absoluto Medio):**

La función **MAPE** se utiliza para calcular el Error Porcentual Absoluto Medio (MAPE) entre las observaciones reales y las predicciones generadas por un modelo ARIMA. El MAPE proporciona una medida de la precisión de las predicciones expresada como un porcentaje del error con respecto a las observaciones reales. Un MAPE más bajo indica predicciones más precisas y es una métrica esencial para la selección del modelo óptimo.

1. **Función test\_stationarity (Prueba de Dickey-Fuller para Estacionariedad):**

La función test\_stationarity se emplea para evaluar si una serie temporal es estacionaria o no. La estacionariedad es un requisito fundamental para aplicar modelos ARIMA, ya que estos modelos asumen que la estadística de la serie no cambia con el tiempo.El Test de Dickey-Fuller comprueba si el valor p es lo suficientemente bajo como para rechazar la hipótesis nula de no estacionariedad.

1. **Función residcheck (Evaluación de Residuos):**

La función residcheck se utiliza para evaluar los residuos generados por un modelo ARIMA. Los residuos son las diferencias entre las observaciones reales y las predicciones del modelo. Esta función verifica si los residuos cumplen con condiciones deseables, como ser ruido blanco (sin patrones discernibles), tener una media cercana a cero y una varianza constante. También evalúa la autocorrelación, normalidad y estacionariedad de los residuos para garantizar que el modelo haya capturado adecuadamente la información de los datos.

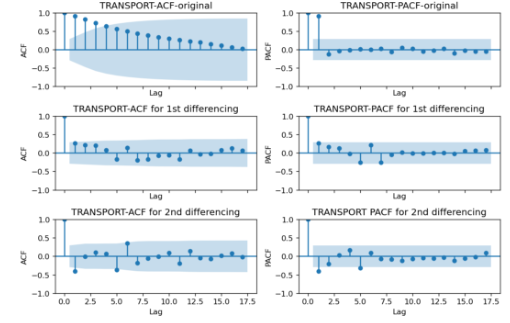
1. **Función order\_aic\_bic (Selección de Parámetros p y q):**

Esta función realiza un bucle a través de diferentes valores de 'p' (orden de autoregresión) y 'q' (orden de media móvil) en un rango de 0 a 4. Para cada combinación de 'p' y 'q', se ajusta un modelo ARIMA a los datos y se calculan los valores de AIC y BIC. Los resultados se almacenan en una lista junto con los valores de 'p' y 'q', y luego se presenta una tabla de modelos ordenados según los valores más bajos de AIC y BIC.

**Búsqueda del valor “d”:**

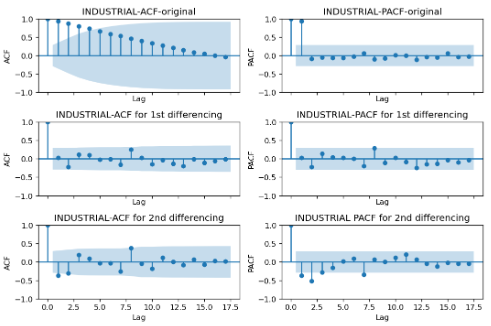
Realizamos una búsqueda para la selección apropiada del valor de 'd' (orden de diferenciación) para cada sector. Para esto, utilizaremos la prueba de estacionariedad de Dickey-Fuller hasta que se alcanza la estacionariedad (cuando el p-valor es menor que 0.05).

Los cuáles nos dan los siguientes resultados para cada sector:

**TRANSPORT:**

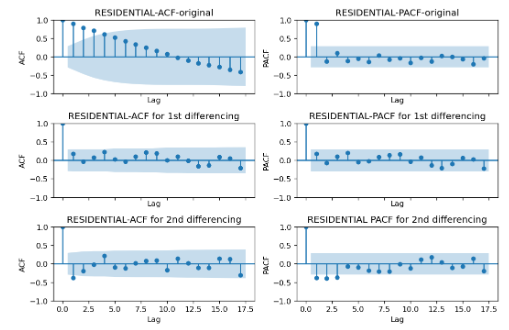
* Test Statistic (Estadístico de Prueba) - 2.756864 (1st diff), -3.444595 (2nd diff)
* p-value (Valor p): 0.064704 (1st diff), 0.009529 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**INDUSTRIAL:**

****

* Test Statistic: -5.317197 (1st diff), -3.912924 (2nd diff)
* p-value: 0.000005 (1st diff), 0.001942 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**RESIDENTIAL:**

****

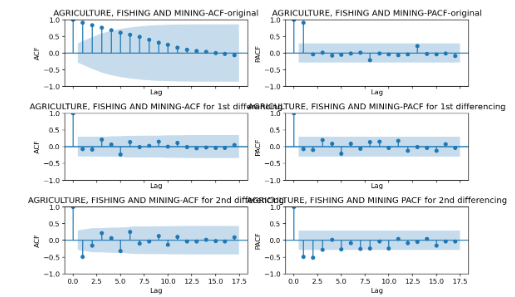
* Test Statistic: -5.425666 (1st diff), -7.081189 (2nd diff)
* p-value: 0.000003 (1st diff), 4.661825e-10 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**COMMERCIAL, SERVICES, PUBLIC:**

****

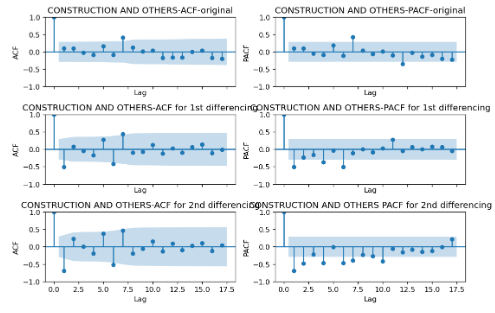
* Test Statistic: -5.187659 (1st diff), -5.083537 (2nd diff)
* p-value: 0.000009 (1st diff), 0.000015 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**AGRICULTURE, FISHING AND MINING:**

****

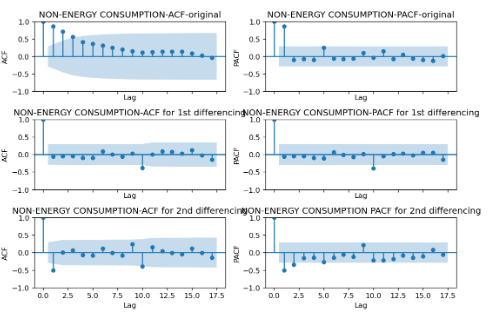
* Test Statistic: -6.949322e+00 (1st diff), -6.809117e+00 (2nd diff)
* p-value: 9.785090e-10 (1st diff), 2.138701e-09 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**CONSTRUCTION AND OTHERS:**

****

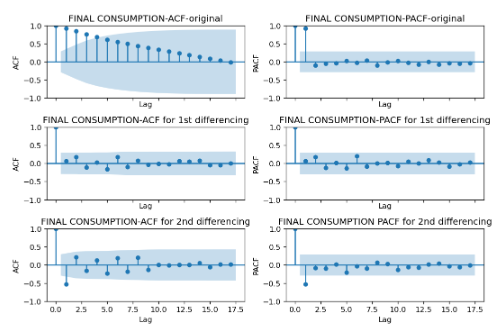
* Test Statistic: -5.843043 (1st diff), -5.001523 (2nd diff)
* p-value: 3.738493e-07 (1st diff), 0.000022 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**NON-ENERGY CONSUMPTION:**

****

* Test Statistic: -2.172822 (1st diff), -2.724450 (2nd diff)
* p-value: 0.216257 (1st diff), 0.069902 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

**FINAL CONSUMPTION:**

****

* Test Statistic: -1.457273 (1st diff), -4.365760 (2nd diff)
* p-value: 0.554532 (1st diff), 0.000341 (2nd diff)
* Conclusión: La segunda diferenciación (2nd diff) hace que la serie sea estacionaria con un p-value menor que 0.05. Se selecciona d = 2.

Para todos los sectores, la segunda diferenciación (d = 2) hace que las series sean estacionarias, lo que indica que se aplicará una diferenciación de orden 2 en la modelización ARIMA de cada sector para lograr la estacionariedad requerida.